PODPIS ZAUFANY
BOGDAN ALEKSANDER
RUSZCZAK
02.12.2025 13:37:19 GMT+1
Dokument podpisany elektronicznie
podpisem zaufanym



Politechnika Opolska

Wydział Informatyki

Autoreferat

na temat dorobku
i osiągnięć w pracy naukowo-badawczej

Autor: dr inż. Bogdan Ruszczak

Opole, grudzień 2025

Spis treści

Wstęp **3**

1 Informacje o wykształceniu i zatrudnieniu **4**
1.1 Dane osobowe 4
1.2 Przebieg pracy zawodowej i naukowej 4
1.3 Uzyskane tytuły oraz stopnie naukowe 4
1.4 Informacje dodatkowe 5

2 Osiągnięcia naukowe stanowiące podstawę wniosku o wszczęcie postępowania habilitacyjnego **6**
2.1 Wybrane publikacje 6
2.2 Opis osiągnięcia 8
2.3 Podsumowanie 33
2.4 Szczegółowe statystyki bibliograficzne 34
2.5 Lista pozostałych publikacji po doktoracie 37

3 Pozostałe osiągnięcia **42**
3.1 Projekty badawczo rozwojowe 42
3.2 Staże badawcze 45
3.3 Profesor wizytujący w uczelniach partnerskich 45
3.4 Opublikowane zbiory danych 46
3.5 Nagrody 47
3.6 Recenzje i prace redakcyjne 48
3.7 Praca z dyplomantami 49
3.8 Osiągnięcia dydaktyczne 50
3.9 Inne osiągnięcia organizacyjne 51
3.10 Uczestnictwo i wygłoszenie artykułu w międzynarodowych i krajowych konferencjach naukowych 52
3.11 Artykuły opublikowane na konferencjach 53
3.12 Członkostwo w organizacjach 54

Literatura **55**

A Załączniki **58**

Wstęp

W niniejszym autoreferacie przedstawiam wszystkie informacje dotyczące mojej działalności naukowo-badawczej i dydaktycznej z okresu obejmującego czas po uzyskaniu stopnia naukowego doktora (tj. po roku 2012), do dnia dzisiejszego. W opracowaniu znaleźć można dane dotyczące dorobku naukowego, zestawienie wszystkich artykułów, rozdziałów w monografiach i materiałach konferencyjnych. Omówiony jest dorobek dydaktyczny, w tym informacje o pracy z dyplomantami, prowadzonych zajęciach i uzyskanych nagrodach. W rozdziale drugim prezentuję dokładny opis osiągnięć, będących główną częścią tego autoreferatu. Informatyka zawsze była i wciąż pozostaje w obszarze moich najważniejszych zainteresowań. Z tego powodu zdecydowałem się na studia w tym kierunku, które okazały się być solidną podstawą w trakcie poznawania jej zagadnień, ale też dostarczyły aparatu badawczego dla badań prowadzonych na rzecz innych dyscyplin naukowych.

W badaniach, które prowadzę, zasadniczym elementem są kompleksowo wykorzystywane metody uczenia maszynowego (ang. *machine learning* – ML). Przeglądając prezentowane zestawienie dorobku, warto zwrócić uwagę, że jest on spójny. Przytaczając wybrane prace, chciałem zwrócić uwagę na naukowe i praktyczne wyzwania związane z implementacją metod ML dla rzeczywistych problemów, oraz ewaluacją tych metod w powtarzalny sposób, pozwalający na zachowanie wiarygodności raportowanych wyników.

Na prezentowane osiągnięcia składa się **dziesięć prac**, stanowiących tematycznie spójną całość. Niektóre z wątków prowadzonych badań są wciąż kontynuowane, a część wyników jest już opublikowana, i prace te również mogłyby być uwzględnione w tym opracowaniu. Przykładowo artykuł opublikowany na konferencji GECCO 2023 dotyczący ewolucyjnej konfiguracji hiperparametrów konwolucyjnej sieci do prognozowania cen energii [B20], artykuł z konferencji ECML–PKDD 2025 dotyczący nienadzorowanego detektora anomalii wspomaganego filtrowaniem morfologicznym [B3] czy opublikowany w czasopiśmie IEEE Geoscience and Remote Sensing Magazine artykuł [B9], podsumowujący udany projekt badawczy [P11], w którym brałem udział i dotyczący estymacji parametrów gleby z użyciem obrazowań hiperspektralnych, również przedstawiają moje ważne osiągnięcia, ale nie zmieściły się w przedstawianym do oceny zestawieniu.

Moja praca dydaktyczna również jest ściśle powiązana z dziedziną uczenia maszynowego oraz inżynierii danych. Wśród przedmiotów, dla których opracowałem program zajęć i które prowadzę, znajdują się *zastosowania sztucznej inteligencji*, w których studenci wykorzystują narzędzia i metody uczenia maszynowego oraz implementują je w konkretnych zadaniach. Wykładam również kurs z zakresu *Zaawansowanych baz danych*, podczas których studenci poznają współczesne narzędzia inżynierii danych. Szczegółowe dane dotyczące mojej pracy dydaktycznej (podrozdział 3.8) oraz informacje statystyczne uzupełniające obraz mojego dorobku, zamieściłem w rozdziale trzecim tego opracowania.

1 Informacje o wykształceniu i zatrudnieniu

1.1 Dane osobowe

a. Imię i nazwisko: Bogdan Ruszczak

[redacted]

1.2 Przebieg pracy zawodowej i naukowej

- 2018–obecnie – **adiunkt w Instytucie Informatyki** (od 2020, w Katedrze Informatyki, a od 2025 Katedrze Systemów Informatycznych i Sterowania), na Wydziale Elektrotechniki, Automatyki i Informatyki (od 2025 Wydziale Informatyki) Politechniki Opolskiej, kierowane [P1] i realizowane projekty badawcze: [P5]
- 2022–2025 – **specjalista ds. uczenia maszynowego** w KP Labs sp. z o.o., Gliwice (od 2019 status centrum badawczo-rozwojowego Ministerstwa Rozwoju i Technologii RP), realizowane międzynarodowe projekty badawcze: [P4, P6, P7, P8]
- 2018–2022 – **ekspert ds. uczenia maszynowego, kierownik projektu B+R** w QZ Solutions sp. z o.o., Opole (od 2019 status centrum badawczo-rozwojowego Ministerstwa Rozwoju i Technologii RP), kierowane [P2, P3] i realizowane projekty badawcze: [P11, P12]
- 2018 – **starszy programista** w Future Processing sp. z o.o., Gliwice
- 2014–2018 – **adiunkt** w Katedrze Ekonomii, Finansów i Badań Regionalnych, na Wydziale Ekonomii i Zarządzania Politechniki Opolskiej
- 2009–2014 – **asystent** w Katedrze Ekonomii i Badań Regionalnych, na Wydziale Zarządzania i Inżynierii Produkcji Politechniki Opolskiej (od 2013, na Wydziale Ekonomii i Zarządzania)
- 2005–2009 – **asystent** w Instytucie Inżynierii Produkcji, na Wydziale Zarządzania i Inżynierii Produkcji Politechniki Opolskiej

1.3 Uzyskane tytuły oraz stopnie naukowe

- listopad 2012 – doktor inżynier:
  - Politechnika Wrocławska, Wydział Informatyki i Zarządzania,

  - dziedzina nauk ekonomicznych, dyscyplina: nauki o zarządzaniu, specjalność: organizacja i zarządzanie,
  - tytuł rozprawy: „Metoda wspomagania oceny efektywności inwestycji w farmy wiatrowe",
  - promotor: prof. dr hab. inż. Zdzisław Szalbierz, prof PWr, recenzenci: prof. dr hab. inż. Rafał Weron (Politechnika Wrocławska), prof. dr hab. Andrzej Graczyk (Uniwersytet Ekonomiczny we Wrocławiu),
  - praca dotyczyła: opracowania, implementacji i ewaluacji metody obliczeniowej dla badania efektywności inwestycji, z uwzględnieniem czynników ryzyka (z zastosowaniem symulacji historycznych dla rozkładów statystycznych dla rzeczywistych danych dotyczących inwestycji w farmy wiatrowe, związanych m.in. ze zmienną charakterystyką warunków pogodowych, awaryjnością instalacji oraz cen energii).

- lipiec 2006 – magister inżynier:
  - Politechnika Opolska, Wydział Elektrotechniki, Automatyki i Informatyki,
  - kierunek: **Informatyka**, specjalność: Sieci komputerowe i systemy baz danych, tryb dzienny,
  - promotor: prof. dr hab. Krzysztof Latawiec,

- lipiec 2005 – magister inżynier:
  - Politechnika Opolska, Wydział Zarządzania i inżynierii Produkcji,
  - kierunek: **Zarządzanie i marketing**, specjalność: Komputerowo Zintegrowane Systemy Zarządzania Produkcją, tryb dzienny,
  - promotor: prof. dr hab. Ryszard Knosala,
  - **praca obroniona z wyróżnieniem**.

1.4 Informacje dodatkowe

- Odbyte szkolenia i posiadane certyfikaty:
  – 2010 – kurs pedagogiczny dla pracowników uczelni wyższej (2 semestry, Wydział Pedagogiczny Wyższej Szkoły Zarządzania i Administracji w Opolu),

- Znajomość języków obcych:
  – angielski: płynnie w mowie i piśmie.

2 Osiągnięcia naukowe stanowiące podstawę wniosku o wszczęcie postępowania habilitacyjnego

Do dorobku wchodzącego w skład osiągnięć naukowych stanowiących podstawę złożenia wniosku habilitacyjnego wchodzi 10 publikacji monotematycznych [A1–A10]. Składa się na niego 5 artykułów z listy JCR [A1, A2, A4, A7, A9], które są indeksowane w bazach Web of Science (WoS) i Scopus, i posiadają sumaryczny wskaźnik Impact Factor (IF) równy 17,4. Druga grupa wybranych publikacji to 5 artykułów z konferencji naukowych posiadających ranking CORE A [A3, A5, A6, A8, A10], które również indeksowane są w bazie Scopus. W publikacjach tych jestem pierwszym autorem i autorem korespondencyjnym. Dwie z nich są jednoautorskie, a dla pozostałych załączono oświadczenia związane z opisem wkładu pracy współautorów oraz jego oszacowaniu procentowym.

Razem składają się one na cykl pt.:

Metody uczenia maszynowego do analizy rzeczywistych danych wielowymiarowych

2.1 Wybrane publikacje

[A1] **B. Ruszczak**, K. Smykała, i K. Dziubański, "The detection of Alternaria solani infection on tomatoes using ensemble learning," *Journal of Ambient Intelligence and Smart Environments*, tom 12, s. 407–418, 2020, ✉ [**MNiSW: 40 pkt, IF 2020: 2,105, cytowania Scopus: 10, WoS: 10; udział: 70%**]
Indywidualny wkład B. Ruszczaka: opracowanie koncepcji badania, manuskryptu, implementacja algorytmów obliczeniowych, zgłoszenie artykułu i odpowiedzi na recenzje, współudział w procesie gromadzenia i przetwarzania danych, analizie wyników, poprawianiu manuskryptu, i pozyskaniu finansowania na publikację[1].

[A2] **B. Ruszczak**, K. Smykała, M. Tomaszewski, i P. J. Navarro Lorente, "Various tomato infection discrimination using spectroscopy," *Signal, Image and Video Processing*, 5 2024, ✉ [**MNiSW: 70 pkt, IF 2024: 2,100, cytowania Scopus: 3, WoS: 2; udział: 40%**]
Indywidualny wkład B. Ruszczaka: współpraca przy opracowaniu koncepcji badania, opracowaniu metodyki, implementacji algorytmów obliczeniowych, współudział w analizie

[1]Przy ustalaniu liczby punktów MNiSW i współczynnika IF dla wszystkich artykułów oraz dla artykułów konferencyjnych rankingu CORE tych konferencji przyjęto wartości obowiązujące w roku publikacji.
✉ – tym symbolem oznaczono status autora korespondencyjnego dla danej publikacji

wyników, pisaniu i poprawianiu manuskryptu, oraz zgłoszenie artykułu i odpowiedzi na recenzje.

[A3] **B. Ruszczak**, "Reducing high-dimensional feature set of hyperspectral measurements for plant phenotype classification," w *Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Lizbona, Portugalia*, S. Silva i L. Paquete, Red. ACM, 2023, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A, cytowania Scopus: 3, WoS: 3; udział: 100%**].

[A4] **B. Ruszczak**, A. M. Wijata, i J. Nalepa, "Unbiasing the estimation of chlorophyll from hyperspectral images: A benchmark dataset, validation procedure and baseline results," *Remote Sensing*, tom 14, no. 21, 2022, ✉ [**MNiSW: 100 pkt, IF 2022: 5,000, cytowania Scopus: 14, WoS: 9; udział: 60%**]
Indywidualny wkład B. Ruszczaka: opracowanie zbioru danych, przygotowanie i oprogramowanie metod i modeli oraz wizualizacji danych, autorstwo wstępnego artykułu oraz współautorstwo końcowej wersji, formalnej analizy oraz przeglądu literatury.

[A5] **B. Ruszczak**, A. M. Wijata, i J. Nalepa, "Estimating chlorophyll content from hyperspectral data using gradient features," w *23rd International Conference on Computational Science - ICCS, Praga, Czechy*, ser. LNCS, J. Mikyska, C. de Mulatier, M. Paszynski *et al.*, Red., tom 14074. Springer, 2023, s. 196–203, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A, cytowania Scopus: 2; udział: 60%**]
Indywidualny wkład B. Ruszczaka: opracowanie zbioru danych, przygotowanie i oprogramowanie metod i modeli oraz wizualizacji danych, wykonanie i ewaluacja eksperymentów, propozycja i imlementacja metody automatycznego preprocesingu z użyciem filtrowania Savitzkiego-Golaya, autorstwo wstępnego artykułu oraz współautorstwo końcowej wersji artykułu, formalnej analizy oraz przeglądu literatury, zaprezentowanie pracy na konferencji międzynarodowej.

[A6] **B. Ruszczak**, S. Kubiciel-Lodzińska, K. Rudnik, Z. Bougriche, A. M. Wijata, i J. Nalepa, "Single- and multi- objective genetic optimization of hyperspectral image pre-processing," w *Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Malaga, Hiszpania.* ACM, 2025, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A; udział: 60%**]
Indywidualny wkład B. Ruszczaka: opracowanie metody preprocesingu obrazów hiperspektralnych z użyciem algorytmów genetycznych dla jedno i wielokryterialnej optymalizacji, opracowanie procedury ewaluacji algorytmów, wykonanie i ocena eksperymentów, wizualizacja wyników, przygotowanie, publikacja i utrzymanie zbioru danych, implementacja algorytmów uczenia nadzorowanego dla zadania regresji zawartości chlorofilu, prezentacja artykułu na konferencji międzynarodowej, współautorstwo i edycja artykułu.

[A7] **B. Ruszczak**, K. Kotowski, J. Andrzejewski, C. Haskamp, i J. Nalepa, "OXI: an online tool for visualization and annotation of satellite time series data," *SoftwareX*, tom 23, p. 101476, 2023, ✉ [**MNiSW: 200 pkt, IF 2023: 2,400, cytowania Scopus: 5, WoS: 3; udział: 30%**]
Indywidualny wkład B. Ruszczaka: opracowanie testów metody, współautorstwo założeń metody, implementacji, wizualizacji i finalnej wersji artykułu.

[A8] **B. Ruszczak**, K. Kotowski, J. Andrzejewski, A. Musial, D. Evans, V. Zelenevskiy, S. Bammens, R. Laurinovics, i J. Nalepa, "Machine learning detects anomalies in OPS-SAT telemetry," w *23rd International Conference on Computational Science - ICCS, Praga, Czechy*, ser. LNCS, J. Mikyska, C. de Mulatier, M. Paszynski *et al.*, Red., tom 14073. Springer, 2023, s. 295–306, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A, cytowania Scopus: 9; udział: 50%**]
Indywidualny wkład B. Ruszczaka: implementacja nadzorowanych detektorów anomalii,

zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych, opracowanie metody generowania pośrednich syntetycznych cech danych i doboru cech modelu detektora, prezentacja pracy na konferencji międzynarodowej, współautorstwo artykułu.

[A9] **B. Ruszczak**, J. Nalepa, K. Kotowski, i D. Evans, "The OPS-SAT benchmark for detecting anomalies in satellite telemetry," *Scientific Data*, tom 12, no. 1, s. 1–14, 2025, ✉ [**MNiSW: 140 pkt, IF 2024: 6,900, cytowania Scopus: 3, WoS: 2; udział: 50%**] *Indywidualny wkład B. Ruszczaka: implementacja algorytmów nienadzorowanych i nadzorowanych detektorów anomalii, zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych i przygotowanie ich publikacji, przygotowanie wstępnej wersji artykułu i współautorstwo finalnej wersji artykułu, oraz zgłoszenie artykułu i odpowiedzi na recenzje.*

[A10] **B. Ruszczak**, "Coupling temporal convolutional networks with anomaly detection and CMA-ES for the electricity price forecasting," w *Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Melbourne, Australia*, X. Li i J. Handl, Red. ACM, 2024, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A, cytowania Scopus: 1; udział: 100%**].

Dodatkowo, w tym opracowaniu wymieniam swoje pozostałe publikacje [B1-B55] (podrozdział 2.5) oraz międzynarodowe i krajowe projekty badawcze, którymi kierowałem [P1-P3] lub byłem wykonawcą [P1-P12] (podrozdział 3.1). Szczególnie ważnymi dla mnie dodatkowymi osiągnięciami są przeniesione do domeny publicznej repozytoria danych, których jestem współautorem [D1-D5] (podrozdział 3.4). Przekazane są w nich weryfikowane z udziałem ekspertów branżowych zbiory danych oraz kod, dla znacznej części prezentowanych tutaj badań. Zbiory te pozwalają na reprodukcję raportowanych osiągnięć, ale także na proponowanie i ewaluację nowych algorytmów przez innych badaczy.

2.2 Opis osiągnięcia

Przeprowadzenie wiarygodnego eksperymentu ekstrakcji wiedzy z użyciem modeli ML wydaje się dość dobrze poznanym tematem. Istnieje wiele wysoko rozwiniętych architektur dla takich modeli ML, dla których rezultaty przedstawiane są z użyciem popularnych, ale niewielkich zbiorów danych, przygotowywanych w wyniku jedynie ograniczonych eksperymentów laboratoryjnych lub z użyciem danych symulowanych. Często, modele oparte na takich zbiorach, w odniesieniu do rzeczywistych zadań, okazują się niewystarczające. Niekiedy również wyniki tych modeli ML nie są wyznaczane w sposób pozwalający na powtórzenie eksperymentów i przygotowanie algorytmów pracujących na podobnym poziomie skuteczności dla danych rzeczywistych.

Dla prawidłowo przeprowadzonej procedury ekstrakcji wiedzy, wymagane jest zachowanie kilku koniecznych warunków. W tym opracowaniu, zapisane zostały w postaci postulatów i omówione przy okazji prezentacji poszczególnych osiągnięć habilitanta. Przedstawiono je z użyciem kilku kompleksowo przeanalizowanych rzeczywistych problemów badawczych. Były to takie zadania, dla których wymagane jest użycie wielu wymiarów danych wejściowych, oraz wykorzystanie informacji która często jest ukryta. Dane w takich zadaniach obarczone są naturalnie występującymi zakłóceniami i szumem, nierzadkimi podczas badania rzeczywistych zjawisk. Analiza takich danych wymaga też niekiedy stosowania wielu dodatkowych indywidualnie dopasowywanych technik wspomagających

oraz wyszukania nietrywialnych wzorców w danych i analizy współzależności pomiędzy badanymi wieloma wymiarami tych danych.

Problemy te, w poniżej omawianych pracach, występują w wielu dziedzinach nauki i przemysłu (teledetekcji, wykrywania anomalii, prognoz rynkowych). Choć nie są głównym tematem tego opracowania, wykorzystane są w nim do przedstawienia wyzwań związanych z implementacją modeli ML, oraz zaproponowanych dla nich rozwiązań.

W pracach [A1–A5] polem badawczym była eksploracja pomiarów spektrograficznych i obrazów hiperspektralnych wykorzystywanych do opracowania i ewaluacji modeli ML do zadań teledetekcyjnych. Do rozważań nad algorytmami pozyskano dane z trzech różnych projektów badawczych, w których habilitant również brał udział. Jednym z kluczowych zadań było zaplanowanie oraz późniejsze wykonanie i ewaluacja procesu akwizycji wielowymiarowych, rzeczywistych danych. Projekty te dotyczyły między innymi: (*i*) teledetekcji chorób roślin [A1, A2], [B19, B35], (*ii*) identyfikacji nowych odmian roślin [A3] (w pracach tych liczba cech w procesie szkolenia modeli ML wynosiła 2151), (*iii*) a także zdalnego określania zawartości chlorofilu [A4, A5] (dla których wielowymiarowy model danych wyjściowych opierał się na 150 cechach ekstrahowanych z poszczególnych pasm obrazów hiperspektralnych).

Dla tych tematów zidentyfikowano kilka niedostatecznie wypełnionych przestrzeni badawczych. W literaturze pojawiają się badania określające wstępne rozwiązania, np. dla detekcji chorób roślin w sposób zautomatyzowany. Ale często propozycje te dotyczą diagnostyki chorób w późnych stadiach, wtedy kiedy w oczywisty sposób możliwa jest obserwacja symptomów chorób na obrazach rejestrowanych w świetle widzialnym [1, 2]. Istotnym ich ograniczeniem jest też stosowanie jedynie prostych metod statystycznych, lub niewielkich zbiorów obserwacji, a przez to pomijanie w badaniach bardziej zaawansowanych wzorców w analizowanych zbiorach danych [3, 4]. Dodatkowo badania te charakteryzują się ograniczoną uniwersalnością. Opisywane w nich wyniki trudno powtórzyć a opisywane metody przenieść na zadania teledetekcyjne dotyczące innych gatunków roślin, czy ich odmian. Badania te prowadzone są dla konkretnych roślin [5–7] i często jedynie w warunkach laboratoryjnych [8, 9]. Przedstawiane opracowania limitowane są dla konkretnych gatunków lub odmian [10–12], bez różnicowych testów czy uniwersalnych rozwiązań i wniosków. Ograniczone jest przenoszenie przedstawianych metod na inne zagadnienia. Przykładowo ewaluowane są detektory chorób z użyciem konkretnych pasm spektralnych [12, 13], bez podpowiedzi jak można by podobnych metod użyć do detekcji odmian, stresu suszowego, czy innych zadań.

Dlatego w przedstawionych pracach [A1, A2] oraz w [B19, B35] zaproponowano kompleksowe podejście do procesu ekstrakcji wzorców z analizowanych zbiorów danych, wiarygodnego wnioskowania, oraz budowy modeli ML zdolnych do skutecznego i w miarę możliwości wczesnego diagnozowania zadanych chorób. Opisywane działania zestawiają wskazówki metodyczne w zakresie konstrukcji badania mającego na celu przygotowanie wiarygodnych eksperymentów pomiarowych i szkolenia modeli ML. Zapisano je aby pomóc uporządkować późniejsze szkolenie wiarygodnych i skutecznych modeli ML, oraz przygotowywanie oznaczeń danych uczących, które powinny być jednoznaczne i wiarygodne. Wielu badaczy zwraca uwagę na powszechną dostępność dużych zbiorów danych (np. wielokanałowych obrazów satelitarnych, czy innych obrazowań), dla których brak jest jednak zweryfikowanych oznaczeń.

A1

Bazą do eksperymentów przedstawionych w publikacji [A1] były pomiary spektrofotometryczne wraz z przydzielonymi do nich dwoma wersjami etykiet. W pracach [A1,A2] wykorzystane dane treningowe pochodziły z pomiarów spektrofotometrycznych liści dwóch różnych odmian pomidora (aby podwyższyć uniwersalność szkolonych modeli), infekowanych w izolowanych warunkach wybranymi patogenami (izolowanych aby zapewnić jednoznaczność oznaczeń klas pomiarów dotyczących chorób oraz zdrowych roślin). W publikacji [A1] badanym patogenem była alternarioza, która jest najpowszechniej występującą chorobą, z którą borykają się plantatorzy pomidorów. W drugim z wymienionych badań, obserwacje rozszerzono o kolejne 4 patogeny, a w obu badaniach występowała też kontrolna próba zdrowych roślin. Pomiary hiperspektralne wykonane zostały w projekcie NCBiR [P12], w którym habilitant brał udział, we współpracy z Uniwersytetem Przyrodniczym w Poznaniu, i z pomocą zespołu dr. hab. Włodzimierza Krzesińskiego.

Na potrzeby prowadzonego badania opracowano szczegółową procedurę obejmującą: separowany proces hodowania i zakażania roślin, ale także ich skanowania, każdorazowej kalibracji instrumentu pomiarowego z użyciem dedykowanego spektralonu, oraz wykorzystania doświetlających lamp o znanej charakterystyce. Kluczowym aspektem dla potrzeb późniejszych prac nad algorytmami w tym badaniu było wykonanie akwizycji danych w separowanych warunkach dla każdego z badanych patogenów, oraz dla niezakażonej próby osobno, aby rejestrowane później próbki można było jednoznacznie przyporządkować do badanych klas.

W obu publikacjach zasadniczym wyzwaniem było opracowanie modelu uczenia maszynowego, pozwalającego na wnioskowanie ze zbioru pomiarów hiperspektralnych o dużej rozdzielczości spektralnej. Jako że dodatkowym celem badania była identyfikacja użytecznych domenowo zakresów reflektancji, do akwizycji danych wykorzystano wysoko rozdzielczy radiospektrofotometr, rejestrujący reflektancję w zakresie od 350 do 2500 [nm] z rozdzielczością 1 [nm], co w prostym przełożeniu odpowiada 2151 różnym rejestrowanym pasmom spektralnym, a w przedstawianej pracy odpowiada 2151 cechom w procesie przygotowania modelu. Tak liczny zbiór cech może być kłopotliwy w procesie szkolenia modeli ML, zwłaszcza w przypadku wysokiego wzajemnego skorelowania wartości reflektancji sąsiadujących ze sobą kanałów spektralnych. W pracy [A1], zdecydowano się na implementację modeli ML wspierających możliwość interpretacji kontrybucji cech wejściowych, dlatego wykorzystane zostały algorytmy oparte o drzewa decyzyjne [14] i lasy losowe [15]. Dodatkowo dla zastosowanych modeli lasów losowych uzyskano wysokie wyniki dokładności (0,987) oraz współczynnika F_1–score (0,986).

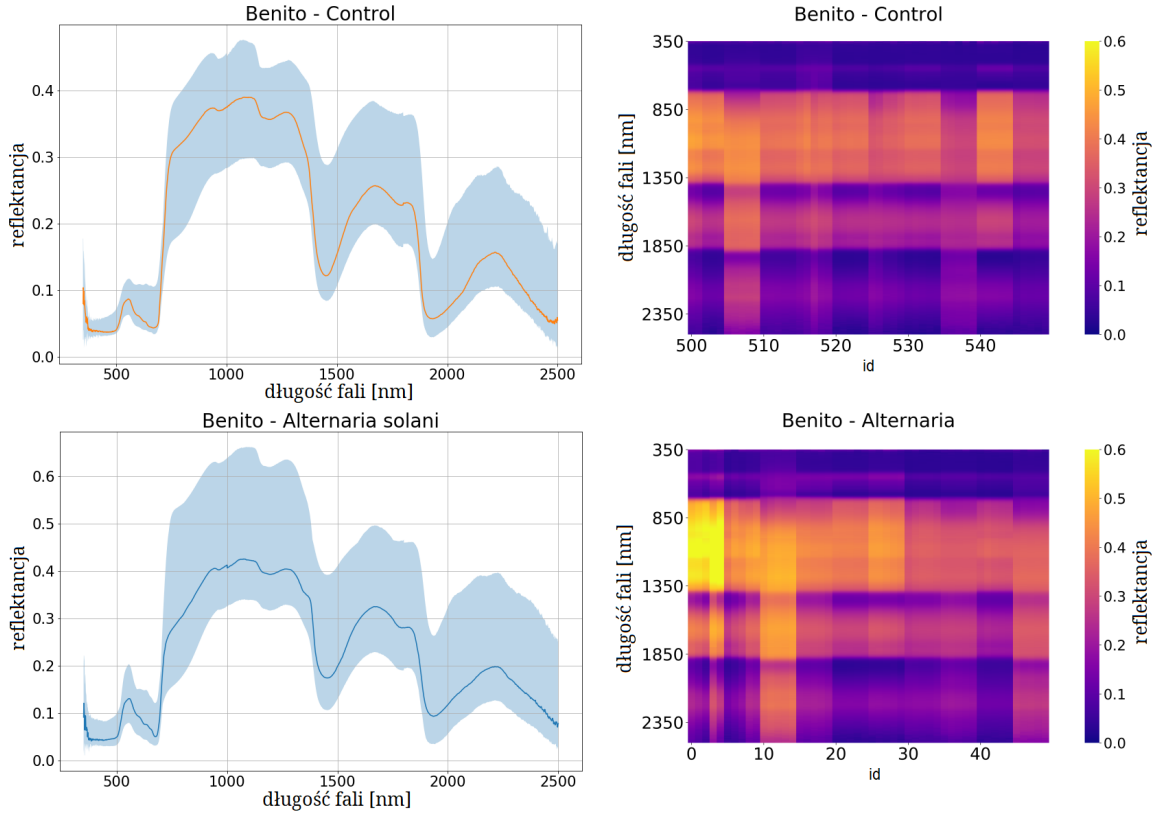
Dodatkową trudnością, której zniwelowanie pozwoliło uzyskać oczekiwany w badaniu model do interpretacji użyteczności poszczególnych pasm spektralnych, była filtracja wartości odstających i błędnych pomiarów. Zastosowane urządzenie (spektrofotometr radiometryczny) cechuje się pewną bezwładnością. Zgodnie z zalecaną przez producenta procedurą jego obsługi, w trakcie długich pomiarów urządzenie było wielokrotnie kalibrowane a każdy z pomiarów był powtarzany dziesięciokrotnie. Mimo to część z uzyskanych pomiarów należało odfiltrować po procesie akwizycji danych. Z uwagi na charakter prowadzonych rejestracji – skanowane były różne liście, w kolejnych fazach rozwojowych, wielu roślin pomidora, w różnych punktach liścia, krzywe przedstawiające tą samą klasę (np. zakażoną roślinę pomidora danej odmiany lub zdrową roślinę pomidora innej odmiany), nie mogły być filtrowane automatycznie. Powód był prozaiczny – krzywe spektralne nie wyglądają identycznie dla każdego pomiaru tej samej klasy a jednocześnie wyniki rejestrowane

są dla wielu wymiarów (tutaj 2151 różnych pasm spektralnych). Na rysunku 1, w lewej części przedstawiono histogram pomiarów obu klas, odpowiadający pomiarom zdrowych roślin (klasa: *Control*: »0«), oraz dla pomiarów liści chorych roślin (klasa: *Alternaria solani*: »1«). Dodatkowo rozróżniane są wartości wszystkich pomiarów (oznaczone jasnym niebieskim kolorem), oraz medianowe wartości pomiarów (oznaczonej ciemnoniebieską pogrubioną linią). Zaobserwować można na nim również duże zróżnicowanie pomiarów w poszczególnych pasmach spektralnych, który jest różny dla poszczególnych pasm rejestrowanego światła odbitego od roślin.

Do wstępnego przetwarzania wykonanych pomiarów zaproponowano w pracy wykorzystanie diagramów całkowitej reflektancji (ang. *absolute reflectance diagram*), a ich implementacja w takim kontekście jest dodatkowym osiągnięciem w prezentowanym zestawieniu. Diagramy te pozwalają na szczegółową analizę wykonanych pomiarów, i dopiero ich włączenie do procesu wstępnego przygotowania danych pozwoliło na eliminację części nieudanych pomiarów. Na rysunku 1 (po prawej stronie), zamieszczono wizualizacje wykonane z użyciem diagramów całkowitej reflektancji dla jednej z obserwowanych odmian, gdzie zauważyć warto ich zmienną charakterystykę spektralną, które wykonywano kontrolnie po procesie filtracji danych. Narzędzie to pozwala również skutecznie odróżnić wartości odstające, oraz błędy pomiarowe od poprawnych pomiarów. Na przedstawionych przykładach wizualizowano dwie próbki danych, a w obrazie tym wyróżniają się szerokie grupy pasów, odpowiadające pomiarom dla tej samej klasy, ale skanowane na materiale z innej rośliny lub w innym dniu. Dodatkowo powiedzieć można, że wzorzec danej klasy, w kolejnych rundach pomiarowych nie jest trywialnie identyfikowalny, a docelowa metoda klasyfikacji powinna umożliwić jego ekstrakcję. Diagramy absolutnej reflektancji są zatem zasadniczo odmienne od *map cieplnych* (ang. *heat map*). Wartości poszczególnych pól są inaczej wyznaczane – w przypadku map cieplnych wykorzystuje się funkcje częstości, i prezentuje się gęstość rozkładu występowania badanych zmiennych. Inaczej niż w przypadku diagramów absolutnej reflektancji, tutaj zintegrowanych do prezentacji spektrogramów, aby możliwe było badanie wielowymiarowych i wielokrotnych pomiarów w czytelnej formie.

Dla zdefiniowanego w badaniu problemu detekcji, i dla zbioru danych przygotowanego z użyciem omówionego narzędzia, wykonano ewaluację szeregu klasyfikatorów. Weryfikowano użyteczność algorytmów opartych o lasy losowe, wzmacnianie adaptacyjne (ang. *Adaptive Boosting* [16]), drzewa ekstremalnie losowe (ang. *Extremely Randomized Trees* [17]), maszyny wzmacniania gradientowego [18], oraz agregowane modele zespołowe z grupowaniem i głosowaniem większościowym. Finalnie uzyskany wysoko skuteczny model, oparty był o architekturę lasów losowych z dokładnością na poziomie 98% i F_1–score równym 0,939 (najwyższą dla stosunkowo niewielkiej puli 50 składowych modeli). Dla modelu tego wyselekcjonowano też najbardziej użyteczne pasma spektralne (905, 982, 1063, 525, 688, 1708, 1826 [nm] – uporządkowane według użyteczności cech badanego modelu, szerzej skomentowane w [A1]) oraz wykonano analizę podgrup analizowanego widma światła odbitego.

Pomyślne wyniki badania [A1], dla ewaluowanych modeli pozwoliły osiągnąć dokładność i współczynnik F_1–score na poziomie odpowiednio 0,987 i 0,986 dla diagnozowania alternariozy. Dodatkowe wnioski i interpretacje domenowe dla tych prac zawarto w [B35], zostały rozwinięte w dwóch kierunkach. Jednym z nich było zbadanie możliwości opracowania metody dla wczesnej detekcji patogenów. Ponadto zaproponowano objaśnienia poziomu wiarygodności takich detektorów oraz zweryfikowano skuteczność oznaczania stopnia wczesności rośliny. Ten wątek opisany został szerzej w pracy [B19], dla której modele



Rysunek 1: Diagramy całkowitej reflektancji (po prawej) i rozkładu pomiarów (po lewej) dla zbioru pomiarów wykorzystywanych w [A1]. Przedstawiono pomiary dla dwóch klas, dane pomiarowe w zakresie od 350 do 2500 [nm] z rozdzielczością 1 [nm]. Przedstawiono wizualizację dla dwóch klas ale jednej odmiany (Benito), w komentowanej pracy badano również inne odmiany.

ML ewaluowane były z użyciem rozszerzonego zbioru danych, uzupełnionego o pomiary na różnych etapach rozwoju roślin i w różnych momentach od inseminacji patogenów.

A2

Kolejny kierunek, w którym prowadzono badanie, opisany został w [A2] i miał dwa praktyczne cele. Pierwszym było umożliwienia detekcji większej liczby patogenów. W badaniu [A1], z użyciem binarnego klasyfikatora, potwierdzono jedynie skuteczną detekcję dla alternariozy. Drugie zadanie zakładało opracowanie użytecznych objaśnień dla opracowanych modeli. Takie objaśnienia powinny wspomagać określenie wymaganych do detekcji pasm spektralnych, tak aby docelowo detekcja mogła być wykonywana uproszczonym sensorem (zamiast spektrofotometru, którego zastosowanie na przemysłową skalę jest kosztowne i niepraktyczne).

Do osiągnięcia tego celu przygotowany został rozbudowany zbiór danych pomiarowych, który obejmował, oprócz kontrolnej klasy zdrowych roślin, 5 wskazanych przez ekspertów domenowych najważniejszych i często występujących patogenów. Oprócz alternariozy (ang. *Early Blight* – EB), przygotowany zbiór zawierał jeszcze: antraknozę (ang. *Anthracnose* – AN), bakteryjną cętkowatość (ang. *Bacterial Speck* – BS), zarazę ziemniaka (ang. *Late Blight* – LB), oraz septoriozę (ang. *Septoria Leaf Spot* – SL), dla których eksperci przygotowali również materiał roślinny do skanowania. Do pomiarów poszczególnych prób roślinnych wykonywanych z użyciem radiospektrofotometru przypo-

rządkowano klasy wskazujące poszczególne patogeny (AN, BS, EB, LB, SL).

Główne zidentyfikowane w pracy wyzwanie miało mocną praktyczną przyczynę. Wstępnie przygotowany wieloklasowy detektor, bazujący na pomiarach hiperspektralnych, cechował się niską skutecznością dla danych testowych. Dodatkowo nie umożliwiał on analizy relacji pomiędzy poszczególnymi pasmami spektralnymi i badanymi klasami patogenów.

Dlatego w [A2] zaproponowano inną metodę, polegającą na zbudowaniu komitetu kilku składowych modeli dyskryminujących poszczególne patogeny. Dla wydzielonych podzbiorów danych ewaluowano binarne klasyfikatory, aby dopiero potem wykonywać rozróżnianie patogenów. Kolejnym osiągnięciem opisanym w pracy jest również kilkuetapowa procedura budowania modelu. W niej, po wstępnym etapie przetwarzania danych, wykonywana jest sekwencyjna, iteracyjną redukcja cech modelu, by na finalnie wyłonionym zestawie 50 cech (zamiast początkowych 2151 cech) wyszkolić model dla konkretnej klasy i z dostrojonymi hiperparametrami.

Dla tak przygotowanych modeli wykonano pogłębioną analizę wybranych cech spektralnych, różnych dla badanych patogenów. Dla trzech z pięciu analizowanych chorób pomidora i dokumentowanych przez innych badaczy modeli ML. Wykonano porównawcze badanie, które ukazało szereg podobieństw dla analizowanych cech i różnych metod detekcji. Dodatkowo, z uwagi na bogatą bazę pomiarów, i zastosowanie spektrofotometru o wysokiej rozdzielczości i szerokim zakresie rejestrowanego pasma, pozwoliło ujawnić inne pasma spektralne, które są mocnymi predyktorami dla opracowanych modeli ML.

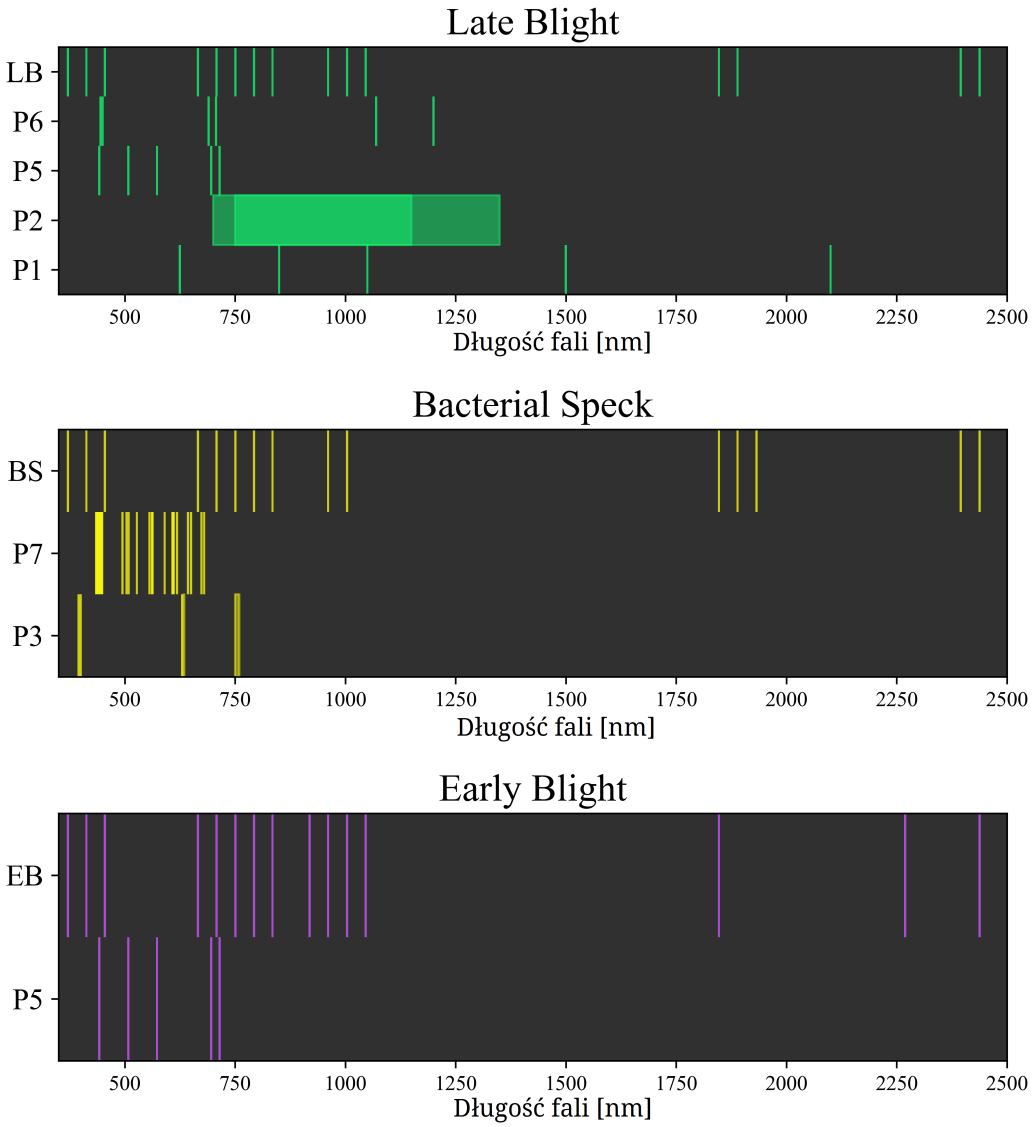
Efekty przeprowadzonej procedury wyselekcjonowania cech modelu klasyfikacyjnego, dla wybranych w badaniu patogenów przedstawiono na rysunku 2. Zestawiono tam odpowiadające cechom modelu pasma spektralne z wynikami innych, publikowanych wcześniej zakresów spektralnych.

A3

Osiągnięciem prezentowanym w pracy [A3] jest wieloetapowa metoda przygotowania modelu ML, ze wstępnych przetworzeniem danych spektralnych (filtrowania wejściowych krzywych spektralnych metodą Savitzkiego-Golaya [19]) z jednoczesną ewolucyjną redukcją liczby cech i krzyżową walidacją skuteczności pracy szkolonych modeli.

W pracach [A1, A2] przedstawiono metody detekcyjne, oparte na wysokorozdzielczych pomiarach spektralnych, dla których zidentyfikowano dodatkową trudność, która ujawnia się w sytuacji, gdy w zbiorze danych występuje niewiele pomiarów. Nieduża liczba próbek danych do szkolenia modelu ML (w praktyce pozyskiwanie wielu przykładów danych uczących bywa kłopotliwe i kosztowne) w zestawieniu z wielowymiarowym modelem danych (dla analizowanego przypadku było to 2151 kanałów spektralnych) może w znacznym stopniu utrudniać przeprowadzenie szkolenia modelu ML. Dodatkowo obserwowane silne skorelowanie sąsiednich kanałów spektralnych oraz pośredni charakter pomiaru (mierzone jest nie właściwe zjawisko, a jedynie rejestrowany jest prezentowany przez roślinę pochodny obraz zjawiska), może wymagać redukcji liczby wymiarów analizowanego zbioru danych (redukcja pomiarów). Taka redukcja może mieć dodatkowo cenny domenowo charakter, gdyż wynikowa selekcja pasm może prowadzić do uproszczenia późniejszej metody badawczej (gdyby udało się zidentyfikować wartościowe pasma spektralne konieczne do zaobserwowania badanego zjawiska).

Aby zbadać możliwość wykorzystania komentowanej metodyki nie tylko do detekcji chorób roślin, w arykule [A3] zaprezentowano próbę implementacji wcześniej omówionych



Rysunek 2: Porównanie pasm spektralnych sugerowanych w literaturze z wynikami badania przeprowadzonego dla wysoko rozdzielczych pomiarów spektralnych i analizy istotności cech przygotowanych modeli klasyfikacyjnych dla wybranych patogenów.

technik dla zadania rozpoznawania nowych odmian roślin. W pracy tej zestaw pomiarów hiperspektralnych wykonany został dla 50 różnych kultywarów (nowych odmian roślin). Akwizycja pomiarów wykonywana była na eksperymentalnych uprawach ziemniaka prowadzonych w IHAR PIB, dla których dane pozyskano przy okazji badania [P2]. Agronomiczna część eksperymentu, przygotowanie roślin oraz wykonanie referencyjnych pomiarów więdnięcia, zostały wykonane przez Panią dr hab. Dominikę Boguszewską-Mańkowską z IHAR PIB, natomiast akwizycja i przetwarzanie danych spektralnych przez habilitanta.

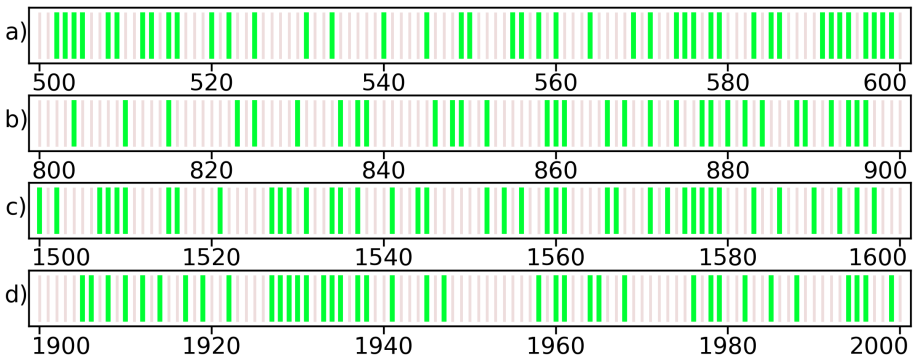
Dla 50 badanych roślin wykonane zostały pomiary spektrofotometryczne liści roślin dla każdej z tych odmian, a pod koniec sezonu zmierzona ich odporność na stres suszowy, który mierzony był w trzy stopniowej skali, $\Delta \in \{0, 1, 2\}$. Na podstawie przygotowanego zbioru danych opracowany został wstępny model klasyfikacyjny, który pozwala badane nowe odmiany roślin ziemniaka z użyciem pomiarów hiperspektralnych ich liści przydzielać odpowiednio do jednej z trzech klas badanej odporności Δ. Z perspektywy agronomicznej pozwala on na znaczące przyśpieszenie prowadzenia badań nowych roślin – nie jest wy-

magane oczekiwanie na rezultat do końca sezonu agronomicznego, przydział takiej rośliny do grupy odporności można wykonać wcześniej, gdyż pomiary wykonać można gdy tylko pojawią się liście tej rośliny wystarczające do wykonania pomiarów spektrofotometrem.

W pracy [A3] zaproponowano zbudowanie modelu ML w oparciu o modele liniowe oraz metodę lasów losowych, dla których można stosunkowo łatwo badać kontrybucję poszczególnych cech danych. Wstępnie przygotowane modele cechowały się niewysoką skutecznością, na poziomie F_1–score 0,615 dla lasów losowych, i 0,653 dla modelu liniowego z regularyzacją L_2. **Dopiero zaproponowane w tej pracy połączenie wstępnego filtrowania danych dla liniowego modelu ML, oraz zastosowanie ewolucyjnej redukcji cech pozwoliło na znaczące podniesienie poziomu skuteczności modelu.**

Dla przetwarzanego wielowymiarowego zbioru pomiarów zidentyfikowano najważniejsze ograniczenia: (*i*) niewielka liczba pomiarów (216 pomiarów różnych roślin; każda z nich była nowo opracowaną odmianą i wymagała wielotygodniowego przygotowania), (*ii*) wysoki stopień korelacji pomiarów dla sąsiednich pasm oraz (*iii*) średni poziom zaszumienia pomiarów, (*iv*) dużo wyższa od liczby pomiarów liczba cech modelu (odpowiadających każdemu z 2151 pasm spektralnych), oraz (*v*) mocno niezbalansowany rozkład badanych klas (108 pomiarów roślin z grupy $\Delta = 0$, 24 pomiary dla $\Delta = 1$ i 84 pomiary dla $\Delta = 2$).

Do rozwiązania tych problemów oraz uzyskania docelowej wyższej skuteczności przydziału pomiaru do odpowiedniej klasy, wykonano modelowanie z użyciem wspomnianego liniowego modelu z regularyzacją, po wstępnym przetworzeniu danych z użyciem filtrowania metodą Savitzkiego-Golaya [19] i równoczesną redukcją cach przekazywanych modelowi ML (główne osiągnięcie przedstawianej pracy). Wspomniane filtrowanie pozwoliło znacznie ograniczyć stopień zaszumienia pomiarów, a szerzej działanie tej techniki skomentowane jest w następnej sekcji (przy okazji badania [A5], którego zasadniczym efektem jest przygotowanie modeli ML wspomagane tego rodzaju filtrowaniem). Weryfikacja modeli prowadzona była z użyciem pięciokrotnej walidacji krzyżowej, podczas której wyznaczane były uśredniane wartości metryk klasyfikacyjnych.



Rysunek 3: Wyselekcjonowane cechy najskuteczniejszego modelu klasyfikacyjnego [A3] odpowiadające pasmom spektralnym z analizowanych pomiarów spektrograficznych; na zielono wyróżniono cechy modelu wybrane w finalnej ewolucji (μ, λ).

Do redukcji przestrzeni cech modelu (odpowiadającej odpowiednim rejestrowanym pasmom spektralnym) wykorzystano algorytmy ewolucyjne budowane z użyciem dwóch odmiennych strategii: $(\mu + \lambda)$ or (μ, λ) [20]. W metodzie tej ewoluowany był binarny wektor (osobnik), za pomocą którego zakodowano kolejne cechy modelu ML, odpowiadające poszczególnym kanałom spektralnym tak, że dla kanałów $K \in \{1, 2, 3, \ldots, 2151\}$, z zakresu od 350 do 2500 [nm], przyporządkowywano odpowiednio {0 lub 1}. Dla każdego

z ewoluowanych osobników szkolono i walidowano model ML z użyciem jedynie tych cech (odpowiednich pasm), dla których odpowiadająca im wartość wektora równa była 1. Dodatkowo dla modeli strojono jedynie parametr regularyzacji ($\alpha \in \{10^{-9}, \ldots, 10^{-1}\}$), natomiast dla prowadzonych eksperymentów ewolucyjnych modyfikowano warunki początkowe w celu sprawdzenia możliwego stopień redukcji liczby cech modelu, przy zachowaniu możliwie wysokich wyników klasyfikacji. Dlatego prowadzono kilka niezależnych rund ewolucji w obu scenariuszach, dla początkowej liczby wykorzystywanych cech N_F, z zastosowaniem bardzo agresywnej redukcji cech ($N_F < 25\%$), braku redukcji ($N_F \in 100\%$), oraz wariantów pośrednich ($N_F < 37,5\%$, $N_F < 50\%$, $N_F < 75\%$), które były potem ewoluowane.

Najwyższy wynik zanotowano dla modelu liniowego, $\alpha = 10^{-8}$, strategii (μ, λ), i początkowo mocno ograniczonej losowanej puli cech przedstawionych modelowi ($N_F < 37,5\%$). Dla tego modelu uzyskano finalną dokładność równą 0,867 i równie wysoki współczynnik F_1–score równy 0,864, przy wykorzystaniu jedynie 783 pasm (z puli 2151 dostępnych). Na rysunku 3 zilustrowano pasma wybrane dla opisywanego modelu.

Oprócz opracowania metody przygotowywania modeli ML dla przedstawionego zadania i omówionych optymalizacji procesu wnioskowania habilitant w projekcie [P2] odpowiedzialny był również za zaplanowanie badań i przygotowanie skutecznej aplikacji o finansowanie projektu (NCBiR) oraz późniejsze nadzorowanie prac, a także ich raportowanie i rozliczenie.

A4

W kolejnym artykule zaproponowano procedurę uczenia maszynowego do wiarygodnej ewaluacji regresyjnych modeli ML dla wielowymiarowych danych rzeczywistych [A4]. W przytoczonej pracy procedura ta została zdefiniowana z wykorzystaniem zbioru danych [D2] opracowanego przez habilitanta w kierowanym przez niego projekcie [P2], dla teledetekcyjnego zadania estymacji kilku parametrów na podstawie obrazowań hiperspektralnych.

Pierwsze wyzwanie, które zidentyfikowano w tej pracy, to brak otwartych, przygotowanych w transparentny sposób zbiorów danych i dedykowanych im procedur testowania modeli dla zadań teledetekcyjnych opartych na rzeczywistych danych wielowymiarowych do estymacji obserwowanych parametrów (zawartości chlorofilu w roślinach dla tego badania). Drugie wyzwanie zidentyfikowano po wstępnej analizie przygotowanego zbioru. Dotyczy ono konieczności translacji danych z postaci obrazów hiperspektralnych do przestrzeni cech gradientowych, aby możliwa była skuteczna analiza wzorców w tych danych z użyciem modeli ML z zachowaniem możliwości interpretacji kontrybucji poszczególnych wymiarów danych. Metoda takiej translacji powinna pozwalać objaśniać działanie modeli na postawie poszczególnych pasm obrazów hiperspektralnych.

Głównym osiągnięciem opisywanym w pracy jest zdefiniowana procedura testowa, który zestawia otwarty, udostępniony zbiór danych, wyselekcjonowany i zweryfikowany z pomocą ekspertów domenowych zbiór pomiarów referencyjnych, oraz reprodukowalną procedurę szkolenia i ewaluacji modeli dla zaproponowanych dla nich indywidualnych zestawów konfiguracji (możliwych ustawień hiperparametrów poszczególnych modeli ML).

W pracy przedstawiono też ewaluację dostarczonego zbioru danych i modeli ML zgodnie z tą procedurą. W tym celu, z dużo większej grupy mierzonych zjawisk, wybrano 4 parametry odpowiadające cechom liści, które zobrazowano na obrazach hiperspektralnych. Wybrane parametry: SPAD, FvFm, PI (związane z zawartością chlorofilu w zebra-

nych liściach), oraz RWC (odpowiadający zawartości wody w próbkach), odpowiadały laboratoryjnym pomiarom wykonanym po zlecanych w projekcie nalotach z kamerą hiperspektralną.

Dodatkowym osiągnięciem pracy jest przygotowany wspólnie ze specjalistami z dziedziny agronomii i udostępniony na otwartej licencji zbiór danych [D4] opisany w towarzyszącym mu artykule [B29]. Przekazano w nim wykonywane w rzeczywistych warunkach agronoomicznych obrazy. Zbiór zawiera zarówno całe ortofotomapy hiperspektralne w wysokiej rozdzielczości jak i wycięte fragmenty (kostki hiperspektralne), dla których przyporządkowane są poszczególne zestawy pomiarów laboratoryjnych (SPAD, FvFm, PI, RWC). Próbkowanie pikseli w tym zbiorze to 2.2 cm GSD (ang. *Ground Sampling Distance*), odpowiada 2,2 cm^2 na piksel hiperspektralny. W odróżnieniu od opisywanej kolekcji, dostępne do tej pory zbiory danych o podobnej tematyce, opierają się najczęściej na szacowanych, a nie mierzonych bezpośrednio wartościach referencyjnych. Dodatkowo obrazowanie dla wcześniejszych zbiorów danych wykonywane było innymi metodami (było to najczęściej obrazowanie satelitarne, o wielokrotnie niższej rozdzielczości przestrzennej i spektralnej). Dodatkowym wkładem publikacji jest szeroki przegląd metod detekcyjnych dla szacowania zawartości chlorofilu i wody w roślinach, oraz rodzajów stosowanych do ich monitorowania sensorów, wraz z wyszczególnieniem raportowanych tamże skuteczności. Z uwagi na to, że zbiór danych oparty został na rzeczywistych obrazowaniach w wysokiej rozdzielczości przestrzennej, i dokładnych, laboratoryjnych referencyjnych oznaczeniach modelowanych parametrów, może być on wykorzystany jako referencja, do weryfikacji lub kalibracji innych metod teledetekcyjnych pozwalających na szacowanie takich parametrów.

W pracy zaproponowano też metodę translacji danych treningowych modeli, z postaci obrazów hiperspektralnych, do postaci krzywych spektralnych, z użyciem przestrzeni cech gradientowych. Taka translacja, umożliwia wykorzystanie do budowania modeli regresyjnych klasycznych modeli ML, które w dalszym etapie prac, bardziej nadają się do objaśnienia domenowych zadań. Algorytmy, takie jak konwolucyjne sieci neuronowe, które mogą być szkolone na całych obrazach hiperspektralnych, ale ich rezultatem są bardzo złożone zestawy warstw i wag sieci, często są przez to dość trudne w interpretacji. W odróżnieniu od nich, algorytmy ML pracujące na cechach gradientowych mogą być wykorzystane do badania kontrybucji poszczególnych cech (w tym zadaniu odpowiadającym poszczególnym pasmom spektralnym) do finalnego modelu.

Wyniki przeprowadzonej analizy dla kompletnej procedury ewaluacji modeli z użyciem przekształconych danych spektralnych i zestawu kilkunastu architektur modeli ML zostały przedstawione w pracy. Procedura zawiera również wytyczne odnośnie zidentyfikowanych konfiguracji niezbędnych do szkolenia zaproponowanych modeli ML. Informacje te zostały uzupełnione o analizę wrażliwości badanych modeli ML na ich najważniejsze parametry (stopień regularyzacji α dla modeli liniowych oraz γ i C dla parametrów regularyzacji i funkcji jądra dla maszyn wektorów nośnych).

Przedstawiona procedura może pozwolić na rozwijanie teledetekcyjnych, inaczej niż w używanych w agronomii inwazyjnych ale nieefektywnych do wielokrotnego użycia na dużą skalę metod pomiaru chlorofilu w liściach roślin. Praca dostarcza otwarty zbiór i procedurę, wspierają rozwijanie wysokowydajnych i weryfikowalnych algorytmów, w odróżnieniu od podobnych komercyjnych ale trudnych do badania rozwiązań. Choć obecnie obserwuje się dynamiczny rozwój aplikacji opartych o uczenie głębokie, w pracy zaprezentowano alternatywę, w postaci wydajnych i niewymagających dużych zasobów (zapo-

trzebowania na moc obliczeniową czy pamięć operacyjną). Co jest istotne w szczególności w przypadku aplikacji przeznaczonych na urządzenia brzegowe, a aplikacje teledetekcyjne w warunkach produkcyjnych używane są choćby na pokładzie satelitów, gdzie zasoby obliczeniowe są mocno ograniczone.

A5

W wyniku badania [A4] zidentyfikowano dodatkową trudność techniczną, wynikającą z charakteru prowadzonych akwizycji. Dane obrazowe rejestrowane poza laboratorium, a takie zawiera zbiór [D2], charakteryzują się dużo wyższą zmiennością i zawierają różnego rodzaju utrudnienia. W raportowanym przez habilitanta projekcie [P2], obloty w celu akwizycji danych obrazowych wykonywano w rzeczywistych warunkach, a zatem bez kontroli oświetlenia (przy zastanym promieniowaniu słonecznym i zmiennym zachmurzeniu), za to z kamerą hiperspektralną o dużej rozdzielczości spektralnej (150 rejestrowanych równocześnie kanałów z zakresu 460–902 nm), na niedużej wysokości nalotów (80 metrów). Pozwoliło to na uzyskanie wysoko rozdzielczych obrazów o rozdzielczości przestrzennej 2,2 cm GSD. W badaniu zaplanowano i skrupulatnie prowadzono procedurę oblotów i akwizycji obrazów, z użyciem paneli korekcyjnych w polu widzenia kamery, każdorazowej kalibracji kamery, selekcjonowania pór nalotów z uwzględnieniem natężenia promieniowania słonecznego w zadanych oczekiwanych zakresach oraz korekt nanoszonych na obrazy w procesie ich przetwarzania po oblotach (korekt automatycznie wyznaczanych z użyciem paneli kalibracyjnych w polu widzenia kamery).

Mimo to, dla skrupulatnie przygotowywanych obrazów hiperspektralnych zidentyfikowano istotny poziom zaszumienia poszczególnych pasm spektralnych i zmienną ich charakterystykę w kolejnych dniach obrotów, co było głównym wyzwaniem badawczym w artykule [A5]. Kluczowym osiągnięciem opisywanym w tej publikacji jest metoda filtracji rzeczywistych danych spektralnych, opracowana właśnie w celu zmniejszenia poziomu zaszumienia takich danych.

W wielu pracach problem ten występuje w niewielkim natężeniu. Zwłaszcza gdy badacze korzystają z obrazów rejestrowanych w kontrolowanych warunkach laboratoryjnych, przy stałym oświetleniu. Również gdy korzystają ze standaryzowanych obrazów satelitarnych, dla których wykonywane jest wieloetapowe wstępne przetwarzania i ich kalibracja. Często do takiego wstępnego przetwarzania obrazów hiperspektralnych wykorzystuje się klasyczne wygładzanie krzywej spektralnej (np. z użyciem średniej kroczącej), lub różniczkowanie analizowanych krzywych spektralnych (do modelu przekazywane są dane w postaci pierwszej czy drugiej pochodnej wejściowego sygnału spektralnego).

Jednakże na podstawie rzeczywistych danych hiperspektralnych (tych samych które przedstawiono w pracy [A4] wykazano, że metoda dedykowanej, dostrojonej filtracji tych danych, pozwala wyszkolić modele ML o zdecydowanie wyższej skuteczności. Efektem tej pracy jest metoda wstępnego przetwarzania ekstrachowanych z rzeczywistych, zaszumionych obrazów hiperspektralnych, krzywych spektralnych, oparta o filtrowanie Savitzky-Golay [19]. W komentowanym badaniu, dla sygnału h analizowanych krzywych spektralnych, można zapisać przekształcenie z użyciem filtra w postaci:

$$y[\pi] = \sum_{m=-\lambda}^{\lambda} h[m] * [\pi - m] = \sum_{m=\pi-\lambda}^{\pi+\lambda} h[\pi - m] * [m]. \tag{1}$$

Takie filtrowanie metodą Savitzkiego-Golaya wykorzystywane jest w zadaniach in-

terpolacyjnych. Natomiast zastosowanie tego filtrowania przy jednoczesnej manipulacji parametrami filtra (π, λ, μ), jako metody wstępnego przetwarzania danych modelu ML pracującego na krzywych spektralnych (cechach gradientowych) pozwala znacznie ograniczyć poziom zaszumienia danych.

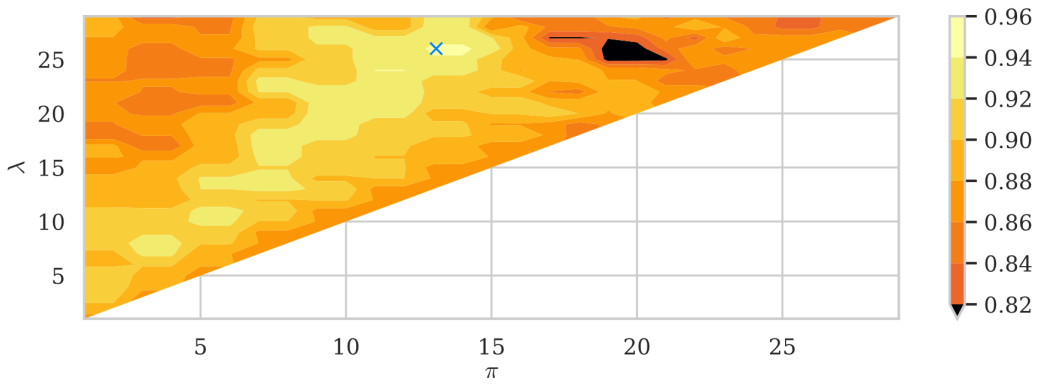
Dla analizowanych w [A4] 4 parametrów (SPAD, FvFm, PI i RWC) i obrazów hiperspektralnych wykazano, że włączenie dodatkowego kroku wstępnego przygotowania danych z użyciem strojonego dla tych danych filtra pozwala znacznie podnieść skuteczność modeli ML. Dla modelowanych w zbiorze 4 parametrów obniżono przeciętny błąd bezwzględny modelu regresyjnego o 53,5% dla SPAD, o 23,3% FvFm, o 30,9% dla PI i o 25,0% dla RWC. Aby uniknąć wpływu zaawansowanych mechanizmów uczenia głębokiego na przetwarzane obrazy, do badania przydatności filtrowania wykorzystano jedynie modele liniowe (linowe modele ML z regularyzacją lub maszyny wektorów nośnych z liniowym kernelem) oraz dane i procedurę testową z [A4] (przedstawione tam wyniki brano jako punkt odniesienia).

Dla zobrazowania procedury, proces strojenia filtra (π, λ, μ) dla jednego z modelowanych parametrów przedstawiono w postaci mapy rozwiązań (rysunek 4) wykreślono wyniki modelu (dla mierzonej wartości SPAD, i kolejnej iterowanej wartości $\mu = 2$) z różnymi ustawieniami filtra (modyfikowane były też: π i λ) z zaznaczeniem najwyższego uzyskanego współczynnika determinacji. Mapa taka odzwierciedla różne przeszukiwane rozwiązania modelu, które jak łatwo zauważyć, znacznie różnią się skutecznością dla alternatywnych konfiguracji filtra. Ale jeszcze ważniejszym faktem jest możliwość znacznego podniesienia mocy predykcyjnej rzeczywistych danych treningowych z użyciem takiej metody filtrowania.

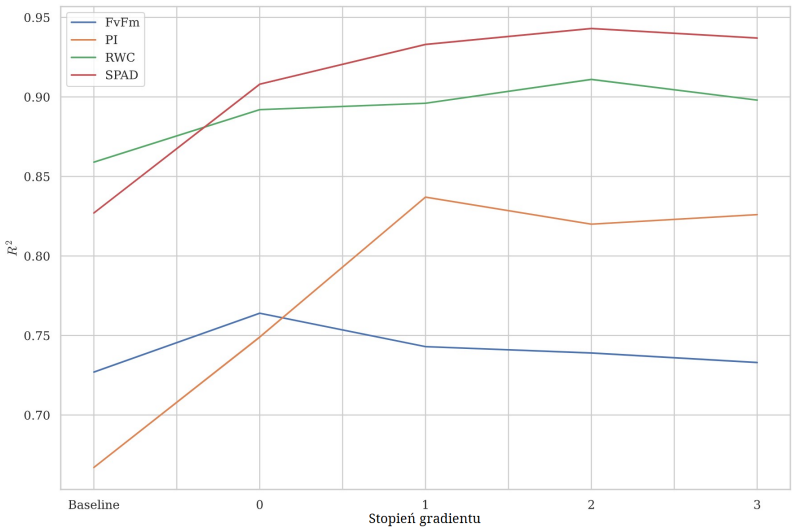
W każdym z komentowanych badań [A1–A5], gdzie analizowane były dane hiperspektralne (obrazowania hiperspektralne oraz pomiary spektrofotometryczne) zadbano także o zróżnicowany charakter pobieranego materiału roślinnego. W celu zapewnienia wyższej uniwersalności prowadzonych eksperymentów, dla każdego z badań dostarczano próbki roślin różnych odmian (charakteryzujące się, chociażby innym wyglądem i zabarwieniem liści), rejestrowane z różnych odległości, dotyczyły różnych gatunków (pomidora, ziemniaka), a pomiary prowadzone były w różnych fazach sezonu agronomicznego i na roślinach na innych etapach rozwoju, aby zapewnić zróżnicowanie po stronie badanych przyporządkowanych klas i umożliwić wydzielenie wiarygodnych podzbiorów danych i ich stratyfikację (względem odmian roślin, gleby na której były one hodowane czy stosowanych zabiegów irygacyjnych).

A6

Ostatnie dotyczące wielowymiarowych danych spektralnych przedstawione w tym referacie badanie [A6] składa wcześniej opracowane techniki w całościowy proces przygotowania wielowymiarowych danych spektralnych i krzyżowej ewaluacji modeli ML oraz proponuje równoczesne wykonanie optymalizacji złożoności obliczeniowej tego procesu. W publikacjach [A3, A5] wykazano, że wstępne przetwarzanie danych spektralnych (krzywych spektroskopowych czy obrazów hiperspektralnych) może istotnie poprawić skuteczność szkolonych modeli ML (ewaluowane z użyciem danych dla problemów teledetekcyjnych – wyższe odczytywane wartości wskaźnika F_1–score [A3] dla klasyfikacji odmian oraz niższe poziomy błędów dla modeli estymacyjnych w [A5] dla estymacji parametrów związanych z estymacją chlorofilu). W badaniach tych jednak do ustalenia korzystnej konfiguracji 3 parametrów badanej metody wstępnego przetwarzania tych danych wykonano



Rysunek 4: Mapa wyników wartości współczynnika determinacji R^2 dla różnych konfiguracji filtrowania. Niebieski marker × wskazuje na mapie wariant z najwyższym wynikiem dla $\mu = 2$ [A5], parametry: π, λ i μ są całkowitoliczbowe, dlatego przejścia pomiędzy poszczególnymi wynikami również są skokowe, przestrzenie pomiędzy poszczególnymi ustawieniami parametrów zostały zakolorowane jedynie, aby ułatwić odczytywanie mapy.



Rysunek 5: Poprawa skuteczności modelu dla poszczególnych estymowanych parametrów (SPAD, FvFm, PI i RWC) [A5], oraz różnych ustawień π, λ i μ.

wyczerpujące sprawdzenie wszystkich możliwych konfiguracji dla zadanej puli parametrów. Takie działanie pozwoliło wybrać najlepszą z puli konfigurację, jednak wiązało się z koniecznością szkolenia i ewaluacji modelu ML dla każdej z testowanych konfiguracji, co jest kosztownym i długotrwałym rozwiązaniem.

Osiągnięciem tej pracy jest zintegrowana ewolucyjna metoda przeszukiwania przestrzeni konfiguracji parametrów wstępnego przetwarzania danych spektralnych z użyciem filtrowania metodą Savitzkiego-Golaya. W pierwszej części pracy [A6] wykazano użyteczność klasycznego algorytmu genetycznego (ang. *genetic algorithm* – GA) z mechanizmem elityzmu [21].

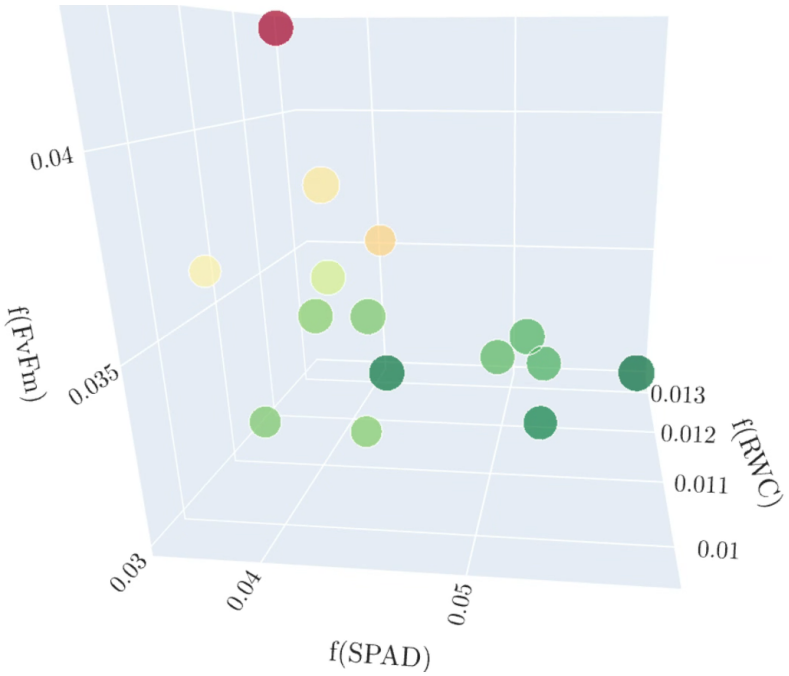
Aby pozwolić na wiarygodną ewaluację skuteczności pracy różnych konfiguracji wstępnego przetwarzania danych, w pracy wyszkolono liniowe modele ML. Zaimplementowano cztery alternatywne podejścia, zróżnicowane pod względem stosowanej metody regularyzacji (L_1, L_2, $L_1 + L_2$) oraz maszyny wektorów nośnych z liniową funkcją jądra. Podczas

krzyżowej ewaluacji modeli ML wyznaczano dla nich szereg metryk skuteczności, i dla szacowanych rzeczywistych wartości wyznaczano przeciętny bezwzględny błąd procentowy (MAPE), przeciętny błąd bezwzględny (MAE), i przeciętny błąd kwadratowy (MSE).

Dodatkowo, aby ułatwić porównywanie wyników optymalizowanych modeli ze skutecznością modeli zaraportowanych w pracy [A4] zaproponowana została relatywna metryka rMAE. Metryka ta pozwala wygodnie śledzić relatywny dystans wyniku optymalizowanego modelu wobec modelu bazowego. Zestawia ona wyniki z tego badania, dla każdej *i-tej* ewaluowanej konfiguracji znalezionej z użyciem algorytmu ewolucyjnego dla którego zmierzony błąd wyniósł $\mathrm{MAE}_{GA,i}$ z wynikami z pracy [A4] $\mathrm{MAE}_{B,i}$. Finalnie współczynnik ten jest liczony z użyciem formuły:

$$rMAE_i = MAE_{GA,i} \,/\, MAE_{B,i}$$

W odróżnieniu od pierwszych trzech statystyk, które na różne sposoby mierzą błąd szkolonych modeli ML, metryka rMAE pozwala zweryfikować czy rezultat procesu optymalizacji konfiguracji modelu zakończył się pełnym sukcesem (wtedy wartość tej metryki wynosi 1), lub wybrać modele dla których optymalna konfiguracja nie została znalezione (rMAE$<$ 1).



Rysunek 6: Wizualizacja wyników. Wykres przedstawia zawężoną do trzech wymiarów analizę znalezionych konfiguracji dla modeli ML dla estymacji $SPAD$, RWC, $FvFm$ [A6]

Na jedną sprawdzaną konfigurację przedstawianej metody składają się parametry omawianego wcześniej filtrowania: rozmiar okna filtra (λ) i stopień wielomianu (π), oraz krotność wykonywanej operacji gradientowej (μ) (co odczytywać można jako stopień pochodnej przekształcanych danych spektralnych; dlatego dla $\mu = 0$ brane są oryginalne dane, bez filtrowania. W obu wariantach optymalizacji, jedno i wielokryterialnej wykorzystano całkowite kodowanie tych parametrów i sprawdzano następujące ich zakresy: $\mu \in (0 \ldots 5), \lambda \in (0 \ldots 31)$, and $\pi \in (1 \ldots 30)$ (aby zachować pełną porównywalność z wynikami przedstawionymi w [A5]).

Do zaimplementowania procesu ewolucyjnego wykorzystano populację $N = 100$ osobników którą którą ewoluowano do pułapu kolejnych generacji $G_{\max} = 500$ (proces przerywano jeśli funkcja przystosowania nie poprawiała się w kolejnych 75 generacjach). Dla każdej generacji, pary *rodziców* do operacji krzyżowania wybierano z jednakowym prawdopodobieństwem – dwoje *potomków* generowano, tak aby każdy z nich dziedziczył wartości sprawdzanych parametrów (μ lub λ lub π) z jednego z rozwiązań *rodziców*. Następnie każdy z generowanych potomków poddawany był operacji *mutowania* z prawdopodobieństwem $\mathcal{P}_m$ poprzez losowe zmiany sprawdzanych parametrów (parametry wymieniano na losowe wartości z zakładanych w badaniu zakresów). Kolejna generacja rozwiązań zawierała N rozwiązań dla których dodatkowo stosowano mechanizm elityzmu, i do kolejnej rundy promowano jedynie 5% najlepszych *osobników* z poprzedniej generacji.

Jakość weryfikowanych parametrów określano na podstawie funkcji przystosowania. Jej wartość sprawdzana była dla każdego zestawu parametrów po wyznaczeniu nadzorowanego modelu ML i ustaleniu jego skuteczności w procesie dwukrotnej krzyżowej walidacji. Celem było przeprowadzenie weryfikacji modelu za każdym razem z użyciem danych na które model nie był eksponowany w procesie szkolenia, aby umożliwić kwantyfikację zdolności generalizacyjnych całego procesu.

W drugiej części tego samego badania wykazano, że można rozszerzyć tę metodę w kierunku wielokryterialnej optymalizacji tego zadania. Dla takiego problemu zaimplementowano algorytm genetyczny z mechanizmem nie zdominowanego sortowania (ang. *non-dominated sorting GA* – NSGA-II) [22]. W przypadku problemów dla których występować mogą różne optymalne konfiguracje zaproponowanej metody przetwarzania danych, dla różnych rozwiązywanych zadań, zaproponowano wyszukiwanie puli rozwiązań (rozwiązań optymalnych w zrozumieniu Pareto, pozwalających uzyskać najwyższe wyniki dla poszczególnych kryteriów) dla każdego z zadań z użyciem metody NGSA-II. Wprowadzenie wielokryterialnego podejścia pozwala uniezależnić proces budowy modelu ML od konkretnego rozwiązywanego zadania i żmudnego przeszukiwana wariantów konfiguracji parametrów filtrowania. Rozwiązanie tego zadania wymagało równoczesnego wyznaczania czterech różnych funkcji przydatności dla każdego z *osobników* podczas ich ewoluowania (osobno dla każdego z mierzonych parametrów teledetekcyjnych: $SPAD, RWC, FvFm, PI$). Do tego zadania w metodzie NSGA-II, zaimplementowany został adaptacyjny operator mutacji, jak sugerowane jest w [23], a prawdopodobieństwo mutacji ustalono na $\mathcal{P}_m = 0.2$. Na rysunku 6 przedstawiono wizualizację wyników wielokryterialnej analizy. Przedstawiony na nim został zestaw znalezionych konfiguracji optymalnych w sensie Pareto (ang. *Pareto front*).

Wymiernym osiągnięciem w tym przypadku jest wykazanie, że metoda ta może w znacznym stopniu ograniczyć złożoność obliczeniową całego procesu przetwarzania danych, poszukiwania optymalnej konfiguracji i szkolenia modelu ML dla wielowymiarowych danych spektralnych. Wykazana w badaniu korzyść z takiego rozwiązania to redukcja przeszukiwań przestrzeni parametrów konfiguracyjnych z prawie dwóch i pół tysięcy do osiemdziesięciu (ewaluowano 2480 konfiguracji dla każdej kombinacji jednego z 4 różnych modeli ML i 4 różnych kryteriów). Dla wariantu badania w którym wykorzystano 3 niezależne kryteria pula wynikowych rozwiązań została ograniczona do jedynie 15 konfiguracji. W przypadku łącznego czasu przetwarzania danych i szkolenia modeli ML dla poszczególnych konfiguracji czas ten ograniczono między 91,5 a 96,2% przy zachowaniu skuteczności szkolonych modeli ML (dla większości z 16 ewaluowanych wariantów model-parametr wynikowe rMAE wynosiło 1).

A7

W trzech kolejnych artykułach zaproponowano zastosowanie narzędzi ML do ekstrakcji wiedzy z wielowymiarowych danych rejestrowanych w postaci szeregów czasowych. Dane do prezentowanych prac pozyskiwano w postaci zapisów telemetrycznych z rzeczywistych urządzeń, dla których zidentyfikowano ważne wyzwanie natury praktycznej. Do wyszkolenia i ewaluacji wielu algorytmów ML dla tego rodzaju danych wymagane jest posiadanie rozległych zbiorów danych i zestawów etykiet. W najkorzystniejszym wariancie etykiety powinny być wykonywane (lub przynajmniej weryfikowane) przez ekspertów domenowych, potrafiących poprawnie ocenić odzwierciedlane w danych sytuacje. Dlatego częstą potrzebą w rozwijaniu aplikacji ML jest efektywna współpraca z takimi ekspertami, najlepiej w trakcie wspólnej pracy na tych samych danych. W projekcie [P7], którego dodatkowym efektem jest praca [A7], w konsultacji z inżynierami odpowiedzialnymi za utrzymanie ruchu satelitów w Europejskiej Agencji Kosmicznej opracowywano zbiór danych dla danych telemetrycznych z trzech kilku- do kilkunasto-letnich misji kosmicznych.

We wstępnej fazie projektu zauważono brak narzędzia informatycznego do pracy z dużymi zbiorami danych telemetrycznych, dla których możliwe byłyby: swobodna i płynna eksploracja danych, wykonywanie zaznaczeń, praca z kilkoma szeregami danych (dane z misji kosmicznych to często wielowymiarowe struktury składające się nawet z kilkuset sygnałów). Dla kilkuletniego szeregu sygnału próbkowanego co sekundę (lub co kilka sekund) niemożliwe była efektywna praca i wykorzystanie istniejących narzędzi, w których nie było możliwości wyświetlenia tak licznego zbioru danych ani płynnej z nimi pracy.

Osiągnięciem opisanym w tej pracy jest zaproponowane rozwiązanie tego problemu – od projektu, definicji wymagań funkcjonalnych – po implementację i testy tego rozwiązania. Wynikowe narzędzie, jako aplikacja na otwartej licencji przekazana do publicznego użytku i jest wykorzystywana przez społeczność do pracy nad innymi zbiorami danych. W prowadzonych testach potwierdzono pracę aplikacji OXI (która dostępna jest na licencji MIT pod adresem `https://oxi.kplabs.pl`) ze zbiorami wielowymiarowych telemetrii, zawierających nawet 2 miliony punktów. Taki zbiór może być płynnie renderowany na ekranie przeglądarki. Zastosowano optymalizacje renderowania i składowania struktury danych w pamięci. Podniesiono także użyteczność pracy z aplikacją dla użytkownika (ang. *web usability*), zaimplementowane i przetestowane zostały wygodne dla użytkowników przybory do wykonywania zaznaczeń, a także zestaw ikon i skrótów klawiszowych ułatwiające żmudny proces etykietowania zbiorów. Od użytkownika opracowanego narzędzie nie są wymagane zaawansowane umiejętności, tak aby mógł skupić się na analizie wyświetlanych danych, a po skończonej pracy mógł sprawnie i bez opóźnień zapisać wykonane zaznaczenia.

Przestawiona metoda inspekcji danych pozwala na pracę z wielowymiarowymi szeregami, które dodatkowo mogą być próbkowane ze zmienną częstotliwością. W metodzie optymalizowana była liczba wyświetlanych na ekranie punktów, tak aby możliwe było przesuwanie widoku inspekcji oraz jego swobodne przybliżanie i oddalanie. Możliwa jest praca z jednym bądź z wieloma szeregami naraz. Zaimplementowano również różne narzędzia wynonywania oznaczeń anomalii (1D – pracujące jednowymiarowo w wymiarze czasowym i 2D – pozwalające na dwuwymiarowe zaznaczanie według czasu i wartości). Metoda ta pozwala również na tryb pracy wieloetykietowej, aby umożliwić definiowanie zbiorów danych, dla których identyfikowane są różne zdarzenia czy anomalie różnych typów.

W artykule [A7] nie zanotowano wniosków metodycznych istotnych z punktu widzenia usprawnienia algorytmów ML, i może być traktowany jako praca uzupełniająca, przedstawiająca kompleksowe podejście do kolejnych trzech przedstawianych badań. Jedną z najważniejszych trosk, w przypadku pracy z algorytmami ML, jest wymagana wysoka jakość przetwarzanych danych i dostarczanych oznaczeń. Przedstawiona metoda inspekcji danych była istotna zwłaszcza w kontekście kolejnych prac będących częścią opisywanego cyklu, a związane były one z detekcją anomalii w wielowymiarowych danych telemetrycznych [A8–A10] a także [B16, B22, B23, B26, B33], dla których narzędzie opisywane w tej pracy było esencjonalne w procesie przygotowania wykorzystywanych zbiorów danych.

A8

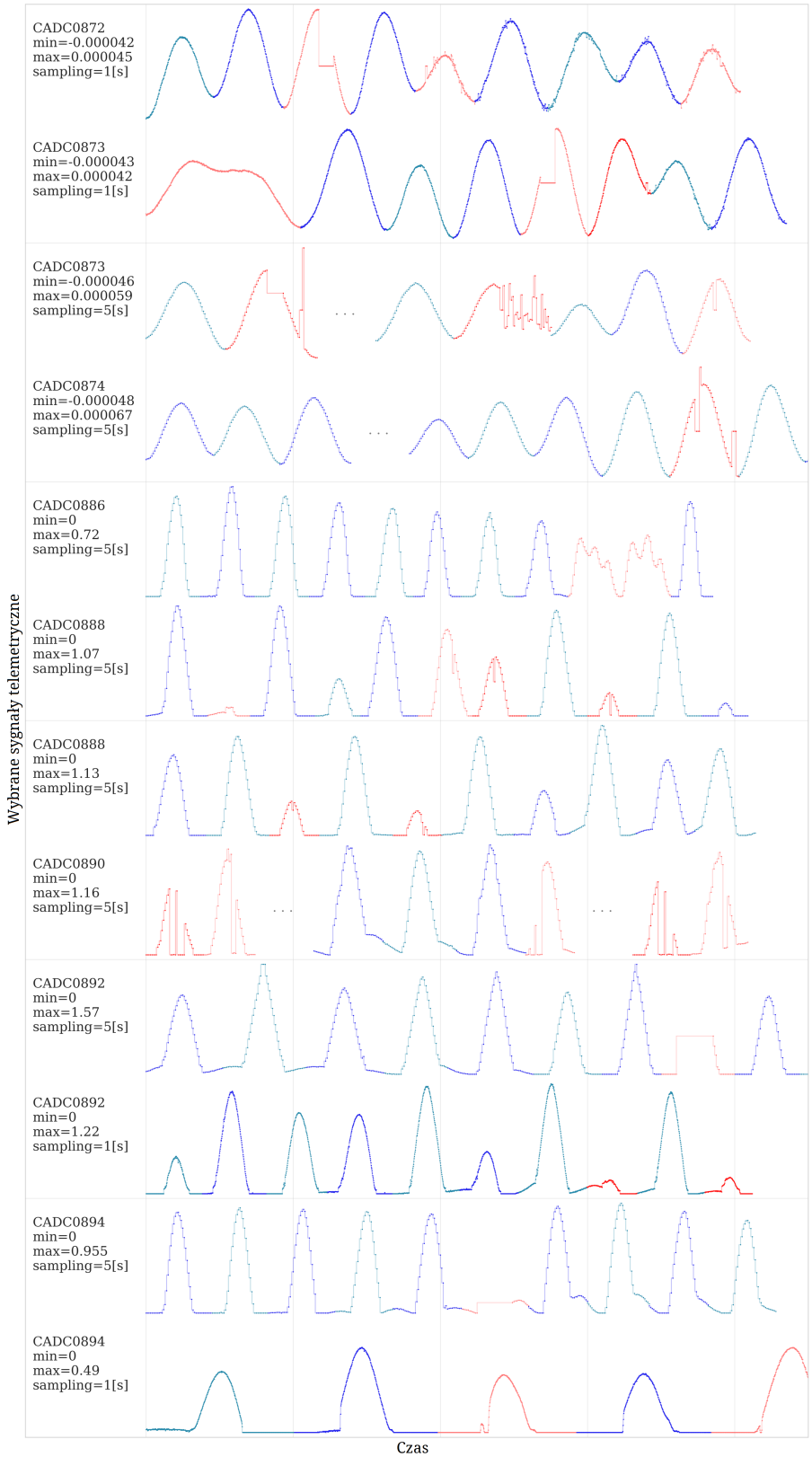
Kolejne osiągnięcie to algorytm do detekcji anomalii w rzeczywistych telemetrycznych danych satelitarnych ewaluowany z użyciem jedno i wielowymiarowych danych, z użyciem metody generowania pośrednich statystycznych cech analizowanych danych oraz metodami poszerzania zbioru danych (ang. *data augmentation*).

W artykule [A8] opisano opracowane rozwiązanie dla detekcji anomalii w rzeczywistych sygnałach telemetrycznych pozyskanych z satelity OPS-SAT należącego do Europejskiej Agencji Kosmicznej. Był to unikalny satelita (umieszczony w przestrzeni kosmicznej w 2019 roku a deorbitowany planowo w 2024), gdyż jego pokład inżynierowie Agencji udostępniali do wykonywania zgłaszanych im eksperymentów. Opracowany algorytm do detekcji anomalii przeszedł pomyślnie pełen proces walidacji i został dopuszczony do wykonania bezpośrednio na operującym satelicie.

Wiele współcześnie implementowanych algorytmów do detekcji anomalii wymaga długich szeregów czasowych, niezbędnych do wyszkolenia modelu ML, i przez to nie można ich zaimplementować do obsługi tego zbioru danych. Specyficzny charakter pracy urządzenia, miał swoje odzwierciedlenie w danych, które z niego pozyskano. Dane telemetryczne pochodzące z tego satelity są nieciągłe w wielu fragmentach. Eksperymenty wykonywane na pokładzie miały różną długość i występowały pomiędzy nimi wielokrotnie przerwy różnej długości (niektóre wielogodzinne). Mają też zmienny charakter, występują radykalne zmiany rozkładów danych dla poszczególnych sensorów (sensory były kalibrowane inaczej i ich odczyty były radykalnie różne dla kolejnych eksperymentów) oraz zmienia się interwał ich próbkowania dla rejestrowanych sygnałów. Przykłady sygnałów z analizowanego zbioru przedstawiono na rysunku 7. Skomplikowany charakter przygotowanej kolekcji danych telemetryczych i ich oznaczeń składa się na główne wyzwanie badawcze towarzyszące opracowaniu skutecznej metody detekcji dla tego zbioru.

Głównym osiągnięciem przedstawionym w publikacji [A8] była zaproponowana dwuetapowa metoda detekcji anomalii w rzeczywistych seriach czasowych danych satelitarnych. W pierwszym etapie generowano syntetyczne cechy dla krótkich segmentów danych telemetrycznych. Dopiero na podstawie tych syntetycznych cech, w drugim etapie, szkolony i ewaluowany jest model ML. Do budowy modeli detekcji anomalii w takim procesie można zastosować dowolny klasyfikator, a w opisywanym artykule przestawiono wyniki ewaluacji skuteczności pracy kilku nadzorowanych modeli ML. Najwyższą dokładność i współczynnik F_1–score, na poziomie 0,929 i 0,823, uzyskiwał model oparty o algorytm lasów losowych.

Zaproponowana w badaniu metoda generowania syntetycznych cech wykorzystuje 18 różnych parametrów. Są to przede wszystkim statystyczne wskaźniki liczone dla kolejnych segmentów, określających czas trwania, liczbę punktów, średnią wartość, wariancję i od-



Rysunek 7: Segmenty danych telemetrycznych satelity OPS-SAT, zawierające przykłady zakłóceń rejestrowanych sygnałów, w tym nieoczekiwane piki i spadki wartości sygnału, szum (`CADC0873`), nieregularności cykliczności (`CADC0886`), krótkie (`CADC0892`, `CADC0894`) i długie przerwy w danych (`CADC0874`). Segmenty oznaczone jako anomalie na wykresie zaznaczono kolorem czerwonym.

chylenie standardowe danego segmentu. Dodatkowo zliczana jest liczba niespodziewanych, znacznych odchyleń sygnału (ang. *peaks*). Zliczane są odstające wartości, o prominencji przynajmniej 10% dla danego segmentu danych. Odchylenia te, zliczane są w metodzie osobno dla oryginalnego sygnału, dla jego wygładzanej, interpolowanej wersji, oraz pierwszej i drugiej pochodnej tego samego segmentu danych. Następnie liczona jest wariancja segmentu, której wartość również wyznaczana jest dla pierwszej i drugiej pochodnej badanego segmentu. Do zbioru wprowadzana jest też liczba brakujących danych, która dodatkowo podnoszona jest do kwadratu, aby wzmocnić znaczenie tej cechy. W końcu zbiór uzupełniany jest o kilku ważonych cech. Długości i wariancja sygnału, ważone są względem próbkowania sygnału, aby pozwolić później szkolonym modelom rozróżniać wzorce dla sygnałów o różnym próbkowaniu w czasie.

Dodatkowym wkładem publikacji jest przygotowany reprezentatywny zbiór danych i procedura ewaluacji modeli z wydzielonym podzbiorem testowym. Zbiór ten pozwala odzwierciedlać rzeczywiste, praktyczne sytuacje, które występują w telemetrycznych sygnałach na urządzeniach satelitarnych, takich jak przerwy w sygnałach, braki danych, zaszumienie, czy radykalne zmiany sygnału (np. dryf danych, czy odzwierciedlenia zdarzeń albo wykonywanych na pokładzie planowanych, albo nieoczekiwanych manewrów). Do przygotowania tego zbioru danych [D1] wykorzystano rzeczywiste dane telemetryczne pochodzące z satelity OPS-SAT, udostępnione przez Europejską Agencję Kosmiczną w ramach projektu [P8], a dodatkowe weryfikacje zaznaczeń i konsultacje zbioru wykonywane były we współpracy z inżynierami ESA jeszcze w projekcie [P7]. Ważnym elementem projektu [P7] było uruchomienie wykonanych algorytmów bezpośrednio na pokładzie satelity OPS-SAT, w warunkach rzeczywistych, z zachowaniem restrykcyjnych procedur ESA i z uwzględnieniem ograniczeń sprzętowych (m.in. wymagany nieduży pobór mocy i nieduża zajętość pamięci do pracy algorytmu).

Wykorzystanie eksperymentalnie dobranych ekstraktorów cech, pozwalających na analizę sygnałów niezależnie od ich długości, próbkowania, przerw czy zakłóceń, pozwoliło na przygotowanie wzorcowych algorytmów detekcji anomalii o wysokiej dokładności 92,9%, wyznaczonej dla zbioru testowego, nieprzedstawianego modelowi w procesie jego szkolenia. Dodatkowo przygotowany model ML charakteryzował się zwinnym działaniem, niedużym stopniem skomplikowania, małymi wymaganiami w odniesieniu do czasu procesora i zajętości pamięci, i przez to mógł zostać zaakceptowany do pracy w ograniczonych zasobach komputera pokładowego satelity OPS-SAT.

W pracy [A8] wykazano również, że podwyższenie skuteczności modelu jest możliwe w procedurze przygotowania nadzorowanego modelu ML (ewaluowano architektury oparte na algorytmie lasów losowych, liniowe modele, maszyny wektorów nośnych z liniowymi kernelami, metodę k–najbliższych sąsiadów i wielowarstwowe perceptrony) z użyciem metod rozszerzania danych (ang. *augmentation techniques*). Dla wejściowych segmentów sygnałów telemetrycznych zastosowano metody zwielokrotnienia danych wejściowych poprzez dogenerowanie większej liczby przypadków. Segmenty zostały poddane odbiciom w osi OX, odbiciom w osi OY, oraz przesunięciom w osi OX z różnym marginesem przesunięć. Weryfikowano również równoczesne użycie wielu sposobów rozszerzania zbioru treningowego. Warto zaznaczyć że testową część zbioru pozostawiono bez zmian, oraz nie stosowano metod rozszerzania na segmentach ze zidentyfikowanymi anomaliami. Dla opracowanych technik i przeprowadzonej procedury ewaluacji uzyskano model, który charakteryzował się dokładnością 0,984, wyższą o 5,9% od najlepszego wcześniej uzyskanego wyniku.

W dalszej części tego podrozdziału przedstawiona jest dodatkowe badanie, wykonane jako uzupełniająca weryfikacja metody [A8] z użyciem wielowymiarowego zbioru sygnałów telemetrycznych ESA-ADB [D2]. W przedstawionym artykule [A8], opisano metodą detekcji anomalii w sygnałach telemetrycznych z użyciem wielowymiarowego zbioru cech syntetycznych. W przywołanej publikacji metoda ta była weryfikowana na zbiorze zawierającym 9 różnych kanałów telemetrycznych, dla których 18 cech wyznaczano dla kolejnych segmentów sygnału.

Na potrzeby tego autoreferatu wykonano dodatkową weryfikację tej metody w odniesieniu do wielowymiarowych sygnałów. Jej celem było sprawdzenie czy przedstawiona metoda pozwala na detekcję anomalii również w przypadku równoczesnej analizy wielu, obserwowanych dokładnie w tym samym czasie sygnałów.

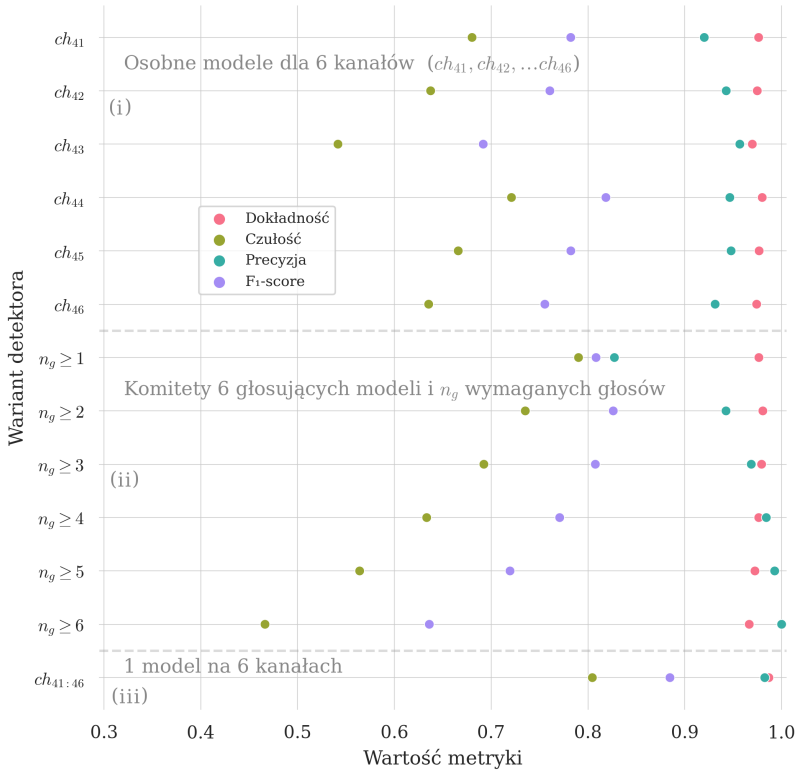
Na potrzeby tego zadania wykorzystano autorski wielowymiarowy zbiór danych ESA-ADB [D2] [B16]. Jest to zbiór danych telemetrycznych z kilku dużych, wieloletnich misji Europejskiej Agencji Kosmicznej, przygotowany w projekcie [P7]. Zbiór ten zawiera kilkadziesiąt równolegle rejestrowanych sygnałów telemetrycznych, dla akwizycji prowadzonych przez ponad 10 lat z interwałem sekundowym. Szczegółowe informacje na temat zbioru, w tym specyfikacja misji, sygnałów, opracowywania oznaczeń anomalii i wyniki dla wielu zbadanych detektorów przedstawiono w publikacji [B16]. Publikacja wskazuje również podzbiór danych do ewaluacji algorytmów, który zawiera wieloletnie dane z jednej misji i sześciu sygnałów telemetrycznych ($ch_{41}, \ldots, ch_{46}$) – i tego właśnie podzbioru użyto do weryfikacji zaprezentowanych tutaj eksperymentów.

W tym dodatkowym badaniu weryfikowano wcześniej przygotowany detektor [A8], pracujący z użyciem 18 pośrednich syntetycznych cech sygnału, jednak tym razem w odniesieniu do wielowymiarowych danych. Wykonano trzy niezależne rundy eksperymentu:

(i) – niezależne detektory dla każdego z testowanych kanałów telemetrycznych ch_i,
(ii) – model komitetowy, który agregował odpowiedzi składowych modeli, ewaluowanych na osobnych kanałach telemetrycznych,
(iii) – oraz z jednym modułem detektora, który operował na wszystkich kanałach telemetrycznych jednocześnie.

Warianty (*ii*) i (*iii*) można określić jak wczesną i późną fuzję odpowiedzi modelu ML. Dla detektorów budowanych w wariancie (*ii*) końcowa odpowiedź modelu ustalana była na zasadzie głosowania, z różnym wymaganym progiem akceptacji detekcji (n_g). W wariancie (*iii*) budowany model wykorzystywał symultanicznie dane z wszystkich kanałów. Szkolono go na 104 cechach wejściowych (z użyciem każdej z 18 syntetycznych cech wyznaczonych dla każdego z 6 rozpatrywanych sygnałów $ch_{41}, \ldots, ch_{46}$).

Weryfikacja wariantów wielowymiarowych (*ii*) i (*iii*) ujawniła wyższą ich skuteczność w odróżnieniu od detektorów pracujących niezależnie (*i*). W tabeli 1 zestawiono statystyki osiąganej skuteczności dla poszczególnych wariantów detektorów, na podstawie których zauważyć można zdecydowaną przewagę detektora pracującego w wariancie (*iii*), oraz dość wysoką skuteczność modelu komitetowego wymagającego zgodności decyzji przynajmniej dwóch detektorów ($n_g \geq 2$). Wyniki poszczególnych detektorów zilustrowano na rysunku 8. Wyraźnie odznacza się na nim dystans dzielący detektory pracujące indywidualnie w odniesieniu do wielowymiarowego podejścia. Można to interpretować jako dodatkowe potwierdzeniem możliwości wykorzystania metody opracowanej w [A8] do pracy z wielowymiarowymi satelitarnymi danymi telemetrycznymi.



Rysunek 8: Porównanie wyników dla modeli opartych o wielowymiarowy zbiór rzeczywistych sygnałów satelitarnych ESA-ADB [B16].

Tabela 1: Skuteczność detekcji w wariancie wielowymiarowym z użyciem zbioru ESA-ADB (uporządkowane według r_ϕ). Najwyższy wynik dla każdej z metryk zaznaczono **pogrubioną** czcionką.

Model	Dokładność	Precyzja	Czułość	F_1–score	r_ϕ
– RF dla sygnałów OPSSAT [A8]	0,9294	0,8826	0,8050	0,8233	0,7830
(*i*) Modele bazujące na jednym kanale telemetrycznym *ch* ze zbioru ESA-ADB [B16]					
– model dla ch_{41}	0,9763	0,9201	0,6802	0,7822	0,7798
– model dla ch_{42}	0,9749	0,9428	0,6375	0,7606	0,7639
– model dla ch_{43}	0,9697	**0,9568**	0,5418	0,6918	0,7076
– model dla ch_{44}	**0,9800**	0,9465	**0,7210**	**0,8185**	**0,8165**
– model dla ch_{45}	0,9768	0,9478	0,6660	0,7823	0,7838
– model dla ch_{46}	0,9742	0,9313	0,6354	0,7554	0,7575
(*ii*) Komitet 6 modeli na 6 różnych kanałach *ch* ze zbioru ESA-ADB [B16], zliczana liczba głosów: n_g (detekcji anomalii poszczególnych detektorów w komitecie)					
– przynajmniej 1 głos: $n_g \geq 1$	0,9765	0,8273	**0,7902**	0,8083	0,7961
– przynajmniej 2 głosy: $n_g \geq 2$	**0,9806**	0,9426	0,7352	**0,8261**	**0,8230**
– przynajmniej 3 głosy: $n_g \geq 3$	0,9793	0,9687	0,6925	0,8076	0,8095
– przynajmniej 4 głosy: $n_g \geq 4$	0,9764	0,9842	0,6334	0,7708	0,7795
– przynajmniej 5 głosów: $n_g \geq 5$	0,9724	0,9928	0,5642	0,7195	0,7375
– wszystkie głosy $n_g \geq 6$	0,9666	**1,0000**	0,4664	0,6361	0,6711
(*iii*) Model dla wielowymiarowego zbioru danych dla zbioru ESA-ADB [B16]					
– model dla 108 cech i $ch_{41} \ldots ch_{46}$	**0,9869**	**0,9826**	**0,8045**	**0,8847**	**0,8827**

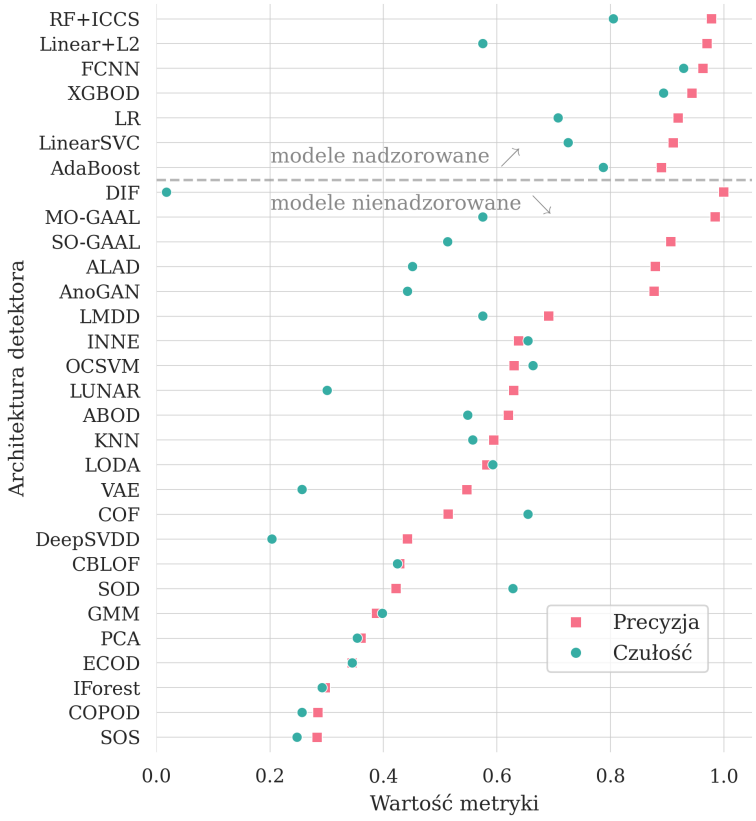
A9

Artykuł [A9] również eksploruje dziedzinę detekcji anomalii w telemetrycznych sygnałach satelitarnych i przynosi trzy istotne kontrybucje. Jego bazą jest ten sam zbiór danych [D1], który wykorzystywany był w publikacji [A8]. Opublikowanie tej publikacji zrodziło wiele zapytań o udostępnienie danych, dlatego po uzyskaniu niezbędnych zgód z ESA, zbiór ten został opublikowany na otwartej licencji, z intencją umożliwienia również innym badaczom wykorzystanie go do testowania ich algorytmów detekcyjnych.

Drugim ważnym efektem rozszerzonych prac było zaproponowanie szeregu nienadzorowanych algorytmów ML dla zadania detekcji anomalii, które weryfikowane były z użyciem zbioru OPSSAT-AD [D1]. Potrzeba zdefiniowana przez operatorów systemów satelitarnych wyraźnie określa wymagania dla algorytmu detekcyjnego. Powinien on adaptować się do nowych zdarzeń. Wstępnie przygotowany algorytm, zweryfikowany z użyciem rzeczywistego zbioru, powinien później pracować nadzoru lub z minimalnym wymaganym nadzorem. Dodatkowo praca ta niekiedy wykonywana będzie w wymagających warunkach (na urządzeniach brzegowych, z ograniczonymi zasobami, np. mniejszą dostępną mocą obliczeniową, mniejszą ilością pamięci czy ograniczonym chłodzeniem lub zasilaniem).

W artykule przedstawiono wyniki ewaluacji 30 algorytmów, a porównanie uwzględniało wcześniej opracowane nadzorowane architektury modeli ML, oraz ponad dwadzieścia algorytmów nienadzorowanych. W poniższym zestawieniu (rysunek 9) zaprezentowano wyniki poszczególnych modeli, a znaleźć można tam metryki wyznaczone zarówno dla modeli ML jak i modeli uczenia głębokiego. Dla wykorzystanego, zdefiniowanego w [A8] zbioru testowego, widać wyraźnie niższą skuteczność modeli nienadzorowanych, jednakże, biorąc pod uwagę całkiem inny proces budowania reprezentacji wzorców anomalii, które nie są w ich przypadku jawnie identyfikowane, część z modeli cechuje się imponującymi rezultatami.

Kolejnym ważnym efektem tego badania było opracowanie zestawienie otwartych problemów i określenie kierunków niezbędnych dalszych prac, dla których zidentyfikowano istotne wyzwania badawcze natury zarówno metodycznej jak i praktycznej. Dla wielu obecnie stosowanych algorytmów nie notuje się wyników na wystarczającym poziomie skuteczności detekcji anomalii, zwłaszcza dla danych rzeczywistych, które nie są przygotowywane do pracy w trybie nadzorowanym. Dlatego dwa istotne działania, które powinny być obecnie rozwijane to budowa algorytmów uczenia ciągłego (ang. *Continual Learning* [24]) oraz algorytmów posiadających zdolność identyfikacji wzorców, na małej liczbie próbek (ang. *Few-Shot Learning* [25]), również w danych wielowymiarowych. Aby dostarczyć rozwiązania tego zagadnienia prowadzone są obecnie przez habilitanta prace związane z adaptacją algorytmów DeepSAD (ang. *Deep Semi-Supervised Anomaly Detection* [26]) oraz *DevNet* (ang. *Deviation Networks* [27]), do skutecznej ekstrakcji wzorców anomalii, z wielowymiarowych danych satelitarnych, z użyciem cech zbioru [D1]. Ważnym zadaniem, które uzupełnia listę bieżących prac, jest też pogłębienie zrozumienia działania implementowanych modeli, wyznaczenie mechanizmów wspomagających wyjaśnialność modeli, tak aby wspomagać budowanie zaufania docelowych operatorów do wykonywanych detekcji, a także wspierać weryfikację i wdrażanie opracowywanych algorytmów ML w praktyce operacyjnej.



Rysunek 9: Ranking modeli nadzorowanych i nienadzorowanych, uszeregowany według uzyskanej precyzji poszczególnych detektorów anomalii [A9].

A10

Ostatni uwzględniony w tym zestawieniu artykuł dostarcza rozwiązanie dla jeszcze innego zadania związanego z ekstrakcją wzorców z danych wielowymiarowych. Dla popularnego problemu budowy modelu predykcji dla wielowymiarowych szeregów czasowych, w tym badaniu weryfikowanego z użyciem rynkowych danych dotyczących cen energii. Dla zadania, dla którego już wcześniej zaproponowano implementację konwolucyjnej sieci neuronowej do przetwarzania rzeczywistych danych rynkowych notowanych w postaci szeregów czasowych [B20], zaproponowano dodatkowy mechanizm detekcyjny pozwalający automatycznie sterować modelem predykcyjnym w odniesieniu do wykrywanych niestandardowych wzorców danych. Badanie zawiera również weryfikację podejścia z użyciem wielowymiarowego modelu danych, które obejmowały 16 szeregów (oprócz giełdowych cen z Towarowej Giełdy Energii, również dane z rynków gazu, uprawnień CO_2, bieżącej podaży energii odnawialnej wiatrowej i solarnej, a także prognozowanych warunków meteorologicznych oraz danych z systemu elektroenergetycznego, tzw. konektorów, pozwalających na przesył energii z sąsiednich krajów) i były zanotowane z godzinnym interwałem dla okresu 435 dni.

Do opracowania modelu głębokiego wykorzystano architekturę TCN (ang. *Temporal Convolutional Network* [28]). W procesie sekwencyjnego szkolenia tej sieci podawano wielowymiarowe dane w postaci szeregów czasowych – głównego, dotyczącego modelowanych cen, oraz dodatkowych (ang. *covariates* [29]), dla których wzorce w kolejnych epokach treningowych, identyfikowane są z użyciem konwolucyjnych przekształceń szeregów na

kolejnych warstwach modelu. W badaniu zidentyfikowano i potwierdzono szansę poprawy działania modelu prognostycznego, poprzez rozbudowę modelu danych oraz zastosowanie dodatkowego mechanizmu detekcji anomalii.

Głównym osiągnięciem opisanym w pracy [A10] jest zaproponowana zintegrowana architektura modelu predykcyjnego, uzupełniona o równolegle pracujący detektor anomalii, który sprawdza bieżące dane rynkowe, i przekazuje je zwrotnie do bazowego predyktora w postaci dodatkowego szeregu czasowego. Zatem do głównego modelu predykcyjnego, oprócz wcześniej przygotowanych danych treningowych, trafia również wektor detekcji, kodujący informację o stanie rynku (nominalnej lub anomalnej) w kolejnych okresach.

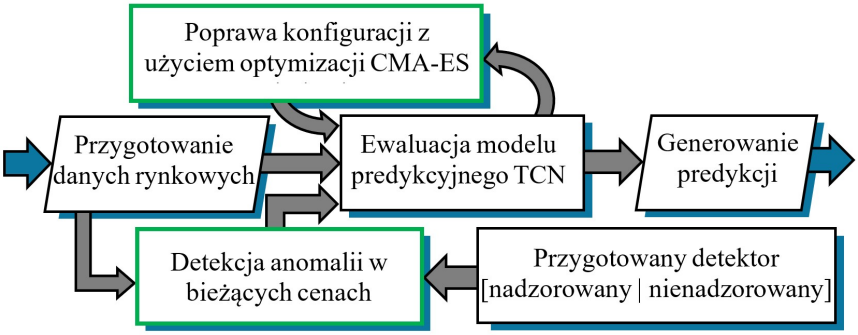
W badaniu ewaluowano zaproponowane rozwiązanie, na identycznym jak w [B20] zbiorze testowym, i zanotowano znaczne usprawnienie pracy predyktora. Podstawowy model predykcyjny szkolony był na danych rzeczywistych, i co zauważono we wstępnych pracach, w sytuacji dostrojenia parametrów sieci TCN, pozwalał na generowanie predykcji obarczonych niewielkim błędem. Pomiędzy podstawową konfiguracją sieci TCN a jej dostrojoną wersją zanotowano poprawę wyników, **z 16,344** do **10,239** dla sMAPE [30] i z **0,603** do **0,376** dla rMAE (metryki relatywnej porównującej błąd modelu w odniesieniu do modelu naiwnego liczone zgodnie z zaleceniami [31]).

Podobnie jak w [B20], w procesie ustalania optymalnej konfiguracji hiperparametrów predyktora zaimplementowano ewolucyjne strategie oparte o metodę CMA-ES [32] (ang. *Covariance Matrix Adaptation Evolution Strategy*). Przeszukiwana przestrzeń wariantów konfiguracyjnych obejmowała duży zakres możliwych ustawień, w tym kilka parametrów dyskretnych:
– rozmiar kernela sieci konwolucyjnej (3 do 7, z krokiem 2),
– liczbę filtrów konwolucyjnych (od 2 do 20),
– liczbę warstw (od 1 do 5),
– dylatację (od 1 do 4),
– uruchamianie regularyzacji modelu (włącz lub wyłącz);
– oraz 2 parametry ciągłe, którymi były tempo uczenia i natężenie mechanizmu wygaszania (ang. *dropout* [33]).

Dla każdego treningu przekazywano 10^3 różnych wariantów weryfikowanych ustawień hiperparametrów sieci TCN, a także zaimplementowano mechanizmy zmniejszania pojemności modelu (ang. *pruning* [34]) i wczesnego przerywania treningu. Weryfikowana strategia ewolucyjnego doboru parametrów (CME-ES) porównywana była z metodą wyszukiwania wielokierunkowego (ang. *Multiobjective Tree structured Parzen Estimator* – MoTPE [35]), a znacznie lepsze wyniki uzyskano dla tej pierwszej strategii, podczas weryfikacji obu z użyciem tego samego zbioru testowego.

Dodatkowo zbadano przydatność dwóch alternatywnych metod do zaimplementowania modułu detekcyjnego sprzężonego z modelem predykcyjnym. W pierwszym ewaluowanym wariancie rolę detektora anomalii pełnił nienadzorowany algorytm ML, oparty na metodzie *isolation forest* [36]. W drugim wariancie detekcję prowadzono z użyciem architektury XGBOD (ang. *Extreme Boosting Based Outlier Detection* [37]), która pracuje trybie półnadzorowanym – z nienadzorowanym wyodrębnianiem reprezentacji wzorców (ang. *Unsupervised Representation Learning*). Schemat określający składowe zadania dla opisywanej metody przedstawiono na rysunku 10.

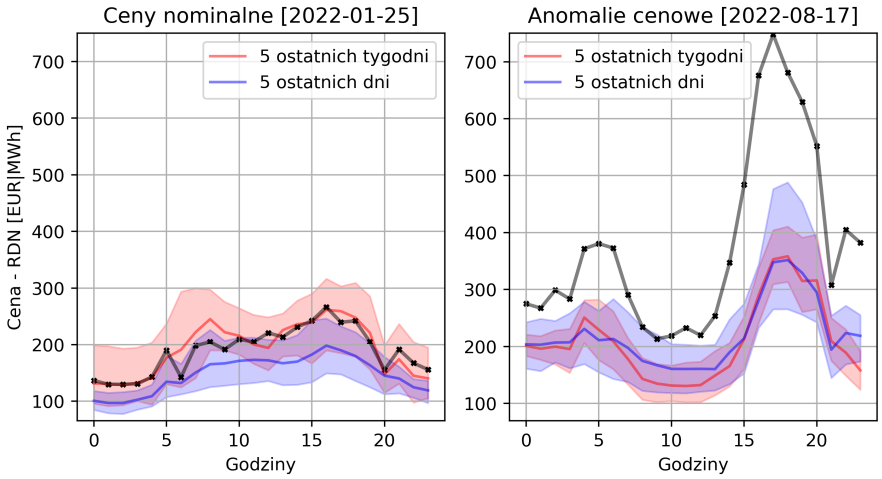
Oba zastosowane rozwiązania, w połączeniu ze strojoną ewolucyjnie siecią TCN pozwoliły istotnie zwiększyć moc predykcyjną finalnego modelu. Błędy predykcyjne dla modelu



Rysunek 10: Metoda predykcji cen wspomagana detektorem anomalii w cenach rynkowych, z modułem strojenia konfiguracji modelu ML [A10].

z nienadzorowanym detektorem wynosiły: **9,511 (sMAPE) i 0,334 (rMAE)**, natomiast dla metody wykorzystującej odpowiedzi modelu XGBOD: **8,190 (sMAPE) i 0,319 (rMAE)**.

Pobocznym efektem przedstawionej publikacji jest dodatkowe narzędzie do identyfikacji anomalii oraz zbiór danych giełdowych z oznaczeniami dla cen z Towarowej Giełdy Energii w Warszawie. Wobec wcześniejszego braku zbioru danych do ewaluacji zaproponowanej metody, habilitant przygotował taki zbiór na potrzeby pracy, jak również narzędzie wizualizacyjne do identyfikacji anomalii w danych rynkowych.



Rysunek 11: Opracowane narzędzie do etykietowania anomalii w szeregach cen giełdowych [A10]. Badany fragment cen oznaczono czarnym kolorem. Kolorowa ciągła linia oznacza średnią cenę w danej godzinie dla badanego okresu, cieniowany obszar oznacza przedziały ufności wykreślone dla $\pm 2 \cdot \sigma$ dla cen z badanego okresu (5 dni – oznaczone kolorem niebieskim, lub 5 godzin – oznaczone kolorem czerwonym).

Do wyszkolenia i ewaluacji detektorów anomalii wymagane było opracowanie zbioru danych, z oznaczeniami każdego fragmentu tych danych, czy przedstawiają nominalną, czy anomalną sytuację rynkową. Na potrzeby projektu przyjęto następującą definicję anomalii, rozpatrywaną osobno dla każdego kolejnego 24-godzinnego okna czasowego danych rynkowych: jako anomalia został oznaczony szereg cen, jeśli dla przynajmniej 6 z 24 godzin ceny były wyższe od $\pm$ dwukrotności odchylenia standardowego (σ) liczonego dla wcześniejszych danych. Dla rozpatrywanych danych rynkowych występuje zarówno dobowa

jak i tygodniowa sezonowość, dlatego w przygotowanym narzędziu wykreślano przedziały ufności (dla: $\pm 2 \cdot \sigma$), i to dla danych z poprzedzających badany moment 5 dni oraz dla zakresu poprzedzających 5 tygodni (cieniowane odpowiednio niebieskim lub czerwonym kolorem). Na rysunku 11 przedstawiono widok okna inspekcji cen dla dwóch różnych przypadków: sytuacja rynkowa z dnia 25 stycznia 2022 oznaczona została jako nominalna, natomiast dla dnia 17 sierpnia oznaczono anomalię rynkową – widać wyraźne, ponad dwukrotne tempo wzrostu cen).

W opisywanym badaniu detekcja anomalii prowadzona była jedynie dla szeregu czasowego z cenami, dlatego właśnie dla niego wykonano oznaczenia, i tylko te dane przetwarzane były przez oba wymienione detektory anomalii. Przedstawiona praca koncentruje się na samej metodzie, a sprawdzenie jej użyteczności w praktyce wymaga na pewno dalszych badań z danymi rynkowymi. Pozytywnie zweryfikowano hipotezę czy zaproponowana metoda, korzystająca dodatkowo z informacji o anomalnej sytuacji rynkowej, mogłaby lepiej adoptować się do sytuacji rynkowej i dostarczać predykcji o niższym błędzie. Natomiast samodzielnie wykonane przez habilitanta oznaczenia anomalii można traktować jedynie jako pilotaż, i dlatego nie zapisywano dla badania interpretacji i wskazówek rynkowych, i wskazano zalecenie do pogłębienia tego badania, oraz szerszej konsultacji oznaczeń anomalii z ekspertami domenowymi.

2.3 Podsumowanie

Powyższy opis dorobku można odczytywać na kilka sposobów. Każdy z nich wymagał odpowiedniego zaplanowania badania, użycia właściwego aparatu, technik i modeli uczenia maszynowego, aby pozwolić na zapisanie wymaganych wniosków na podstawie wiarygodnego badania.

Pierwszym sposobem może być recepta rozwiązania zadania ekstrakcji wzorców dla różnych danych wielowymiarowych dotyczących rzeczywistych problemów. W tej pracy opisane są dwie grupy takich danych: dane wielospektralne (obrazowe, z użyciem obrazów hiperspektralnych i ortofotomap, oraz spektrograficzne, rejestrowane wysokorozdzielczym spektrofotometrem) oraz dane telemetryczne (przedstawiające zmienne w czasie charakterystyki pracy urządzeń, w tej pracy z użyciem danych satelitarnych).

Drugi poziom odczytywania poniższej pracy to proces przygotowania modeli uczenia maszynowego, czyli taki sposób implementacji modeli analitycznych, który pozwala na automatyczne modelowania badanych zagadnień. W tej pracy omówiono zarówno metody uczenia maszynowego jak i uczenia głębokiego. Tryb, w jakim przygotowywane były omawiane modele, może się też w praktyce istotnie różnić, i w tym zestawieniu ujęto zarówno przygotowanie modelu trybie uczenia nadzorowanego [A1–A5, A8–A10], jak i nienadzorowanego [A9, A10], oraz ich łączenia [A10]. W bieżących pracach habilitanta ten wątek jest obecnie intensywnie rozwijany o kolejne warianty tego procesu.

Następna warstwa, na której można rozpatrywać prace, to różne zadania uczenia maszynowego wykorzystywane dla ekstrakcji wzorców w danych przedstawionych artykułach. Są to zarówno problemy klasyfikacyjne (zarówno dla klasyfikacji binarnej [A1, A2] jak i wieloklasowej [A3]), estymacyjne, korzystające z modeli regresyjnych ML, oraz pożytkujące dynamiczny wymiar danych w postaci szeregów czasowych problemy prognostyczne i związane z detekcją anomalii.

Ostatnim spojrzeniem może być ocena przydatności przedstawionych publikacji z punk-

tu widzenia aplikacyjnego – rzeczywiste dane, które wykorzystano w opisanych badaniach, dotyczą domenowych problemów, i dla nich w poszczególnych artykułach również zapisano szereg wniosków praktycznych. Umiejętna komunikacja i współpraca z badaczami z innych dziedzin, którzy pomagali weryfikować oraz interpretować, praktyczną użyteczność zaproponowanych metod uczenia maszynowego i przedstawionych konkluzji domenowych, stanowi równie istotną część prezentowanego dorobku.

2.4 Szczegółowe statystyki bibliograficzne

Sumaryczne statystyki dorobku według baz danych indeksujących oraz szczegółowe dane bibliograficzne (na dzień 21 listopada 2025):

- Scopus (Scopus-ID: 55357278900):
  – liczba artykułów autora indeksowanych w bazie Scopus: 42 (w tym 11 artykułów w czasopismach Q1: [A4, A9], [B1, B9, B19, B27, B28, B30, B39, B50, B55]),
  – sumaryczna liczba cytowań: 272 (bez autocytowań: 211),
  – h-index: 10 (bez autocytowań: 7),

- Web of Science (WoS ID: 2088830):
  – liczba artykułów autora indeksowanych w bazie Web of Science: 35,
  – sumaryczna liczba cytowań: 213 (bez autocytowań: 162),
  – h-index: 9 (bez autocytowań: 7),
  – sumaryczny IF: 71,621 (wyliczony z użyciem IF czasopisma w roku publikacji):
    (1,444 [B55] + 3,342 [B50] + 0,735 [B48] + 1,600 [B39] + 1,995 [B40] + 2,105 [A1] + 1,200 [B29] + 2,700 [B28] + 3,900 [B30] + 5,000 [A4] + 3,800 [B19] + 2,400 [A7] + 3,400 [B27] + 1,000 [B18] + 1,600 [B17] + 2,100 [A2] + 16,400 [B9] + 5,800 [A9] + 7,500 [B1] + 3,600 [B7]) = 71,621

- Google Scholar (user=7gk0kB0AAAAJ):
  – liczba artykułów autora indeksowanych w bazie Google Scholar: 65,
  – sumaryczna liczba cytowań: 417,
  – h-index: 13,

- DBLP (dblp.org/pid/241/6354.html):
  – liczba artykułów autora indeksowanych w bazie DBLP: 29,

- liczba publikacji na międzynarodowych konferencjach naukowych: 33, w tym:
  – 9 publikacji na konferencjach z rankingiem **CORE A**: [A3, A5, A6, A8, A10], [B2, B3, B10, B20],
  – 1 publikacja na konferencji z rankingiem **CORE B**: [B4],
  – 4 publikacje na konferencjach z rankingiem **CORE C**: [B11–B13, B35],

- liczba publikacji według punktacji MNiSW (45 pozycji):

  – **200 pkt**: 2 pozycje, **140 pkt**: 12 pozycji, **100 pkt**: 7 pozycji, **70 pkt**: 5 pozycji, **45 pkt**: 1 pozycja, **40 pkt**: 2 pozycje, **30 pkt**: 1 pozycja, **20 pkt**: 15 pozycji,

  - w tym w dyscyplinie *informatyka techniczna i telekomunikacja* (35 pozycji):
    200 pkt: 2 pozycje: [A7], [B1],
    140 pkt: 12 pozycji: [A3, A5, A6, A8–A10], [B2, B9, B10, B19, B20, B39],
    100 pkt: 5 pozycji: [A4], [B7, B27, B28, B30],
    70 pkt: 4 pozycje: [A2], [B13, B34, B40],
    40 pkt: 2 pozycje: [A1], [B29],
    20 pkt: 10 pozycji: [B4–B6, B11, B12, B21, B35, B37, B38, B41],

- liczba publikacji według wydawnictwa:
  – **IEEE**: 9 pozycji [B6, B7, B9, B11–B13, B15, B50, B54],
  – **Springer**: 7 pozycji [A2, A5, A8], [B37, B38, B41, B43],
  – **Elsevier**: 6 pozycji [A7], [B1, B29, B39, B44, B55],
  – **ACM**: 6 pozycje [A3, A6, A10], [B2, B10, B20],
  – **IOS Press**: 4 pozycje [A1], [B4, B5, B35],
  – **MDPI**: 4 pozycje [A4], [B27, B28, B30],
  – **Nature**: 2 pozycje [A9], [B19],
  – **IET**: 1 pozycja [B40],
  – **Wiley**: 1 pozycja [B48],

- inne bazy informacji naukowych:
  – ORCID: 0000-0003-1089-1778,
  – rejestr naukometryczny pracodawcy kandydata: https://bazawiedzy.po.edu.pl/
  – profil ResearchGate: researchgate.net/profile/Bogdan-Ruszczak.

Tabela 2: Zestawienie cytowanych pozycji bibliograficznych i liczb cytowań do wyznaczenia współczynników h-index.

Publikacja	Liczba cytowań			Punkty	Kwantyl	Indeks	CORE	Rok
	Scopus	WoS	GS	MNiSW	Scopus	IF**	ranking	
[A9]	3	2	6	140	96	6,900	–	2025
[B2]	1	–	–	140	–	–	A	2025
[B9]	7	6	7	140	99	16,400	–	2024
[A2]	3	2	4	70	56	2,000	–	2024
[B12]	3	2	3	20	–	–	C	2024
[B11]	1	1	2	20	–	–	C	2024
[A10]	1	–	1	140	–	–	A	2024
[B19]	**12**	**13**	**14**	140	92	3,800	–	2023
[A8]	9	–	7	140	–	–	A	2023
[B27]	7	6	7	100	83	3,400	–	2023
[A7]	5	3	7	200	64	2,400	–	2023
[A3]	3	3	4	140	–	–	A	2023
[B20]	3	1	2	140	–	–	A	2023
[A5]	2	–	4	140	–	–	A	2023
[B28]	**22**	**22**	**27**	100	75	2,700	–	2022
[A4]	**14**	9	**16**	100	90	5,000	–	2022
[B29]	**12**	9	**15**	40	72	1,200	–	2022
[B30]	8	6	**13***	100	87	3,900	–	2022
[A1]	**14**	**14**	**14**	40	63	2,105	–	2020
[B39]	**26**	**14**	**31**	140	83	1,600	–	2019
[B40]	**10***	8	**16**	70	61	1,995	–	2019
[B37]	9	2	12	20	–	–	–	2019
[B38]	1	1	1	20	–	–	–	2019
[B44]	**28**	**22**	**30**	20	65	–	–	2018
[B43]	**14**	**12**	**17**	20	–	–	–	2018
[B41]	2	2	3	20	–	–	–	2018
[B48]	7	**9***	**31**	20	74	0,735	–	2016
[B50]	**32**	**28**	**36**	45	99	3,342	–	2015
[B55]	9	**10**	**18**	30	91	1,444	–	2013
[B54]	2	2	–	20	–	–	–	2013

*ostatnia pozycja wliczana do wyznaczenia h-index, pozycje przyjęte do wyliczeń dla poszczególnych baz cytowań zostały pogrubione

**wartość indeksu IF czasopisma w roku publikacji

***przy ustalaniu liczby punktów MNiSW i współczynnika IF dla wszystkich artykułów oraz dla artykułów konferencyjnych rankingu CORE tych konferencji przyjęto wartości obowiązujące w roku publikacji

2.5 Lista pozostałych publikacji po doktoracie

[B1] P. J. Navarro, E. Salas, F. Rosique, **B. Ruszczak**, i M. Egea-Cortines, "Conditional Generative Adversarial Network for segmentation and image reconstruction of grape berries in variable illumination conditions," *Expert Systems With Applications*, 2025, [**MNiSW: 200 pkt, IF 2024: 7,500 Scopus: 96**].

[B2] M. Przewozniczek, F. Chicano, R. Tinós, J. Nalepa, **B. Ruszczak**, i A. M. Wijata, "On revealing the hidden problem structure in real-world and theoretical problems using walsh coefficient influence," w *Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Malaga, Hiszpania*. ACM, 2025, s. 1–10, [**MNiSW: 140 pkt, ranking konferencji CORE A, cytowania Scopus: 1, nagroda best paper award**].

[B3] M. Bikhit, S. Kandanaarachchi, K. Kotowski, J. Nalepa, i **B. Ruszczak**, "Morphological leave-one-out kernel density estimates for anomaly detection in satellite telemetry," w *Proceedings of the Machine Learning for Irregular Time Series (ML4ITS) Workshop at European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Porto, Portugalia*. Springer, 2025, s. 1–16, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A**].

[B4] J. Sahs i **B. Ruszczak**, "Configuring the transformer-based electricity market predictions using ant colony optimisation," w *International Workshop on Artificial Intelligence and Machine Learning for Emerging Topics, ALLEGET – Workshop Proceedings of the 21th International Conference on Intelligent Environments, Ambient Intelligence and Smart Environments*, ser. Ambient Intelligence and Smart Environments, tom 34. IOS Press, 2025, s. 1–5, ✉ [**MNiSW: 20pkt, ranking konferencji CORE B**].

[B5] **B. Ruszczak**, "Alleget keynote remarks: From ambient to space and back ... with a little help from machine learning," w *International Workshop on Artificial Intelligence and Machine Learning for Emerging Topics, ALLEGET – Workshop Proceedings of the 21th International Conference on Intelligent Environments, Ambient Intelligence and Smart Environments*, ser. Ambient Intelligence and Smart Environments, tom 34. IOS Press, 2025, s. 94–95, ✉ [**MNiSW: 20pkt, ranking konferencji CORE B**].

[B6] J. Sahs i **B. Ruszczak**, "Enhancing the transformer-based forecasting models by training datasets permutation," w *6th International Conference In Electronic Engineering, Information Technology & Education, EEITE*, Chania, Grecja, 2025, s. 1–5, ✉.

[B7] P. Pałka, T. Śliwiński, P. Kaszyński, M. Kuta, **B. Ruszczak**, M. Malec, P. Saługa, i J. Kamiński, "Planning and scheduling system for electric vehicle charging," *IEEE Access*, tom 13, p. 113905 – 113923, 2025, ✉ [**MNiSW: 100 pkt, IF 2024: 3,600**].

[B8] K. Kotowski, C. Haskamp, J. Andrzejewski, **B. Ruszczak**, J. Nalepa, D. Lakey, P. Collins, A. Kolmas, M. Bartesaghi, J. Martínez-Heras, i G. De Canio, "The making of the european space agency benchmark for anomaly detection in satellite telemetry," w *18th International Conference on Space Operations, SpaceOps*, Montreal, Canada, 2025, s. 1–10.

[B9] J. Nalepa, L. Tulczyjew, B. Le Saux, N. Longépé, **B. Ruszczak**, A. M. Wijata, K. Smykała, M. Myller, M. Kawulok, R. S. Kuzu, F. Albrecht, C. Arnold, M. Alasawedah, S. Angeli, D. Nobileau, A. Ballabeni, A. Lotti, A. Locarini, D. Modenini, P. Tortora, i M. Gumiela, "Estimating Soil Parameters From Hyperspectral Images: A benchmark dataset and the

outcome of the HYPERVIEW challenge," *IEEE Geoscience and Remote Sensing Magazine*, tom 12, no. 3, s. 35–63, 2024, [**MNiSW: 140 pkt, IF 2024: 16,400, cytowania Scopus: 7, WoS: 6**].

[B10] **B. Ruszczak**, M. Myller, L. Tulczyjew, A. M. Wijata, i J. Nalepa, "Particle Swarm optimization meets deep learning for estimating root-zone soil moisture from hyperspectral images," w *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO, Melbourne, Australia*, X. Li i J. Handl, Red. ACM, 2024, p. 695–698, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A**].

[B11] A. M. Wijata, T. Lakota, M. Cwiek, **B. Ruszczak**, M. Gumiela, L. Tulczyjew, A. Bartoszek, N. Longépé, K. Smykala, i J. Nalepa, "Intuition-1: Toward in-orbit bare soil detection using spectral vegetation indices," w *2024 IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, 2024, s. 1708–1712, [**MNiSW: 20 pkt, ranking konferencji CORE C**].

[B12] A. M. Wijata, A. Miroszewski, B. L. Saux, N. Longépé, **B. Ruszczak**, i J. Nalepa, "Detection of bare soil in hyperspectral images using quantum-kernel support vector machines," w *2024 IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, 2024, s. 817–822, [**MNiSW: 20 pkt, ranking konferencji CORE C, cytowania Scopus: 3, WoS: 2**].

[B13] L. Tulczyjew, **B. Ruszczak**, M. Myller, A. M. Wijata, D. Boguszewska-Mańkowska, i J. Nalepa, "Convolutional neural networks estimate root-zone soil moisture from hyperspectral images," w *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2024, ✉ [**MNiSW: 70 pkt, ranking konferencji CORE C**].

[B14] **B. Ruszczak**, A. M. Wijata, i J. Nalepa, "Accelerating machine learning for earth observation using Intel Extension for Scikit-learn," w *Progress in Polish Artificial Intelligence Research 5. Proceedings of the 5th Polish Conference on Artifcial Intelligence, PP-RAI*, 2024, s. 23–28, ✉.

[B15] A. Luckiewicz, **B. Ruszczak**, M. Pelc, D. Mikołajewski, A. Sudoł, B. Grochowicz, E. J. Gorzelanczyk, P. Mendon, i A. Kawala-Sterniuk, "Multi-method analysis of emotional and affective states," w *2024 Progress in Applied Electrical Engineering, PAEE*, ser. IEEE PES, 2024, s. 1–8.

[B16] K. Kotowski, C. Haskamp, J. Andrzejewski, **B. Ruszczak**, J. Nalepa, D. Lakey, P. Collins, A. Kolmas, M. Bartesaghi, J. Martinez-Heras, i G. D. Canio, "European space agency benchmark for anomaly detection in satellite telemetry," s. 245–250, 2024, arXiv.

[B17] J. Maj, **B. Ruszczak**, i S. Kubiciel-Lodzinska, "Towards algorithm-assisted career management - the challenge for new immigration countries. predicting the migrant"s work trajectory using ensemble learning," *Central European Business Review*, 2024, [**MNiSW: 20 pkt, IF 2024: 1,600**].

[B18] K. Mazur-Włodarczyk, A. Rajchel, i **B. Ruszczak**, "Perception and use of new technologies in crafts education and craftsmanship among young craftsmen: a study of sustainable practices," *Economics and Environment*, 2024, [**MNiSW: 100 pkt, IF 2024: 1,000**].

[B19] M. Tomaszewski, J. Nalepa, E. Moliszewska, **B. Ruszczak**, i K. Smykała, "Early detection of solanum lycopersicum diseases from temporally-aggregated hyperspectral measurements using machine learning," *Scientific Reports*, tom 13, no. 1, p. 7671, 2023, [**MNiSW: 140 pkt, IF 2023: 3,800, cytowania Scopus: 12, WoS: 13**].

[B20] **B. Ruszczak**, "Evolutionary configuration of the temporal convolutional network for the electricity price prediction," w *Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO*, S. Silva i L. Paquete, Red. Lizbona, Portugalia: ACM, 2023, ✉ [**MNiSW: 140 pkt, ranking konferencji CORE A, cytowania Scopus: 3, WoS: 1**].

[B21] **B. Ruszczak** i K. Rudnik, "Ensemble-based versus expert-assisted approach to carbon price features selection," w *Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, FedCSIS*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, M. Paprzycki, i D. Slezak, Red., tom 37, Warszawa, 2023, s. 245–250, ✉ [**MNiSW: 20 pkt, ranking konferencji CORE Multi-conference**].

[B22] A. M. Wijata, A. Miroszewski, B. L. Saux, N. Longépé, **B. Ruszczak**, i J. Nalepa, "Annotating large satellite telemetry dataset for ESA international AI anomaly detection benchmark," w *From foresight to impact - Proceedings of the 2023 conference on Big Data from Space, BiDS*, P. Soille, S. Lumnitz, i A. S., Red. Wiedeń, Austria: Publications Office of the European Union, 2023, s. 341–344.

[B23] J. Nalepa, **B. Ruszczak**, K. Kotowski, J. Andrzejewski, A. Musial, D. Evans, V. Zelenevskiy, S. Bammens, i R. Laurinovics, "Look Ma, no ground truth! On building supervised anomaly detection from OPS-SAT telemetry," w *Proceedings of 74th International Astronautical Congress, IAC*, 2023, s. 1–8.

[B24] A. M. Wijata, T. Lakota, M. Cwiek, **B. Ruszczak**, M. Gumiela, Łukasz Tulczyjew, N. Longépé, i J. Nalepa, "Bringing bare soil detection on-board intuition-1 through exploiting data-level digital twins," w *Proceedings of 74th International Astronautical Congress, IAC*, 2023, s. 1–5.

[B25] S. Kubiciel-Lodzińska, J. Maj, K. Widera, P. Bębenek, **B. Ruszczak**, i M. Sokołowska, "Escaping brain waste. work experience and qualifications of highly skilled ukrainians in Poland," *Zeszyty Naukowe Politechniki Śląskiej. Organizacja i Zarządzanie*, tom -, no. 170, s. 267–282, 2023, [**MNiSW: 100 pkt**].

[B26] K. Kotowski, C. Haskamp, **B. Ruszczak**, J. Andrzejewski, i J. Nalepa, "Annotating large satellite telemetry dataset for ESA international AI anomaly detection benchmark," w *Proceedings of the 2023 conference on Big Data from Space, BiDS – From foresight to impact*, P. Soille, S. Lumnitz, i S. Albani, Red. Wiedeń, Austria: Publications Office of the European Union, 2023, s. 341–344.

[B27] **B. Ruszczak**, P. Michalski, i M. Tomaszewski, "Overview of image datasets for deep learning applications in diagnostics of power infrastructure," *Sensors*, tom 23, no. 16, 2023, ✉ [**MNiSW: 100 pkt, IF 2023: 3,400, cytowania Scopus: 7, WoS: 6**].

[B28] D. Boguszewska-Mańkowska, **B. Ruszczak**, i K. Zarzyńska, "Classification of potato varieties drought stress tolerance using supervised learning," *Applied Sciences*, tom 12, no. 4, 2022, ✉ [**MNiSW: 100 pkt, IF 2022: 2,700, cytowania Scopus: 22, WoS: 22**].

[B29] **B. Ruszczak** i D. Boguszewska-Mańkowska, "Deep Potato – the hyperspectral imagery of potato cultivation with reference agronomic measurements dataset: Towards potato physiological features modeling," *Data in Brief*, tom 42, p. 108087, 2022, ✉ [**MNiSW: 40 pkt, IF 2022: 1,200, cytowania Scopus: 12, WoS: 9**].

[B30] **B. Ruszczak** i D. Boguszewska-Mańkowska, "Soil moisture a posteriori measurements enhancement using ensemble learning," *Sensors*, tom 22, no. 12, 2022, ✉ [**MNiSW: 100 pkt, IF 2022: 3,900, cytowania Scopus: 8, WoS: 6**].

[B31] R. Gasz i **B. Ruszczak**, "Image processing methods as an effective identification and diagnostics tool," w *Proceedings of 3rd International Conference in Electronic Engineering, Information Technology & Education, EEITE, Hellenic Mediterranean University*, Chania, Grecja, 2022, s. 76–81.

[B32] **B. Ruszczak** i R. Gasz, "The application of machine learning methodology aiding agriculture measurements," w *Proceedings of 3rd International Conference in Electronic Engineering, Information Technology & Education, EEITE, Hellenic Mediterranean University*, Chania, Grecja, 2022, s. 82–85, ✉.

[B33] J. Nalepa, J. Andrzejewski, **B. Ruszczak**, K. Kotowski, A. Musiał, D. Evans, V. Zelenevskiy, S. Bammens, i R. Laurinovics, "Toward on-board detection of anomalous events from OPS-SAT telemetry using deep learning," w *8th International Workshop on On-board Payload Data Compression, OBPDC*, Ateny, Grecja, 2022, s. 1–7.

[B34] R. Gasz, M. Tomaszewski, S. Zator, **B. Ruszczak**, i A. Kawala-Sterniuk, "Implementation of vision methods for the identification of spatial objects in electrical power-line latticework," w *Proceedings of the 37th International Business Information Management Association Conference*, S. Khalid, Red., 2021, s. 7797–7806, [**MNiSW: 70 pkt,**].

[B35] K. Smykała, **B. Ruszczak**, i K. Dziubański, "Application of ensemble learning to detect alternaria solani infection on tomatoes cultivated under foil tunnels," w *Workshop Proceedings of the 16th International Conference on Intelligent Environments, Ambient Intelligence and Smart Environments, 2020, vol. 28*, C. Iglesias *et al.*, Red. Madryt, Hiszpania: IOS Press, 2020, s. 127–132, ✉ [**MNiSW: 20 pkt, ranking konferencji CORE C, cytowania WoS: 4**].

[B36] **B. Ruszczak** i Z. Kawalec, "The overview of machine learning applications for agriculture based on hyperspectral data," w *Proceedings of 2nd International Conference in Electronic Engineering, Information Technology & Education, EEITE, Hellenic Mediterranean University*, Chania, Grecja, 2020, s. 26–29, ✉.

[B37] P. Michalski, **B. Ruszczak**, i P. J. N. Lorente, "The implementation of a convolutional neural network for the detection of the transmission towers using satellite imagery," w *Information Systems Architecture and Technology: Proceedings of 40th Anniversary International Conference on Information Systems Architecture and Technology, ISAT 2019*, J. Świątek, L. Borzemski, i Z. Wilimowska, Red. Cham: Springer, 2020, s. 287–299, ✉ [**MNiSW: 20 pkt, cytowania Scopus: 9, WoS: 2**].

[B38] M. Słapek, K. Smykała, i **B. Ruszczak**, "Brassica napus florescence modeling based on modified vegetation index using sentinel-2 imagery," w *Artificial Intelligence and Soft Computing*, L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, i J. M. Zurada, Red. Cham: Springer, 2019, s. 80–90, ✉ [**MNiSW: 20 pkt, cytowania Scopus: 1, WoS: 1**].

[B39] M. Tomaszewski, **B. Ruszczak**, P. Michalski, i S. Zator, "The study of weather conditions favourable to the accretion of icing that pose a threat to transmission power lines," *International Journal of Critical Infrastructure Protection*, tom 25, s. 139–151, 2019, ✉ [**MNiSW: 140 pkt, IF 2019: 1,600, cytowania Scopus: 26, WoS: 14**].

[B40] M. Tomaszewski, P. Michalski, **B. Ruszczak**, i S. Zator, "Detection of power line insulators on digital images with the use of laser spots," *IET Image Processing*, tom 13, no. 12, s. 2358–2366, 2019, ✉ [**MNiSW: 70 pkt, IF 2019: 1,995, cytowania Scopus: 10, WoS: 8**].

[B41] R. Gasz, **B. Ruszczak**, M. Tomaszewski, i S. Zator, "The registration of digital images for the truss towers diagnostics," w *Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT 2018*, L. Borzemski, J. Świątek, i Z. Wilimowska, Red. Cham: Springer, 2019, s. 166–177, [**MNiSW: 20 pkt, cytowania Scopus: 2, WoS: 2**].

[B42] S. Kubiciel-Lodzińska i **B. Ruszczak**, "The attractiveness of Poland for economic migrants. assessment by employers: results of empirical studies," *Olsztyn Economic Journal*, tom 13, no. 4, s. 425–440, 2018, [**MNiSW: 13 pkt**].

[B43] P. Michalski, **B. Ruszczak**, i M. Tomaszewski, "Convolutional neural networks implementations for computer vision," w *Biomedical Engineering and Neuroscience*, W. P. Hunek i S. Paszkiel, Red. Cham: Springer, 2018, s. 98–110, [**MNiSW: 20 pkt, cytowania Scopus: 14, WoS: 12**].

[B44] M. Tomaszewski, **B. Ruszczak**, i P. Michalski, "The collection of images of an insulator taken outdoors in varying lighting conditions with additional laser spots," *Data in Brief*, tom 18, s. 765–768, 2018, ✉ [**MNiSW: 20 pkt, cytowania Scopus: 28, WoS: 22**].

[B45] S. Kubiciel-Lodzińska i **B. Ruszczak**, "Labor market segmentation as an important factor in research of labor immigration. comparative analysis based on the opolskie voivodship empirical findings," *Barometr Regionalny*, tom 16, no. 4, s. 61–67, 2018, [**MNiSW: 14 pkt**].

[B46] R. Gasz, P. Michalski, i **B. Ruszczak**, "Możliwości wykorzystania narzędzi computer vision w przedsiębiorstwach," w *Technologie informatyczne we współczesnej rzeczywistości gospodarczej, Studia i Monografie*, J. Wielki, Red. Opole: Oficyna Wydawnicza Politechniki Opolskiej, 2017, s. 45–68, ✉ [**MNiSW: 20 pkt**].

[B47] S. Kubiciel-Lodzińska i **B. Ruszczak**, "Non-salary employment conditions offered to foreigners legally employed in Poland," *Central and Eastern European Journal of Management and Economics*, tom 5, no. 1, s. 77–91, 2017.

[B48] S. Kubiciel-Lodzińska i **B. Ruszczak**, "The determinants of student migration to Poland based on the Opolskie voivodeship study," *International Migration*, tom 54, no. 5, s. 162–174, 2016, [**MNiSW(A): 20 pkt, IF 2016: 0,735, cytowania Scopus: 7, WoS: 9, Google Scholar: 31**].

[B49] **B. Ruszczak**, "Ocena efektywności ekonomicznej i finansowej projektów rozwojowych: wybrane narzędzia," w *Ekonomia: przewodnik dla studentów i doktorantów kierunków technicznych*, K. Malik, Red. Opole: Oficyna Wydawnicza Politechniki Opolskiej, 2016, s. 169–198, ✉.

[B50] **B. Ruszczak** i M. Tomaszewski, "Extreme value analysis of wet snow loads on power lines," *IEEE Transactions on Power Systems*, tom 30, no. 1, s. 457–462, 2015, ✉ [**MNiSW(A): 45 pkt, IF 2015: 3,342, cytowania Scopus: 32, WoS: 28**].

[B51] **B. Ruszczak** i P. Misiurski, "Innovations related to the low emission technologies and renewable sources of energy," w *Effective Transfer of Knowledge from Science to Industry in the Opolskie Voivodeship. Requirements for an Effective Cooperation*, K. Malik i Łukasz Dymek, Red. Opole: Oficyna Wydawnicza Politechniki Opolskiej, 2015, s. 138–148, ✉.

[B52] L. Dzierżanowski, **B. Ruszczak**, i M. Tomaszewski, "Oprogramowanie wspomagające klasyfikację linii i planowanie kosztów eksploatacji sieci elektroenergetycznej," w *Diagnostyka inwestycyjno-eksploatacyjna wybranych urządzeń i sieci energetycznych*, S. Zator i Z. Kabza, Red. Opole: Oficyna Wydawnicza Politechniki Opolskiej, 2014, s. 149–157.

[B53] M. Tomaszewski i **B. Ruszczak**, "Wspomaganie procesu eksploatacji napowietrznych linii elektroenergetycznych przez ograniczenie wpływu zimowych zjawisk atmosferycznych," *Energetyka, problemy energetyki i gospodarki paliwowo-energetycznej*, tom 5, s. 272–282, 2014, ✉.

[B54] L. Dzierżanowski, **B. Ruszczak**, i M. Tomaszewski, "Frequency of power line failures in life cycle cost analysis," w *IEEE International Symposium on Electrodynamic and Mechatronic Systems, SELM*, B. Tomczuk, A. Waindok, i J. Zimon, Red. Opole-Zawiercie: IEEE Xplore, 2013, s. 55–56, [**cytowania Scopus:2, WoS: 2**].

[B55] M. Tomaszewski i **B. Ruszczak**, "Analysis of frequency of occurrence of weather conditions favouring wet snow adhesion and accretion on overhead power lines in Poland," *Cold Regions Science and Technology*, tom 85, s. 102–108, 2013, [**MNiSW(A): 30 pkt, IF 2013 1,444, cytowania Scopus: 9, WoS: 10**].

3 Pozostałe osiągnięcia

3.1 Projekty badawczo rozwojowe

Kierowanie projektami badawczo-rozwojowymi:

- [P1] 2024–2025 – Optymalizacja algorytmów uczenia maszynowego do przetwarzania danych wielowymiarowych, wraz z ich weryfikacją na danych rzeczywistych, grant JM Rektora Politechniki Opolskiej w konkursie DELTA, nr. projektu 272/24,
  - stanowisko: **kierownik projektu, ekspert ds. uczenia maszynowego**,
  - rola: opracowanie algorytmów uczenia maszynowego do analizy wielowymiarowych danych rzeczywistych, ewaluacja modeli ML, autor wniosku projektowego i raportu końcowego,
  - efekty prac w: [A6], [B2–B4, B13],
- [P2] 2020–2022 – DeepPotato: Opracowanie bezinwazyjnego systemu do pomiaru wilgotności gleby na poziomie korzeni dla uprawy ziemniaka w oparciu o nowe algorytmy wykorzystujące metody uczenia maszynowego do analizy danych hiperspektralnych, NCBiR, POIR.04.01.04-00-0009/19,
  - stanowisko: **kierownik B+R, ekspert ds. uczenia maszynowego**,
  - rola: opracowanie algorytmów uczenia maszynowego do analizy obrazowań hiperspektralnych, koordynator współpracy z Katedrą Algorytmiki i Oprogramowania Wydziału Automatyki, Elektroniki i Informatyki Politechniki Śląskiej i Instytutem Hodowli i Aklimatyzacji Roślin – Państwowym Instytutem Badawczym (IHAR PIB), współautor wniosku projektowego i raportu końcowego,
  - efekty prac w: [A3–A5], [B10, B28–B30],
- [P3] 2018–2018 – Opracowanie aplikacji wspomagającej i optymalizującej rozmieszczenie uli pszczelich za pomocą inteligentnej analizy wizyjnej zdjęć satelitarnych, Regionalny Program Województwa Opolskiego RPOP.01.01.00-16-0056/16,

  - stanowisko: **kierownik projektu B+R, ekspert ds. uczenia maszynowego**,
  - rola: planowanie i realizacja badań nad algorytmami i opracowanie algorytmów detekcyjnych,
  - efekty prac w: [B36, B38].

Wykonawca w międzynarodowych projektach badawczo-rozwojowych:

- [P4] 2024–2025 – BOLERO: On-Board Continual Learning for SatCom Systems, Europejska Agencja Kosmiczna (ESA), 4000142406/23/NL/FGL,
  - stanowisko: **specjalista ds. uczenia maszynowego**,
  - rola: opracowanie procesu uczenia ciągłego i ewaluacja modeli uczenia maszynowego na urządzeniach brzegowych KP Labs Leopard DPU oraz komputerze neuromorficznym Intel Akida oraz opracowanie publikacji naukowych,
- [P5] 2024–2024 – Brainsourcing dla Szacowania Uwagi Afektywnej, CHIST-ERA, NCN 2021/03/Y/ST7/0008,
  - stanowisko **specjalista ds. danych i algorytmów**,
  - rola: opracowanie algorytmów uczenia maszynowego i analiza danych,
  - efekty prac w: [B15],
- [P6] 2023–2024 – Few-shot Anomaly Detection In Satellite Telemetry, Europejska Agencja Kosmiczna (ESA), 4000141301,
  - stanowisko: **specjalista ds. uczenia maszynowego**,
  - rola: opracowanie i ewaluacja algorytmów uczenia maszynowego na podstawie ekstremalnie niedoreprezentowanych przypadków do detekcji anomalii w sygnałach telemetrycznych dla rzeczywistych danych satelitarnych,
- [P7] 2022–2023 – ESA Anomalies Dataset for International AI Anomaly Detection Benchmark, Europejska Agencja Kosmiczna (ESA), T709-705OS,
  - stanowisko: **specjalista ds. uczenia maszynowego**,
  - rola: opracowanie i ewaluacja detekcyjnych modeli uczenia maszynowego,
  - efekty prac w: [A7], [B8, B16],
- [P8] 2022–2022 – OPS-SAT – On-board Anomaly detection from the OPS-SAT telemetry using deep learning, Europejska Agencja Kosmiczna (ESA), 4000137339/22/NL/GLC/ov.,
  - stanowisko: **specjalista ds. uczenia maszynowego**,
  - rola: opracowanie procedury ewaluacji oraz detekcyjnych modeli uczenia maszynowego dla analizy telemetrii satelity OPS-SAT, przygotowanie ich do uruchomienia na pokładzie satelity, oraz opracowanie zbiorów danych i publikacji naukowych,
  - efekty prac w: [A8, A9], [B23, B33].

Wykonawca w krajowych projektach badawczo-rozwojowych:

- [P9] 2021–2023 – RoEV: Opracowanie zautomatyzowanego narzędzia wspierającego decyzje kierowcy w zakresie optymalnego wyboru stacji ładowania samochodu elektrycznego, wykorzystującego metody programowania matematycznego oraz uczenia maszynowego, POIR.01.01.01-00-0855/21, NCBiR,
  - stanowisko: **ekspert ds. uczenia maszynowego**:
  - rola: opracowanie algorytmów uczenia maszynowego dla estymacji zmiany stanu naładowania pojazdu, współautor wniosku projektowego i raportu końcowego,
  - efekty prac w: [B7],
- [P10] 2021–2023 – TwinIO: Opracowanie zintegrowanego narzędzia bazującego na metodach programowania matematycznego i uczenia maszynowego umożliwiającego przedsiębiorstwom optymalizację procesu zakupów, produkcji własnej i magazynowania energii elektrycznej, przyjmując jako kryterium optymalizacji minimalizację sumarycznych kosztów energii, POIR.01.01.01-00-0709/21, NCBiR,
  - stanowisko: **ekspert ds. uczenia maszynowego**,
  - rola: opracowanie algorytmów uczenia maszynowego dla prognozowania giełdowych cen energii oraz wytwarzania energii ze źródeł odnawialnych, współautor wniosku projektowego i raportu końcowego,
  - efekty prac w: [A10], [B20],
- [P11] 2021–2021 – Opracowanie systemu pozwalającego na zdalne badanie zasobności gleby w makroelementy oraz odczynu gleby z użyciem nowych algorytmów uczenia maszynowego na podstawie obrazowania hiperspektralnego (SmartSoil), POIR.01.01.01-00-2197/20, NCBiR,
  - stanowisko: **ekspert ds. uczenia maszynowego**,
  - rola: opracowanie algorytmów uczenia maszynowego do badania składu gleby na podstawie satelitarnych danych hiperspektralnych, opracowanie architektury systemu,
  - efekty prac w: [B9, B11, B12, B14],
- [P12] 2019–2022 – Opracowanie nowej odmiany pomidora gruntowego dedykowanego dla przetwórstwa oraz ulepszonej technologii jego uprawy w celu uzyskania wysokiej jakości, powtarzalnej i wydajnej produkcji w polskich warunkach klimatycznych, POIR.01.01.01-00-1317/17, NCBiR,
  - stanowisko: **ekspert ds. uczenia maszynowego**,
  - rola: przygotowanie algorytmów uczenia maszynowego pozwalających na automatyczną identyfikację patogenów, estymacja plonów na podstawie obrazowania multispektralnego,
  - efekty prac w: [A1, A2], [B19, B35].

3.2 Staże badawcze

- 2022–2025:
  - 3 lata stażu w KP Labs sp. z o.o., Gliwice (od 2019 status centrum badawczo-rozwojowego Ministerstwa Rozwoju i Technologii RP),
  - rola: **specjalista ds. uczenia maszynowego** w ramach stażu praca w międzynarodowych projektach badawczych dla Europejskiej Agencji Kosmicznej (ESA): [P4, P6, P7 i P8],
  - opublikowane rezultaty: [A7–A9], [B8, B9, B11, B12, B14, B16, B23, B33],
- 2022–2023:
  - 2 lata stażu w TwinIO Energy sp. z o.o.,
  - rola: **ekspert ds. uczenia maszynowego** w projektach NCBiR [P9 i P10],
  - opublikowane rezultaty: [A10], [B7, B20],
- 2018–2022:
  - 3,5 roku stażu w QZ Solutions sp. z o.o., Opole (od 2019, status centrum badawczo-rozwojowego, Ministerstwa Rozwoju i Technologii RP),
  - rola: **kierownik projektu B+R** w projektach [P2](NCBiR) i [P3],
  - oraz rola: **ekspert ds. uczenia maszynowego** w projektach NCBiR [P11] i [P12],
  - opublikowane rezultaty: [A1–A5], [B10, B19, B28–B30, B35, B36, B38].

3.3 Profesor wizytujący w uczelniach partnerskich

Zadania: wygłoszenie wykładów dotyczących uczenia maszynowego i prowadzonych badań, wymiana doświadczeń dydaktycznych, konsultacje dalszych badań i opracowanie wspólnych publikacji:

- Hellenic Mediterranean University, Chania, Grecja (fundusz Erasmus; 2025, opublikowane artykuły: [B6]),
- Technical University of Cartagena, Hiszpania (fundusz Erasmus; 2023; opublikowane artykuły: [A2], [B1]),
- Hellenic Mediterranean University, Chania, Grecja (fundusz Erasmus; 2022, opublikowane artykuły: [B31, B32]),
- Hellenic Mediterranean University, Chania, Grecja (fundusz Erasmus; 2020, opublikowane artykuły: [B36]),
- University of Almeria, Hiszpania (fundusz Erasmus; 2019),
- Coimbra Institute of Engineering, Portugalia (fundusz Erasmus; 2018),
- Technical University of Cartagena, Hiszpania (fundusz Erasmus; 2017; opublikowane artykuły: [B37]),

- University of Alicante, Hiszpania, (fundusz Erasmus; 2016),
- Technical University of Cartagena, Hiszpania (fundusz Erasmus; 2014),
- Coimbra Institute of Engineering, Portugalia (fundusz Erasmus; 2013),
- Valencia University of Technology, Hiszpania (fundusz Erasmus; 2012),
- Pekin, Chiny (fundusz Instytutu Konfucjusza, 2009).

3.4 Opublikowane zbiory danych

- [D1] 2024 – OPSSAT Anomaly Dataset:
  - zbiór sygnałów telemetrycznych z europejskiego satelity ESA OPS-SAT, wraz ze wzorcowymi opisami anomalii przygotowanymi przez ekspertów, oraz dodatkowo przekształcony do postaci tabelarycznej,
  - opublikowany w `https://zenodo.org/records/12588358` wraz z repozytorium pozwalającym na szybkie wykorzystanie zbioru: `https://github.com/kplabs-pl/OPS-SAT-AD`,
  - publikacja zbioru jest dodatkowym efektem projektu [P8],
  - zbiór komentowany jest bezpośrednio w [A9], jego przygotowanie omówiono w [B23, B33] a wyniki dla przygotowanych z jego użyciem nadzorowanych algorytmów ML omówiono też w [A8],
  - cytowania zbioru – Scopus: 3, Web of Science: 1, Google Scholar: 6,
  - pobrania zbioru – 3275.
- [D2] 2024 – ESA-ADB:
  - zbiór kilkudziesięciu wieloletnich sygnałów telemetrycznych z trzech misji kosmicznych Europejskiej Agencji Kosmicznej, wraz z przygotowanymi przez ekspertów oznaczeniami anomalii,
  - opublikowany w `https://zenodo.org/records/12528696` wraz z repozytorium: `https://github.com/kplabs-pl/ESA-ADB`,
  - publikacja zbioru jest dodatkowym efektem projektu [P7],
  - zbiór komentowany jest bezpośrednio w [B16],
  - cytowania zbioru – Google Scholar: 18,
  - pobrania zbioru – 4407.
- [D3] 2024 – HYPERVIEW:
  - zbiór pomiarów składników odżywczych w glebie i lotniczych obrazów hiperspektralnych (rektyfikowanych ortofotomap oraz wyciętych i skojarzonych z pomiarami fragmentów obrazów), wraz z notatnikami Jupyter,
  - publikacja zbioru jest dodatkowym efektem projektu [P11],
  - zbiór, wraz z opisem najskuteczniejszych modeli ML do estymacji substancji odżywczych w glebie, zostały opisane w [B9],

  - został opublikowany wraz z kodem pozwalającym na obsługę danych w: `https://platform.ai4eo.eu/seeing-beyond-the-visible-permanent`.

- [D4] 2022 – Deep Potato:
  - zbiór obrazów hiperspektralnych w wysokiej rozdzielczości przestrzennej i spektralnej (rektyfikowanych ortofotomap oraz wyciętych i skojarzonych z pomiarami fragmentów obrazów), wraz z bogatymi pomiarami agronomicznymi skojarzonymi z każdym obrazem (m.in. pomiary wilgotności różnymi metodami, zarówno dla roślin jak i gleby, zawartości chlorofilu i innych składników, masy łodyg, liści, bulw), zbiór ten, wraz z notatnikami Jupyter,
  - opublikowany w: `https://data.mendeley.com/datasets/xn2wy75f8m/2`,
  - publikacja zbioru jest dodatkowym efektem projektu [P2],
  - komentowany jest bezpośrednio w [B29],
  - cytowania zbioru – Scopus: 12, Web of Science: 9,
  - pobrania zbioru – 608,
  - wyniki modeli ML opracowanych z użyciem tego zbioru: [A4], [B10, B24, B28, B30].

- [D5] 2018 – „The collection of images of an insulator taken outdoors in varying lighting conditions with additional laser spots“:
  - zbiór obrazów izolatorów energetycznych w wysokiej rozdzielczości wraz z oznaczeniami obiektów oraz precyzyjnymi pomiarami luminancji,
  - wyniki modeli opracowanych z użyciem tego zbioru: [B40],
  - omówienie zbioru i niezbędne do pobrania go linki zamieszczono w: [B44] (cytowania Scopus: 28, Web of Science: 22),
  - cytowania zbioru – Scopus: 10, Web of Science: 8.

3.5 Nagrody

- 2025 – nominacja i nagroda **Best Paper** w: Evolutionary Combinatorial Optimisation and Metaheuristic (ECOM) track na konferencji **Conference on Genetic and Evolutionary Computation, ACM GECCO**, Malaga, Hiszpania, ranking konferencji CORE A, za artykuł [B2],

- 2024 – Nagroda Projakościowa JM Rektora Politechniki Opolskiej za rok 2023, (nagroda za osiągnięcia publikacyjne),

- 2023 – Nagroda Projakościowa JM Rektora Politechniki Opolskiej za rok 2022, (nagroda za osiągnięcia publikacyjne),

- 2022 – Nagroda Projakościowa JM Rektora Politechniki Opolskiej za rok 2021,

- 2021 – Nagroda Projakościowa JM Rektora Politechniki Opolskiej za rok 2020,

- 2020 – Nagroda Projakościowa JM Rektora Politechniki Opolskiej za rok 2019,

- 2015 – Podziękowania Dyrektor Koordynacji Programów Operacyjnych Urzędu Marszałkowskiego Województwa Opolskiego za efekty projektu i wdrożone innowacje wypracowane w projekcie „Efektywny Transfer Wiedzy z nauki do przemysłu w województwie opolskim”,
- 2013 – Nagroda JM Rektora Politechniki Opolskiej, nagroda zespołowa III stopnia za osiągnięcia organizacyjne,

3.6 Recenzje i prace redakcyjne

- 56 recenzji w indeksowanych czasopismach naukowych i 13 recenzji na międzynarodowych konferencjach naukowych, w tym:
  - Engineering Applications of Artificial Intelligence, Elsevier (9 recenzji),
  - Applied Sciences, MDPI (9 recenzji),
  - Sensors, MDPI (6 recenzje),
  - Remote Sensing, MDPI (4 recenzje),
  - Expert Systems With Applications, Elsevier (3 recenzje),
  - Scientific Reports, Nature (3 recenzje),
  - IEEE Access, IEEE (3 recenzje),
  - Journal of Imaging, MDPI (2 recenzje),
  - Electronics, MDPI (2 recenzje),
  - Applied Soft Computing Journal, Elsevier (1 recenzja),
  - Data in Brief, Elsevier (1 recenzja),
  - IEEE Transactions on Power Delivery, IEEE (1 recenzja),
  - International Journal of Information Technology and Decision Making, WS (1 recenzja),
  - Journal of Telecommunications and Information Technology, NIT (1 recenzja),
  - Scientific Data, Nature (1 recenzja),
  - ISPRS Journal of Photogrammetry and Remote Sensing, Elsevier (1 recenzja),
  - International Journal of Applied Mathematics & Computer Science, AMCS (1 recenzja),
  - Signal Processing: Image Communication, Elsevier (1 recenzja),
  - Astrodynamics, Springer (1 recenzja),
  - Advances in Space Research, Elsevier (1 recenzja),
  - Frontiers in Oncology, Frontiers (1 recenzja),
  - Future Internet, MDPI (1 recenzja),
  - Horticulturae, MDPI (1 recenzja),
  - Information, MDPI (1 recenzja),
  - konferencji European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) w 2025 roku (2 recenzje),

  - konferencyjnego warsztatu Machine Learning for Irregular Time Series Workshop na ECML PKDD w 2025 (2 recenzje),
  - konferencyjnego warsztatu Explainable AI in Space (EASi) Workshop na European Conference on Artificial Intelligence (ECAI) w 2025 (5 recenzji),
  - konferencji PP-RAI w 2025 roku (2 recenzje),
  - konferencji IEEE Visual Communications and Image Processing w 2024 roku (2 recenzje),

- sekretarz redakcji czasopisma Przegląd Nauk Stosowanych (dot. numerów: 13/2016, 12/2016, 11/2016, 10/2016).

3.7 Praca z dyplomantami

- praca z doktorantami
  - **promotor pomocniczy** w przewodzie doktorskim **dr. inż. Krzysztofa Smykały** (doktorat obroniony w 2024 roku w dyscyplinie automatyka, elektronika, elektrotechnika i technologie kosmiczne, na Politechnice Opolskiej, promotor dr hab. Michał Tomaszewski). Zadania jako promotora pomocniczego: opracowanie pierwotnego badania, koncepcji przygotowania eksperymentów i danych, współpraca przy akwizycji danych, przygotowaniu zbioru danych, modeli uczenia maszynowego, opracowaniu artykułów przedstawiających wyniki badania. Do efektów wspólnych prac z doktorantem należą m.in.: [A1, A2], [B9, B11, B35, B38],
  - **promotor pomocniczy** w przewodzie doktorskim **mgr inż. Sławomira Szymochy** (promotor dr hab. Michał Tomaszewski). Praca nie jest jeszcze ukończona, dotyczy zastosowania bezzałogowego statku powietrznego do monitorowania i diagnozowania stanu zanieczyszczenia powietrza. Zadania jako promotora pomocniczego: współpraca przy opracowaniu planu badawczego i uczestnictwo w badaniach eksperymentalnych, pomoc merytoryczna w opracowaniu i interpretacji wyników pomiarów, pomoc w opracowaniu algorytmów i metod badawczych w zakresie przetwarzania uzyskiwanych pomiarów.
  - **recenzent** w przewodzie doktorskim **mgr. Jaime Giménez Gallego** (promotor Prof Dr. Roque Torres Sánchez) pt. Sensors and methodologies design for image processing in precision agriculture: crop water status and fruit ripening monitoring, która obroniona została w 2024 roku w Universidad Politécnica de Cartagena.

- opieka nad przygotowaniem prac dyplomowych:
  - **3 prac magisterskich z Informatyki** na Wydziale Informatyki Politechniki Opolskiej,
    – do dodatkowych efektów prac magisterskich należą publikacje: [B3, B4, B6],
  - **11 prac magisterskich z Informatyki** na Wydziale Elektrotechniki, Automatyki i Informatyki Politechniki Opolskiej,
  - **11 prac inżynierskich z Informatyki** na Wydziale Elektrotechniki, Automatyki i Informatyki Politechniki Opolskiej,

- **9 prac licencjackich z Zarządzania** na Wydziale Ekonomii i Zarządzania Politechniki Opolskiej,
- **3 recenzje prac dyplomowych**: **inżynierskich** (2) i **magisterskiej** (1) z Informatyki na Wydziale Elektrotechniki, Automatyki i Informatyki Politechniki Opolskiej.

3.8 Osiągnięcia dydaktyczne

- w ramach ciągłego zatrudnienia na Politechnice Opolskiej (2005 – obecnie) prowadzenie zajęć dydaktycznych, opracowanie materiałów dydaktycznych oraz dokumentacji dydaktycznej dla przedmiotów:
  - od roku 2018 (na Wydziale Elektrotechniki, Automatyki i Informatyki Politechniki Opolskiej):
    * *Zaawansowane systemy baz danych* (**Informatyka**, wykład i laboratoria),
    * *Zastosowania sztucznej inteligencji* (**Informatyka**, wykład i laboratoria),
    * *Technologie projektowania baz danych* (**Informatyka**, wykład i laboratoria),
    * *Systemy ekspertowe w automatyce* (**Automatyka i robotyka**, wykład i laboratoria),
  - przed rokiem 2018 (na Wydziale Ekonomii i Zarządzania Politechniki Opolskiej):
    * *Technologia informacyjna* (wykład i laboratoria),
    * *Analiza ryzyka* (wykład i laboratoria),
    * *Podstawy baz danych* (laboratoria),
    * *Wybrane języki programowania* (laboratoria).

- wyniki ankiet studenckich do prowadzonych zajęć (równa maksymalnie 10,00 średnia ważona ocen wystawianych przez studentów za przygotowanie prowadzącego i za prowadzone zajęcia, liczona osobno w semestrach zimowych i letnich):
  - 2023/24Z – prowadzący: **10,00**, zajęcia: **10,00**,
  - 2022/23L – prowadzący: **9,86**, zajęcia: **9,71**,
  - 2022/23Z – prowadzący: **9,90**, zajęcia: **9,90**,
  - 2021/22L – prowadzący: **10,00**, zajęcia: **9,71**,
  - 2021/22Z – prowadzący: **10,00**, zajęcia: **9,56**,
  - 2020/21L – prowadzący: **10,00**, zajęcia: **9,69**,
  - 2020/21Z – prowadzący: **10,00**, zajęcia: **10,00**,
  - 2019/20L – prowadzący: **9,23**, zajęcia: **9,30**,
  - 2018/19L – prowadzący: **10,00**, zajęcia: **8,49**,
  - 2018/19Z – prowadzący: **9,11**, zajęcia: **9,16**,
  - 2017/18L – prowadzący: **10,00**, zajęcia: **10,00**,

  - 2017/18Z – prowadzący: **9,82**, zajęcia: **9,68**.

- 2023 – obecnie: członek rady dydaktycznej kierunku Analityka danych w biznesie, prowadzonego na Wydziale Elektrotechniki, Automatyki i Informatyki Politechniki Opolskiej,

- opracowanie i ewaluacja sylabusów dla przedmiotów:

  - *Zaawansowane systemy baz danych* (kierunek: Informatyka),
  - *Zastosowania sztucznej inteligencji* (Informatyka),
  - *Analiza danych przestrzennych* (Informatyka),
  - *Przetwarzanie danych w chmurze* (Informatyka),
  - *Geoinformatyka* (Informatyka),
  - *Inżynieria obliczeniowa* (Informatyka),
  - *Przetwarzanie obrazów* (Analityka danych w biznesie),
  - *Przetwarzanie języka naturalnego* (Analityka danych w biznesie),
  - *Zaawansowana analiza szeregów czasowych* (Analityka danych w biznesie).

3.9 Inne osiągnięcia organizacyjne

- 2025 – członek komitetu programowego: Machine Learning for Irregular Time Series (ML4ITS2025): Advances in Generative Models, Global Models and Self-Supervised Learning, na konferencji: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML PKDD, Porto, Portugalia, 2025, ranking konferencji CORE A,

- 2025 – naukowy i techniczny współorganizator: workshop Explainable AI in Space (EASi) na konferencji: European Conference on Artificial Intelligence (ECAI), Bologna, Włochy, ranking konferencji CORE A,

- 2025 – współorganizator konkursu: Spacecraft Anomaly Challenge on ESA dataset na platformie Kaggle (www.kaggle.com/competitions/esa-adb-challenge),

- 2025 – współorganizator hackathonu w European Space Operations Centre, centrum kontroli misji Europejskiej Agencji Kosmicznej, Darmstadt, 13 marca 2025, (53 uczestników, 13 zespołów),

- 2025 – współorganizator konkursu: HYPERVIEW2 challenge na platformie Europejskiej Agencji Kosmicznej (www.ai4eo.eu/portfolio/easi-workshop-hyperview2),

- 2025 – członek komitetu programowego: Workshop on Artificial intelligence and machine learning for Emerging Topics (ALLEGET), na konferencji: 21st International Conference on Intelligent Environments (IE), Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Niemcy, 2025, ranking konferencji CORE B,

- 2016–2018 – członek Rady Wydziału Ekonomii i Zarządzania Politechniki Opolskiej,

- 2015–2018 – wydziałowy koordynator ds. planów zajęć – koordynacja pracy planistów, synchronizacja planów z innymi wydziałami, układanie planów,
- 2006–2014 – planista wydziałowy – układanie planów zajęć,
- 2006–2017 – członek komitetu organizacyjnego 6 ogólnokrajowych konferencji naukowych: w latach 2014 – 2017: XIX, XVIII, XVII i XVI ogólnokrajowej konferencji naukowej z cyklu *Rozwój Regionalny*, oraz w latach 2006 i 2007: IX i X konferencji *Komputerowo Zintegrowane Zarządzanie* w Zakopanem.

3.10 Uczestnictwo i wygłoszenie artykułu w międzynarodowych i krajowych konferencjach naukowych

- Genetic and Evolutionary Computation Conference (GECCO), Malaga, Hiszpania, 2025, [A6], [ranking konferencji CORE A]:
  - nagroda "Best Paper" w: Evolutionary Combinatorial Optimisation and Metaheuristic (ECOM) track [B2],
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML PKDD, Porto, Portugalia, 2025, [B3], [ranking konferencji CORE A]:
  - prowadzenie sesji *Time Series Applications* (Session Chair).
  - współorganizator *Machine Learning for Irregular Time Series Workshop*,
- 21st International Conference on Intelligent Environments (IE), Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Niemcy 2025, [B4], [ranking konferencji CORE B]:
  - Invited Keynote talk – podczas: ALLEGET – Workshop on Artificial intelligence and machine learning for Emerging Topics, [B5],
- 6th Polish Conference on Artificial Intelligence (PP-RAI), Uniwersytet Śląski, Katowice, 2025,
- 6th International Conference In Electronic Engineering, Information Technology & Education (EEITE), Chania, Grecja, 2025 [B6],
- XI Spotkanie polskiej grupy badawczej systemów uczących się: Uczenie maszynowe na potrzeby wizji komputerowej i robotyki (Polish-SIGML), Zachodniopomrski Uniwersytet Technologiczny, Szczecin, 2025,
- Genetic and Evolutionary Computation Conference (GECCO), Melbourne, Australia, 2024, [A10] [B10], [ranking konferencji CORE A],
- IEEE Visual Communications and Image Processing (VCIP), Tokio, Japonia, 2024, [B13], [ranking konferencji CORE C],
- X Spotkanie polskiej grupy badawczej systemów uczących się: Wyzwania uczenia maszynowego i głębokiego: eksploracja możliwości w medycynie i automatyce przemysłowej (Polish-SIGML), AGH, Kraków, 2024,

- 5th Polish Conference on Artificial Intelligence (PP-RAI), Politechnika Warszawska, Warszawa, 2024, [B14],
- Genetic and Evolutionary Computation Conference (GECCO), Lizbona, Portugalia, 2023, [A3] [B20] , [ranking konferencji CORE A],
- 23rd International Conference on Computational Science (ICCS), Praga, Czechy, 2023, [A5, A8], [ranking konferencji CORE A],
- 18th Conference on Computer Science and Intelligence Systems (FedCSIS), (IEEE #57573), Politechnika Warszawska, Warszawa, 17–20.9, 2023, [B21],
- IX spotkanie polskiej grupy badawczej systemów uczących się: Interpretacja i wyjaśnialność uczenia maszynowego (Polish-SIGML), Uniwersytet Ekonomiczny, Katowice, 23–24.10.2023,
- 3rd International Symposium in Electronic Engineering, Information Technology & Education (EEITE), Chania, Grecja, 2022, [B31, B32],
- 2nd International Symposium in Electronic Engineering, Information Technology & Education (EEITE), Chania, Grecja, 2020, [B36],
- 16th International Conference on Intelligent Environments (IE), Madryt, Hiszpania, 2020, [B35], [ranking konferencji CORE C],
- 40th International Conference on Information Systems Architecture and Technology (ISAT), Wrocław, 2019, [B37],
- 3rd International Scientific Conference on Brain-Computer Interfaces (BCI), Opole, 2018, [B43],
- 39th International Conference on Information Systems Architecture and Technology (ISAT), Wrocław, 2018, [B41],
- Energy Finance Christmas Workshop (EFC11), Politechnika Wrocławska, Wrocław, 2011.

3.11 Artykuły opublikowane na konferencjach

- 18th International Conference on Space Operations (SpaceOps), Montreal, Kanada, 2025 [B8],
- IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Ateny, Grecja, 2024, [B11, B12], [ranking konferencji CORE C],
- Progress in Applied Electrical Engineering (PAEE), IEEE , Kościelisko, 2024, [B15],
- 2023 Conference on Big Data from Space (BiDS), Wiedeń, Austria, 2023, [B22],
- 74th International Astronautical Congress (IAC), Baku, Azerbejdżan, 2023, [B23, B24],

- 8th International Workshop on On-board Payload Data Compression (OBPDC), Ateny, Grecja, 2022, [B33],
- 18th International Conference on Artificial Intelligence and Soft Computing (ICAISC), Zakopane, 2019, [B38],
- International Symposium on Electrodynamic and Mechatronic Systems (SELM) Opole, 2013, [B54].

3.12 Członkostwo w organizacjach

- The Association for Computing Machinery (ACM Membership No.: 6133841),
- The Institute of Electrical and Electronics Engineers (IEEE Membership ID 93050247), członek grupy GRSS (Geoscience and Remote Sensing Society),
- ECML PKDD Association,
- członek Polskiej Grupy Badawczej Systemów Uczących się – SIGML,
- ekspert na liście Narodowego Centrum Badań i Rozwoju, NCBiR.

Literatura

[1] D. R. Hammou i M. Boubaker, "Tomato plant disease detection and classification using convolutional neural network architectures technologies," w *Networking, Intelligent Systems and Security*. Springer, 2022, s. 33–44.

[2] A. R. Al-Shamasneh i R. W. Ibrahim, "Classification of tomato leaf images for detection of plant disease using conformable polynomials image features," *MethodsX*, tom 13, s. 1–13, 2024.

[3] A. K. Mahlein, T. Rumpf, P. Welke, H. W. Dehne, L. Plümer, U. Steiner, i E. C. Oerke, "Development of spectral indices for detecting and identifying plant diseases," *Remote Sensing of Environment*, tom 128, s. 21–30, 2013.

[4] D. Ashourloo, A. A. Matkan, A. Huete, H. Aghighi, i M. R. Mobasheri, "Developing an index for detection and identification of disease stages," *IEEE Geoscience and Remote Sensing Letters*, tom 13, no. 6, s. 851–855, 2016.

[5] J. Lu, R. Ehsani, Y. Shi, A. I. de Castro, i S. Wang, "Detection of multi-tomato leaf diseases (late blight, target and bacterial spots) in different stages by using a spectral-based sensor," *Scientific Reports*, tom 8, no. 1, p. 2793, dec 2018.

[6] C. Jones, J. Jones, i W. Lee, "Diagnosis of bacterial spot of tomato using spectral signatures," *Computers and Electronics in Agriculture*, tom 74, no. 2, s. 329–335, nov 2010.

[7] Y. Cen, Y. Huang, S. Hu, L. Zhang, i J. Zhang, "Early Detection of Bacterial Wilt in Tomato with Portable Hyperspectral Spectrometer," *Remote Sensing*, tom 14, no. 12, p. 2882, jun 2022.

[8] S. Appeltans, J. G. Pieters, i A. M. Mouazen, "Potential of laboratory hyperspectral data for in-field detection of phytophthora infestans on potato," *Precision Agriculture*, tom 23, no. 3, s. 876–893, Jun 2022.

[9] J. Fahrentrapp, F. Ria, M. Geilhausen, i B. Panassiti, "Detection of Gray Mold Leaf Infections Prior to Visual Symptom Appearance Using a Five-Band Multispectral Sensor," *Frontiers in Plant Science*, tom 10, no. May, s. 1–14, may 2019.

[10] J. Fahrentrapp, F. Ria, M. Geilhausen, i B. Panassiti, "Detection of gray mold leaf infections prior to visual symptom appearance using a five-band multispectral sensor," *Frontiers in Plant Science*, tom 10, 2019.

[11] E. Bauriegel, A. Giebel, M. Geyer, U. Schmidt, i W. Herppich, "Early detection of Fusarium infection in wheat using hyper-spectral imaging," *Computers and Electronics in Agriculture*, tom 75, no. 2, s. 304–312, feb 2011.

[12] M. Haagsma, C. H. Hagerty, D. R. Kroese, i J. S. Selker, "Detection of soil-borne wheat mosaic virus using hyperspectral imaging: from lab to field scans and from hyperspectral to multispectral data," *Precision Agriculture*, tom 24, no. 3, s. 1030–1048, Jun 2023.

[13] D. Ashourloo, A. A. Matkan, A. Huete, H. Aghighi, i M. R. Mobasheri, "Developing an Index for Detection and Identification of Disease Stages," *IEEE Geoscience and Remote Sensing Letters*, tom 13, no. 6, s. 851–855, 2016.

[14] L. Breiman, J. H. Friedman, R. A. Olshen, i C. J. Stone, *Classification and Regression Trees*. New York: Chapman and Hall/CRC, 1984.

[15] L. Breiman, "Random forests," *Machine Learning*, tom 45, no. 1, s. 5–32, Oct 2001.

[16] T. Hastie, S. Rosset, J. Zhu, i H. Zou, "Multi-class adaboost," *Statistics and Its Interface*, tom 2, no. 3, s. 349–360, 2009.

[17] P. Geurts, D. Ernst, i L. Wehenkel, "Extremely randomized trees," *Machine Learning*, tom 63, no. 1, s. 3–42, Apr 2006.

[18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, tom 29, no. 5, s. 1189–1232, 2001.

[19] R. W. Schafer, "What is a Savitzky-Golay filter?" *IEEE Signal Processing Magazine*, tom 28, no. 4, s. 111–117, 2011.

[20] S. Luke, "Essentials of metaheuristics," s. 1–263, 2013, available at https://cs.gmu.edu/ sean/book/metaheuristics/.

[21] A. Wook i R. Ramakrishna, "Elitism-based compact genetic algorithms," *IEEE Transactions on Evolutionary Computation*, tom 7, no. 4, s. 367–385, 2003.

[22] K. Deb, A. Pratap, S. Agarwal, i T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, tom 6, no. 2, s. 182–197, 2002.

[23] A. Carvalho i A. Araujo, "Improving NSGA-II with an adaptive mutation operator," w *Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 2009, p. 2697–2700.

[24] L. Wang, X. Zhang, H. Su, i J. Zhu, "A comprehensive survey of Continual Learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, tom 46, s. 5362–5383, 2023.

[25] Y. Song, T. Wang, P. Cai, S. K. Mondal, i J. P. Sahoo, "A comprehensive survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities," *ACM Computing Surveys*, tom 55, no. 13s, Jul. 2023.

[26] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K. R. Müller, i M. Kloft, "Deep semi-supervised anomaly detection," w *International Conference on Learning Representations (ICLR)*, 2020, s. 1–23.

[27] G. Pang, C. Shen, i A. van den Hengel, "Deep anomaly detection with deviation networks," w *25th International Conference on Knowledge Discovery & Data Mining (ACM SIGKDD)*. ACM, 2019, s. 353–362.

[28] Y. Chen, Y. Kang, Y. Chen, i Z. Wang, "Probabilistic forecasting with temporal convolutional neural network," *Neurocomputing*, s. 491–501, 2020.

[29] A. Mohammadi, G. Nápoles, i Y. Salgueiro, "Deep learning for time-series forecasting with exogenous variables in energy consumption: A performance and interpretability analysis," *IEEE Access*, tom 13, s. 86 746–86 767, 2025.

[30] C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," *Journal of the Operational Research Society*, tom 66, no. 8, s. 1352–1362, Aug 2015.

[31] K. G. Olivares, C. Challu, G. Marcjasz, R. Weron, i A. Dubrawski, "Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx," *International Journal of Forecasting*, tom 39, no. 2, s. 884–900, 2023.

[32] Y. Akimoto i N. Hansen, "CMA-ES and advanced adaptation mechanisms," w *Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 2022, p. 1243–1268.

[33] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, i R. Fergus, "Regularization of neural networks using dropconnect," w *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta i D. McAllester, Red., tom 28, no. 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, s. 1058–1066.

[34] D. W. Blalock, J. J. G. Ortiz, J. Frankle, i J. V. Guttag, "What is the state of neural network pruning?" w *Conference on Machine Learning and Systems*, I. S. Dhillon, D. S. Papailiopoulos, i V. Sze, Red. MLSys, 2020.

[35] Y. Ozaki, Y. Tanigaki, S. Watanabe, i M. Onishi, "Multiobjective tree-structured parzen estimator for computationally expensive optimization problems," w *Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 2020, p. 533–541.

[36] F. T. Liu, K. M. Ting, i Z. H. Zhou, "Isolation Forest," w *2008 8th IEEE International Conference on Data Mining (ICDM)*. IEEE, 2008, s. 413–422.

[37] Y. Zhao i M. K. Hryniewicki, "XGBOD: Improving supervised outlier detection with unsupervised representation learning," w *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, s. 1–8.

A Załączniki

Załączniki od 1 do 10 zawierają kopie prac stanowiących osiągnięcie naukowe.

Załącznik 11 zawiera oświadczenia współautorów dotyczące merytorycznego wkładu w powstanie cyklu publikacji.

Załącznik 1 – [A1]

Journal of Ambient Intelligence and Smart Environments 12 (2020) 407–418
DOI 10.3233/AIS-200573
IOS Press

The detection of *Alternaria solani* infection on tomatoes using ensemble learning

Bogdan Ruszczak [a,b,*], Krzysztof Smykała [a,b] and Karol Dziubański [a]
[a] *QZ Solutions Sp. z o.o., Ozimska 72A Street, 45-310 Opole, Poland*
[b] *Faculty of Electrical Engineering Automatic Control and Informatics, Opole University of Technology, Prószkowska 76 Street, 45-758 Opole, Poland*

Abstract. This paper presents a detection method of *Alternaria solani* in tomatoes. Several machine learning models were used to detect the pathogen, such as the implementation of decision trees and ensemble learning methods. The use of these methods requires the acquisition of large volumes of data and adequate preprocessing of this data. For the presented study the dataset of hyperspectral measurements of two varieties of tomatoes was used. Measurements were split into two groups: one inoculated with the *Alternaria solani* pathogen and the other one was treated as the reference. Measurements were taken by the spectroradiometer in consecutive measurement series. The main part of the study was the evaluation of the decision trees and the popular ensemble learning algorithms to select the most accurate one. After subsequent iterations of the training process and adjustment of hyperparameters, satisfactory accuracy results, equal to 0.987 for random forest, were obtained. This paper also covers the examination of the spectral range required for *Alternaria solani* identification. From several variants, the accuracy of models based on VIS and NIR spectral range was the closest to the accuracy obtained with the whole spectrum of measured absolute reflectance.

Keywords: *Alternaria solani*, plant disease detection, hyperspectral data, random forest, decision tree, machine learning, ensemble learning

AUTHOR COPY

1. Introduction

Tomatoes (*Solanum lycopersicum*), a native species from South America, are one of the most grown vegetables in the world. It can be grown in various climate conditions from temperate to tropical. Tomatoes are the second most-consumed vegetable in the world [9]. Its annual worlds' supply exceeds 160 million tons. China is the largest producer of tomatoes, ahead of India and the United States with a production of 59 million tons [8]. Tomatoes are a source of well-absorbed vitamin C and A [4].

However, tomatoes are very vulnerable to fungal, bacterial, and viral diseases such as late blight, powdery mildew, or early blight. *Alternaria solani*, the cause of early blight, is commonly referred to as the most dangerous tomato pathogen [21]. Symptoms of the early blight are pathogenic changes on the leaves, stems, petioles, and fruits. Under favorable conditions, this disease may result in premature leaf drying and fruits fall, causing 30–80% losses in crop yield [6,11]. Various methods are used to control *Alternaria solani* on tomatoes e.g. fungicide control and planting plants partially resistant to early blight. However, under a high disease pressure, these methods are not sufficient for the effective protection of tomato plants [18].

Currently, the diagnosis of this disease is based on the assessment of the symptoms of plant tissue damage or assessing the number of pathogen spores under a microscope. The main limitation of such an approach is that it requires individual plant inspection, including the presence of qualified personnel. Therefore, it is necessary to find fast and reliable methods for assessing the degree of damage to plants that could be later used remotely [3]. The main objective of the present study aims to show the possibilities of using ensemble

*Corresponding author. E-mail: b.ruszczak@po.edu.pl.

1876-1364/20/$35.00 © 2020 – IOS Press and the authors. All rights reserved

learning methods for disease classification and crop non-destructive diagnosis. The other desirable factor is an early warning for the occurrence of the pathogen, but the present methodology does not provide such a feature. The information on suspected upcoming infection could let the farmers plan proper preventative measures, such as precise fungicide management.

The paper presents a detection method of *Alternaria solani* in tomatoes cultivated under foil tunnels. Hyperspectral measurements and ensemble learning methods were investigated for the detection purpose. The spectral range of measurements was 350–2500 nm and that allowed one to carry out the comparison of several different algorithms configurations with different spectral band sets. The last part of the described study was aimed to select the best performing detectors for the analyzed scenarios.

2. Background and related work

The issue of detecting plant diseases attracts the interest of many researchers, which is reflected in numerous publications [2,3,10,15,16,20,26]. This is also the case of research on tomato diseases [2,10,15,16,26]. Publications describe solutions based on standard digital camera (RGB color model) [2,20], multispectral imagery [7] and hyperspectral imagery [16,24–26]. Many authors propose solutions based on hyperspectral measurements with a spectrometer invisible (400–765 nm), near-infrared (900 ÷ 1700 nm), short wave (1000–2500 nm), and long-wave infrared wavelengths (2500–5000 nm), especially in the 350–2500 nm range [15,24,27]. Also, vegetation indices can be used for detecting the early stages of plant diseases [1,5,15,19].

The use of hyperspectral data and machine learning algorithms in disease detection appears to be a popular trend in research [10,24,25]. Research conducted by Moghadam et al. suggests that the analysis of hyperspectral data from 400–2500 nm spectrum gives better results than using various indices. 12 indices were taken to the research: NDVI, SR 800/670, SIPI, PSSRa 800/680, PSSRb 800/635, ARI, REP, mCAI, WBI, CAI, NDLI, NDNI. Researchers applied the Support Vector Machine with Radial Basis Function kernel [16]. They reported a final, best-performing model, using full-spectrum RBF SVM, with 0.936 and 0.915 accuracy on VNIR and SWIR imagery, which were higher than the same model used by Zhu et al. [28] who stated 0.817 accuracy with a full spectrum feature vector.

Xie et al. proposed to use feature ranking (FR) for selecting the most important bands, and then use kNN to classify the plant disease [26]. In his further research, Xie indicates the usefulness of using extreme learning machine (ELM) on the entire spectrum 380–1023 nm to classify early blight and late blight on tomatoes [25]. In the same research, Xie shows usage of successive projections algorithm (SPA) for a decreasing number of bands needed for classifications from 477 to 5 bands. As they reported such reduction allowed the process simplification.

To detect late blight in tomato crops, Wang et al. used an artificial neural network with backpropagation (BPNN) [24]. The 3-25-9-1 architecture was used in the study. Data was acquired by spectrometer GER-2600 (400–2500 nm of spectral range) and aerial hyperspectral imagery (AVIRIS). The authors reported that such a shallow neural network architecture provided proper accuracy with their dataset analysis.

Van De Vijver et al. proposed using PLS-DA method (partial least squares – discriminant analysis) and SIMCA (soft independent modeling of class analogy) [23], as well as PLS-DA and SVM based on PCA scores [22] to detect *Alternaria solani* in potato crops. Field research was conducted in conditions similar to natural conditions with a special measuring platform that limits the inflow of external light.

Despite various publications on the usage of machine learning algorithms for analysis of hyperspectral data for plant disease detection, there is a lack of research covering the usage of these methods and data in *Alternaria solani* detection for Benito and Polfast varieties crops growing under foil tunnels in natural conditions. Admittedly, Xie et al. published research on detecting *Alternaria solani* in tomato crops, however in his research ELM algorithm was used, and the dataset covers hyperspectral images in spectral range 380–1023 nm, taken from one variety (Zheza 809) of tomatoes grown in laboratory conditions [25]. For the comparison the present study covers the random forest algorithm usage on dataset consist of hyperspectral reflectance measurements in spectral range 350–2500 nm, taken from two varieties (Benito and Polfast) of tomatos grown in field conditions (foil tunnels).

The automated applications of machine learning techniques are entering the agriculture industry [13], and among them, the detection of tomato diseases seems to be an important task and is subject to plenty of studies. The intense usage of sensors [19], cameras, and UAVs [5] for that area is also increasing. Several methods already implement computer vision methods

AUTHOR COPY

for tomato disease detection. The presented work is dedicated to collect information on how to better solve that issue, using an automated method, that could have the potential to address an early infection identification.

3. Methodology

The methodology of the research includes data collection from experimental fields with tomato crops and the data analysis description with detailed information on applied algorithms and the configuration of the conducted experiments.

3.1. Data collection

Data used for this study was acquired in experimental fields in Poznan (Greater Poland Voivodeship, Poland) by QZ Solutions for a project co-founded by The National Center for Research and Development (POIR.01.01.01-00-1317/17).

During the project, the observations of pathogens affecting two tomato varieties were performed on fields. For the following experiment, two fields (with those different varieties) were split into two sections: one inoculated with the *Alternaria solani* pathogen and the other treated as the reference (control). The reason for such split was to enable the measurement labeling process (the measurements that refer to the infected plants were labeled as *Alternaria*, the rest were marked as control), aimed to distinguish healthy plants and those with progressing infection. In Fig. 1 the plant and picked leaf are presented. The *Alternaria solani* symptoms can be spotted on the bottom image.

The whole process – from plant preparation and pathogen inoculation, through data collection and preprocessing, to the final modeling and detection has been depicted in Fig. 2. The leaves collected from the inoculated plants were later scanned separately, in a different session, to avoid the infection transmission. The measurements were described as healthy or infected as they were carried out in separate sessions, and that resulted in a precise category assignment for the final dataset.

Measurements were taken by spectrometer ASD FieldSpec 4 Hi-Res. For the purpose of the study, two tomato varieties were cultivated: Benito and Polfast. The reflectance was measured of the collected leaves. The leaves were scanned just after picking with no additional storing process, in separate sessions for



Fig. 1. The tomato plant infected with *Alternaria solani*: under a foil tunnel (top), and a leaf collected for measurement (bottom).

healthy and inoculated plants, using the same setup – calibrated spectrophotometer, fixed scanning height (height of 5 cm resulting in 2.22 cm diameter of a leaf area and spectrometer field of view of 25 degrees) and calibrated halogen lighting set. Tomatoes were planted on experimental fields under foil tunnels and the experimental setup is presented in Table 1.

In order to stabilize the measurement system, ASD FieldSpec 4 Hi-Res was turned on 30 minutes before the data acquisition. Then the spectrometer was calibrated with white reference and the calibration was repeated after every series of measurements. Leaves were collected briefly before scanning to prevent any unexpected spectra changes.

The spectrophotometer during the whole process has been powered up each time at least 30 minutes before calibration and measurements. The integration time, that is essential to get the proper reflectance readouts has been set to maximize the spectrometer radiometric range. Each single recorded measurement was a result of five different spectrum reads.

The spectrometer calibration process has been carried out using the white reference spectralon as well as the dark current calibration, before each measurement session, and the reference spectral calibration curve was also recorded in separate files.

AUTHOR COPY

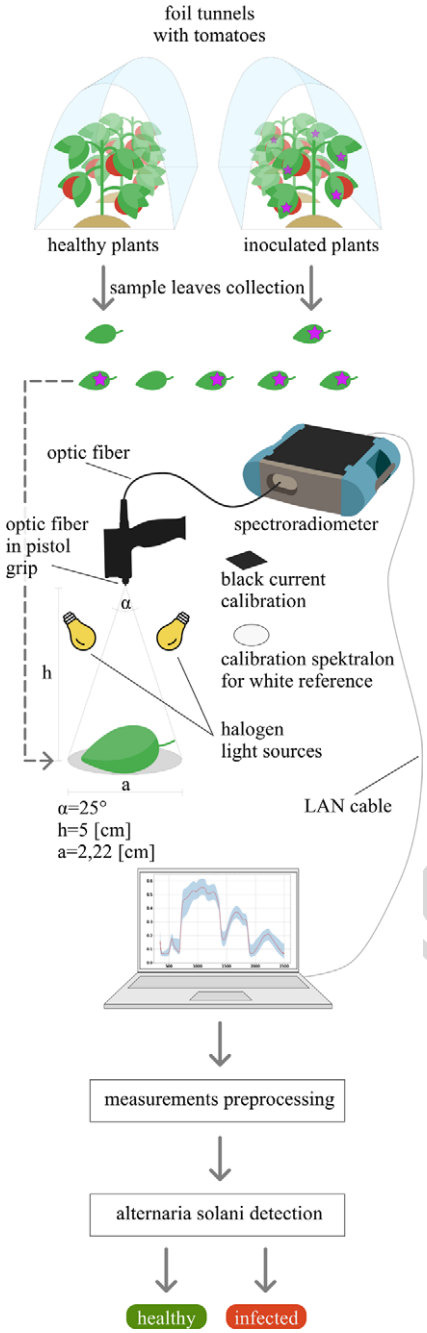


Fig. 2. The process of plants preparation, data collection, postprocessing, and final pathogen detection.

The pathogen has been marked on each leaf to avoid mistaken labeling. Each leaf has been pictured before collection, and later transported to the scanning scene.

Each measurement has been followed by the macroscopic analysis, pathogen isolation, and specie identification. The infected parts of the plant that has been carefully separated were fumigated, using a 2% *Sodium hypochlorite* solution in 3 minutes. Later the fragments were desiccated, cut into four smaller pieces, and placed into Potato Dextrose Agar (PDA) to grow the pathogen. After seven to 10 days the grow cultures has been taken into examination to assign the pathogen and to picture the results.

Table 1
The experiment setup of plants observed for the study

Field ID	Tomato variety name	Pathogen name
6	Benito	Control
18	Polfast	Control
1	Benito	*Alternaria solani*
13	Polfast	*Alternaria solani*

The dataset contains hyperspectral measurements of leaf reflectance for visible, near-infrared, and shortwave infrared wavelengths (350–2500 nm). The spectral resolution of the data is 3 nm at 700 nm and 8 nm at 1400/2100 nm. Both leaves with and without background were included in the dataset. Figure 3 depicts different measurements split into four subsets. As can easily be seen, the consecutive measurements have various spectral characteristics and the distinction of healthy and infected plants may not be obvious. A different perspective of the observed hyperspectral leaves characteristics is presented in Fig. 4 where the specific peaks and lows related to the investigated phenomena can be compared.

3.2. *Random forest*

Random forest is one of the most popular ensemble learning methods. It consists of multiple decision trees. Each decision tree is a hierarchy of some true-or-false statements leading to the decision and the trees are different from each other [12]. Random forest, as an ensemble of decision trees, shares all benefits of decision trees from one hand and making up for some of their deficiencies [12].

Random forest maintains high accuracy on training data and generalization accuracy [14]. This method can be used for regression and classification problems as well. The main flaw of the random forest use is that its hyperparameters need to be tuned to reach high performance and the algorithm does not accept the missing data. Fortunately, the application of spectrophotometer measurement results with a dataset of proper quality.

COPY

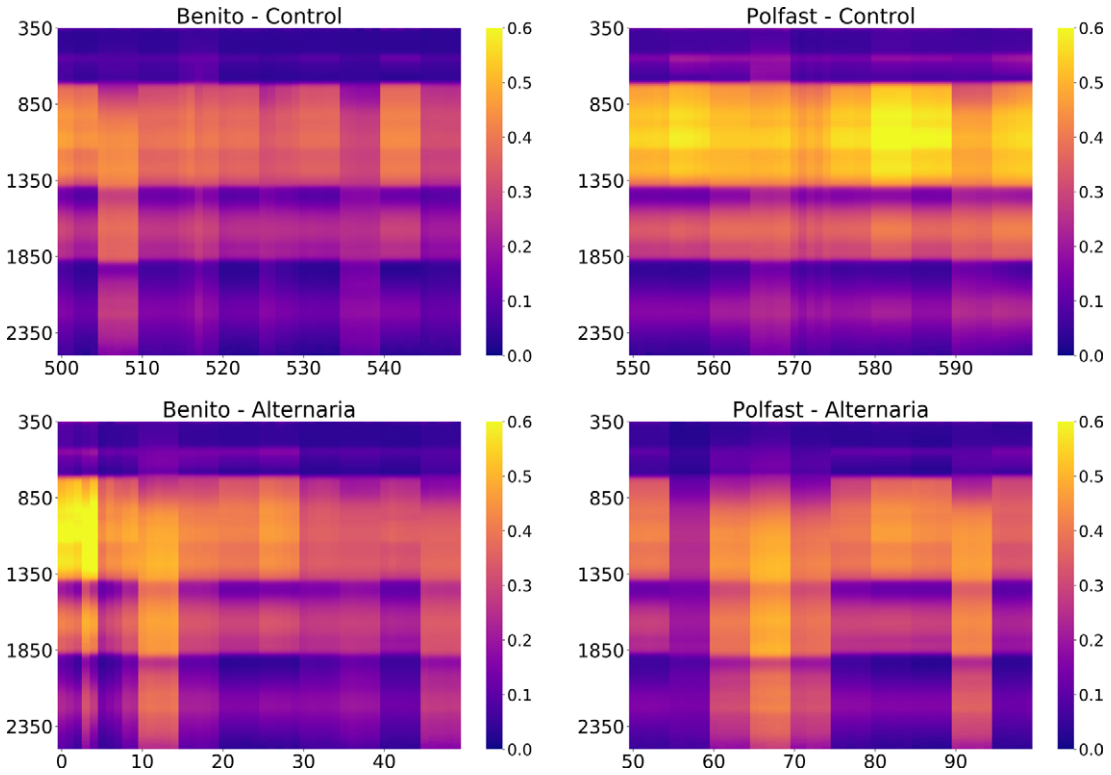


Fig. 3. The absolute reflectance diagram (with the consecutive number of the measurement set on the x-axis and the spectrum on the y-axis) for different experiment subsets – two tomato varieties: Benito and Polfast and two labeled sets: measured on infected leaves, inoculated with *Alternaria solani*, and set of the control group – measurements of leaves of not infected plants.

AUTHOR

This method is a powerful technique that usually works well without parameter tuning and does not require data scaling.

The problem presented in this study is based on learning from the measured leaf absolute reflectance values and classifying the pattern as a leaf inoculated with *Alternaria solani* or as a health leaf. The decision trees, random forest, and other ensemble learning models used for this study are prepared with sklearn framework.

3.3. Experiment configuration

The measurements acquired were converted to a data frame with a structure presented in Table 2 and visualized in Fig. 4. Each data sample consists of the reference information – attributes that describe the group, the variety, and the disease (the information if the sample indicates a tomato with *Alternaria solani* infection or the control group). Every measurement set has its number to allow its unambiguous identification, and it is presented in Table 3 in the ID column. The second group of attributes is the set of reflectance measurements, ranging from 350 to 2500 nanometers. Although every measurement provides 2151 inputs, and the dataset has 2151 columns with the results, the columns from 352.0 to 2498.0 were removed from Table 2 to improve its readability. The dataset consists of 50 samples per each group and each class. Each of those samples is a result of a set of 20 consecutive measurements taken on the field using the spectroradiometer. In the tested dataset, there are 200 samples (based on 1000 measurements in total). For the purpose of later model training the whole dataset was split into train and test sets in proportions 75:25, respectively.

First of all, the decision tree was taken into consideration. The decision hierarchy was determined. The pre-pruning method was used to optimize the model. Moreover, various pre-pruning parameters were analyzed to determine the best value of the maximum depth parameter. Feature importance analysis was carried out as well. The method was applied to the individual decision trees, to improve the overall classifier performance. Such an approach allows finding the optimal maximum depth of a tree or the number of estimators.

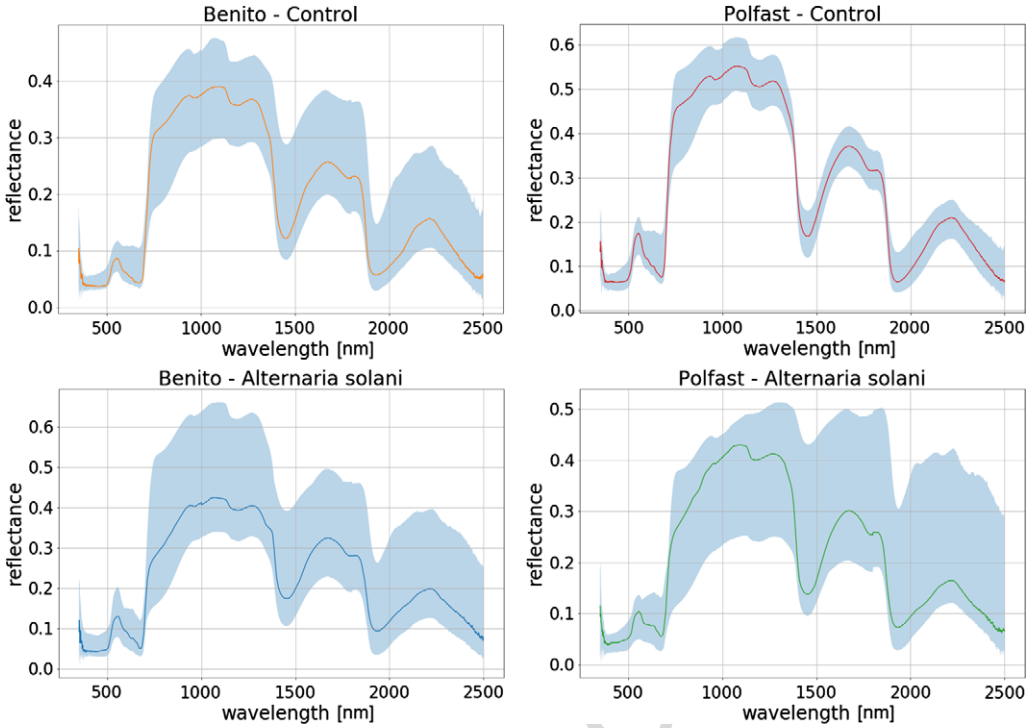


Fig. 4. Hyperspectral representation of each group (Benito–inoculated, Benito–control, Polfast–inoculated, Polfast–control). The median of spectral data is presented as a solid line, the blue background is a visualization of minimum and maximum reflectance.

Table 2

Exemplary measurements extracted from the training set (10 randomly selected out of 1000 measurements). The columns from 350.0 to 2500.0 provide the measured reflectance for the consecutive spectra

ID	Group	Variety	Disease	350.0	351.0	...	2499.0	2500.0
162	18	Polfast	Control	0.14639	0.20135	...	0.05620	0.05734
93	13	Polfast	*Alternaria*	0.13539	0.11347	...	0.23540	0.22830
103	6	Benito	Control	0.11659	0.11340	...	0.02218	0.02074
24	1	Benito	*Alternaria*	0.12174	0.08946	...	0.03038	0.03765
132	6	Benito	Control	0.04130	0.04471	...	0.01830	0.02289
581	18	Polfast	Control	0.16784	0.16610	...	0.06499	0.06480
51	13	Polfast	*Alternaria*	0.13662	0.12204	...	0.04922	0.05339
589	18	Polfast	Control	0.13475	0.15257	...	0.04688	0.04584
534	6	Benito	Control	0.09816	0.10905	...	0.03973	0.03886
71	13	Polfast	*Alternaria*	0.07317	0.09514	...	0.17331	0.18117

Table 3

Comparison of the *Alternaria solani* detection using decision trees and random forest

Method	Accuracy	Class	Precision	Re-call	F1-score
Decision trees	0.780	*Alternaria*:	0.731	0.826	0.776
		Control:	0.250	0.040	0.784
		Weighted:	0.786	0.780	0.780
Random forest	0.987	*Alternaria*:	**0.984**	**1.000**	**0.992**
		Control:	**1.000**	**0.920**	**0.958**
		Weighted:	**0.987**	**0.987**	**0.986**

AUTHOR C

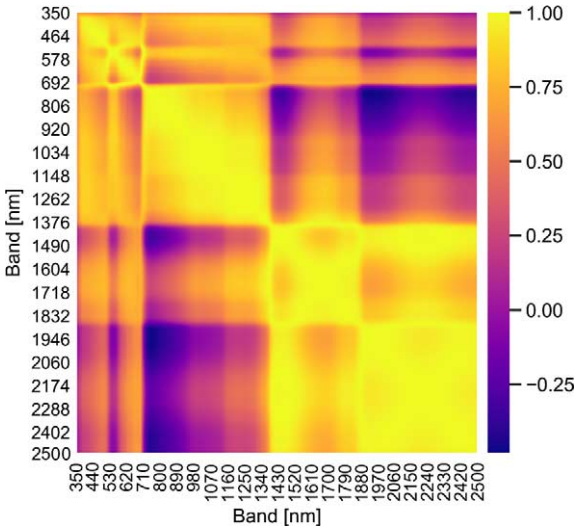


Fig. 5. Measurements correlogram for observed reflectance spectra.

The second approach was the random forest method. An analysis of various numbers of trees in the trained model was performed. As with the decision tree approach, the feature importance analysis was carried out.

4. Results and discussion

4.1. Reflectance analysis

The spectra used in this study cover the range of 350–2500 nm wavelengths with a resolution of 3 nm @ 700 nm and 8 nm @ 1400/2100 nm. Reflectance measurements were taken for four classes based on the variety.

The diagram reflects the association between all measurements split across all studied spectra. It reveals also some strong relationships between some more distant band sets, those relations are also presented in the form of a correlogram in Fig. 5.

4.2. Max-depth parameter adjustment

The first studied approach to the collected measurement classification was a single decision tree algorithm. A decision hierarchy was determined. For full-depth tree (max_depth=7) accuracy of train and test dataset was 1.00 and 0.78 respectively, what could mean the model is overfitted. As shown in Fig. 6, a depth higher than four does not increase accuracy with the test dataset in this particular case.

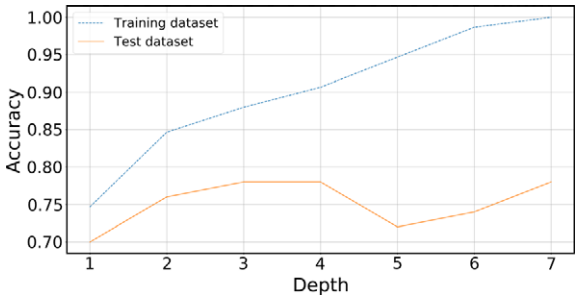


Fig. 6. Accuracy depends on decision tree max_depth.

Accuracy for train and test dataset, for parameter max_depth=4, was ∼0.91 (for the train set) and 0.78 (for the test set) respectively, so the same result on test data can be achieved with the optimized tree.

4.3. Results

The results of the classification using decision trees were insufficient, the accuracy of 0.78 and the precision related to *Alternaria solani* on the level of 0.25 were not promising and did not let to prepare the final detector.

4.3.1. Classification report comparison

The next approach to the problem was the application of several ensemble learning classifiers. The best results were obtained using the random forest algorithm.

The metrics for the classification report were calculated separately for each approach. The report consists of the most informative scores: accuracy, precision, recall, and F1-score.

The applied random forest algorithm achieved performance should be enough to correctly identify plants affected with the *Alternaria solani*. All: precision, accuracy, and re-call are high and are shown in Table 3 (all the metrics were obtained using the test set).

The metrics for the classification report were calculated separately for each approach. The report consists of most informative scores: accuracy, precision, recall, and F1-score (which is a harmonic mean of both precision and recall and could be interpreted easier though). The applied random forest algorithm achieved performance should be enough to correctly identify plants affected with the *Alternaria solani*. All: precision, accuracy, and recall come with great numbers (the greater the result is the better performance is). The difference between both approaches is clearly visible on the normalized confusion matrices that are presented in Fig. 7.

AUTHOR COPY

4.3.2. Results for the random forest algorithm

High accuracy and F1 score were obtained, for the classifier implementing Random Forrest the accuracy reached 0.987 and F1-score 0.986. The method presented herein is not scalable, due to the use of specialized measuring equipment. However, the results encourage further research on the use of ensemble learning to detect *Alternaria solani* and other infections on tomatoes cultivated under foil tunnels. As a next stage of the research new measurements are going to be incorporated into the model to extend the detection to other diseases, such as *Phytophthora infestans*, as well as the use of other machine learning methods associated with ensemble learning.

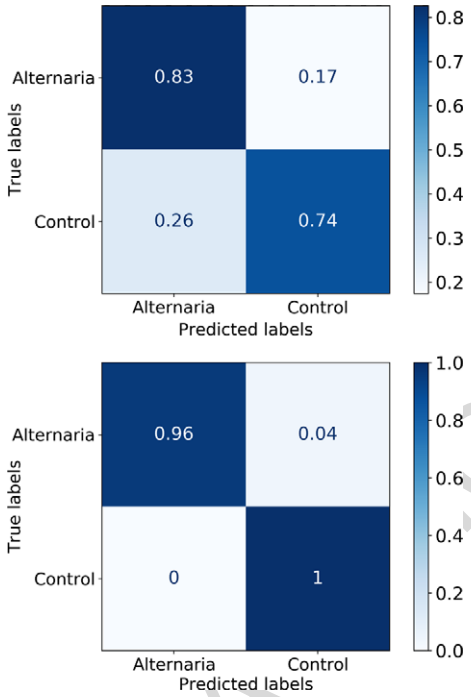


Fig. 7. Normalized confusion matrix for the fitted decision tree (top) and random forest classifier (bottom).

4.4. Feature importance analysis

Feature importance analysis was performed for both decision trees and random forest approaches.

For the decision tree the most important bands in the decision process, with feature importance parameter higher than 0.1, had wavelengths 918 and 2353 nm. Only seven bands were determined as important in the decision process. All the bands of high importance are presented in Fig. 8.

The most important bands in the decision process for random forest approach, with feature importance parameter higher than 0.1, were bands 905 and 982, but in the decision process total number of 32 bands were found useful.

4.5. Hyper-spectral ranges analysis

There are many hyperspectral measuring devices and hyperspectral cameras available. These devices have different specifications and parameters of registered reflectance. The analyzed dataset contained measurements in the spectra range 350–2500 nm, which allowed to separate it to a couple of subsets of the data to verify the accuracy metric of detecting *Alternaria solani* in a few narrower spectral ranges. Various tested channels sets correspond to the range of spectra recorded by the most popular devices (UV, VIS, VNIR, NIR, SWIR). The full list of analyzed ranges and the complete set of results is presented in Table 4.

Quite obviously, the best results were obtained for the data subset marked as UV+VIS+NIR, which refers to the whole spectral range (350–2500 nm). The achieved accuracy result was equal to 0.987. Surpris-

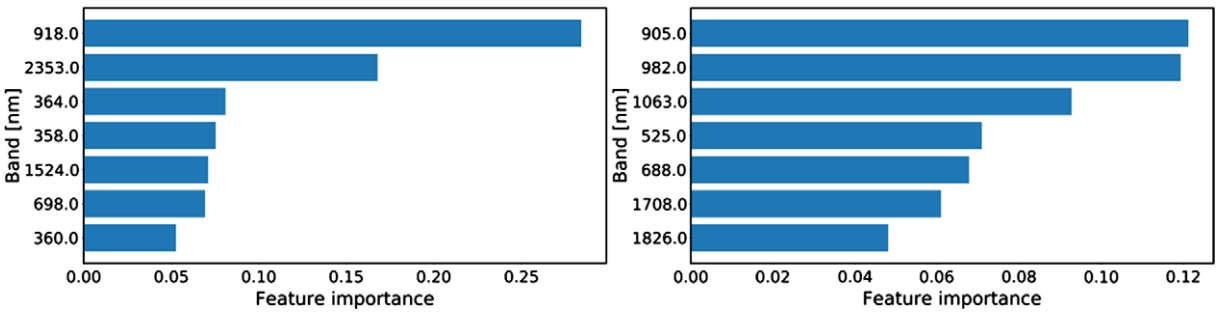


Fig. 8. Seven bands with the highest importance selected for the final decision trees model (left) and seven with the highest importance for the random forest classifier (right). The feature importance value was assigned to the x-axis.

AUTHOR COPY

Table 4
Results of classification using different spectral band ranges

Spectral range name	Bands range [nm]	Total number of bands	Accuracy	Class	Precision	Re-call	F1-score
UV	350–400	51	0.820	*Alternaria*:	0.250	0.040	0.069
				Control:	0.836	0.976	0.900
				Weighted:	0.738	0.820	0.762
VIS	400–765	366	0.967	*Alternaria*:	0.955	0.840	0.894
				Control:	0.969	0.992	0.980
				Weighted:	0.966	0.967	0.966
VNIR	400–1000	601	0.973	*Alternaria*:	1.000	0.840	0.913
				Control:	0.969	1.000	0.984
				Weighted:	0.974	0.973	0.972
NIR	900–1700	801	0.947	*Alternaria*:	0.870	0.800	0.833
				Control:	0.961	0.976	0.968
				Weighted:	0.945	0.947	0.946
SWIR	1000–2500	1501	0.953	*Alternaria*:	0.875	0.840	0.857
				Control:	0.968	0.976	0.972
				Weighted:	0.953	0.953	0.953
UV+VIS+NIR+SWIR	350–2500	2151	0.987	*Alternaria*:	**0.984**	**1.000**	**0.992**
				Control:	**1.000**	**0.920**	**0.958**
				Weighted:	**0.987**	**0.987**	**0.986**

ingly good results were achieved by models based on significantly limited data subset.

Results gained by models based on VIS and VNIR are similar to the best classificatory even though they use a limited range. Interestingly, the model using the VNIR range has achieved the best precision result of detecting *Alternaria solani*.

The only exception is the classifier using the UV range. This model was trained based on only 51 bands, which turned out to be too small a data subset for analysis and the results are unsatisfactory.

4.6. Comparison of classification using ensemble-based methods

The results of the last experiment are presented in Table 5. It is the effect of several different ensemble learning algorithms comparison. During this study, the training with the cross-validation was performed as well as the number of the estimators fit using the grid search method.

The process was done using the same dataset and again among those methods, the random forest had the highest accuracy, re-call, and precision.

Surprisingly the lowest metrics were obtained for the gradient boosting method, with the accuracy equal to only 0.893. Even after the classifier hyperparameters re-fitting the *Alternaria solani* related re-call and F1-score were found less than 0.7.

4.7. Discussion

Random forest classifiers are successfully applied in agriculture e.g. crop classification [17] and early detection of TMV (*tobacco mosaic virus*) on tobacco leaves [28].

The initial analysis of measurements' spectra was not sufficient to enable classification. To distinguish samples related to the infected plant more advanced classifier is needed. Even the selection of specific bands that could be more informative was not easy, as many of them present similar information. The smooth change of neighboring spectra could be observed on the correlogram presented in Fig. 5.

The study shows a comparison of decision tree and random forest approaches. Other ensemble learning methods, such as AdaBoost, Bagging, Extra Trees, Gradient Boosting, have been investigated to determine the most effective method for detecting *Alternaria solani*. An important stage of the study was also the review of the bands' importance level in the decision process, which helps increase the classifier efficiency. It is purposefully worth expanding this study and looking for a set of bands that will allow for the

AUTHOR COPY

Table 5
Results of classification using ensemble-based methods

Ensembling method	Fitted number of estimators	Accuracy	Class	Precision	Re-call	F1-score
AdaBoost	2000	0.947	*Alternaria*:	0.870	0.800	0.833
			Control:	0.961	0.976	0.968
			Weighted:	0.945	0.947	0.946
Bagging	50	0.933	*Alternaria*:	0.941	0.640	0.762
			Control:	0.932	0.992	0.961
			Weighted:	0.934	0.933	0.928
Extra trees	50	0.973	*Alternaria*:	0.920	0.920	0.920
			Control:	0.984	0.984	0.984
			Weighted:	0.973	0.973	0.973
Gradient boosting	1000	0.893	*Alternaria*:	0.714	0.600	0.652
			Control:	0.922	0.952	0.937
			Weighted:	0.888	0.893	0.890
Random forest	50	0.980	*Alternaria*:	**0.958**	**0.920**	**0.939**
			Control:	**0.984**	**0.992**	**0.988**
			Weighted:	**0.980**	**0.980**	**0.980**

construction of a device enabling remote detection of the pathogen.

The results of the research showed the high potential of the pathogen detection method using hyperspectral data. The use of ensemble learning methods turned out to be valuable in the in-depth study and enabled the selection of parameters of the algorithm and appropriate selection of the most significant bands in the decision process. It would be worth testing a simpler sensor, based only on selected bands, which could allow building an inexpensive device in the future work.

5. Conclusions

There is a real need to provide a non-destructive method of plant disease detection for tomato cultivators. The information about the type of infection can determine crop treatment methods and help farmers to prevent yield losses. In this work, a preliminary model for *Alternaria solani* detection was analyzed and constructed using a random forest. To check if the method is able to generalize it was tested against two different tomato varieties (Benito and Polfast) cultivated under foil tunnels. The model was trained using hyperspectral measurements from 350–2500 nm spectral range. The final model resulted in the accuracy and F1 score equal to 0.98 and was found satisfying, as it was able to correctly distinguish most presented cases (infected plants and control samples). In addition, feature importance analysis was performed and 32 bands were determined as most significant for the model.

In additional experiments, classifiers based on various measurement ranges were studied. Quite obviously the model with the full range of spectra measurements gained the highest results. However, models using absolute reflectance measured in VIS (from 400 to 765 nm) and VNIR (between 400 and 1000 nm) ranges had similar detection performance of studied pathogen. This may mean that a spectrometer working in narrower range (VIS and VNIR) can be used to observe this phenomenon, which, however, requires confirmation in subsequent experiments.

In the upcoming season, the presented study will be continued. Future works should result in new measurements, to let the incorporation of the updated model, which could detect other diseases, that will be planted there. The other possible way to extend the usability of the presented methodology is to test it with other tomato varieties and with other plants.

The described study covers the detection of *Alternaria* solani for two different tomato varieties cultivated under foil tunnels. The possible way to extend this study is to add some other pathogen detection trials. Another task that could be tackled is the detection of this pathogen but on the other species, for instance potatoes, that are also highly sensitive to *Alternaria solani*.

The application of the presented detection process to the regular fields could be impractical or even impossible. Most of the farmers do not have access to

AUTHOR COPY

a spectrophotometer, the device is pricy, the measurement process is complex and gives only single point readouts and pathogen warnings. Therefore the development of the sensor that will be able to detect such pathogens has great potential. This requires the extended study of other sensors that could provide imagery in the identified band, such as multi spectra cameras or hyperspectral cameras, and some trails on an unmanned aerial vehicle supported data acquisition. However, that kind of remote measurement, from higher altitudes would result in a signal that could demand the band study reconsideration.

Acknowledgements

This work was partially supported by The National Centre of Research and Development of Poland under project POIR.01.01.01-00-1317/17.

References

[1] J. Abdulridha, Y. Ampatzidis, S.C. Kakarla and P. Roberts, Detection of target spot and bacterial spot diseases in tomato using UAV-based and benchtop-based hyperspectral imaging techniques, *Precision Agriculture* (2019), 1–24.

[2] S. Adhikari, E. Kc, L. Balkumari, B. Shrestha and B. Baiju, Tomato plant diseases detection system using image processing, in: *Kantipur Engineering College Conference*, 2018.

[3] C.H. Bock, G.H. Poole, P.E. Parker and T.R. Gottwald, Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging, *CRC Critical Reviews in Plant Sciences* **29**(2) (2010), 59–107. doi:10.1080/07352681003617285.

[4] F. Bohm, R. Edge and G. Truscott, Interactions of dietary carotenoids with activated (singlet) oxygen and free radicals: Potential effects for human health, *Molecular Nutrition & Food Research* **56**(2) (2012), 205–216. doi:10.1002/mnfr.201100222.

[5] S. Candiago, F. Remondino, M. De Giglio, M. Dubbini and M. Gattelli, Evaluating multispectral images and vegetation indices for precision farming applications from UAV images, *Remote Sensing* **7** (2015), 4026–4047. doi:10.3390/rs70404026.

[6] V.V. Datar and C.D. Mayee, Conidial dispersal of *Alternaria solani* in tomato, *Indian Phytopathology* **35** (1982), 68–70.

[7] J. Fahrentrapp, F. Ria, M. Geilhausen and B. Panassiti, Detection of gray mold leaf infections prior to visual symptom appearance using a five-band multispectral sensor, *Frontiers in Plant Science* **10** (2019), 1–14.

[8] FAOSTAT: Food and Agriculture Organization of the United Nations, Tomato globa production, http://www.fao.org/faostat/en/#data, access: 24.03.2020.

[9] M.R. Foolad, Genome mapping and molecular breeding of tomato, *International Journal of Plant Genomics* **2007** (2007).

[10] K. Golhani, S.K. Balasundram, G. Vadamalai and B. Pradhan, A review of neural networks in plant disease detection using hyperspectral data, *Inf. Process. Agric.* (2018).

[11] J.F.J. Grigolli, M.M. Kubota, D.P. Alves, G.B. Rodrigues, C.R. Cardoso, D.J.H. da Silva and E.S.G. Mizubuti, Characterization of tomato accessions for resistance to early blight, *Crop Breeding and Applied Biotechnology* **11**(2) (2011), 174–180. doi:10.1590/S1984-70332011000200010.

[12] S. Guido and A. Müller, Introduction to Machine Learning with Python: A Guide for Data Scientists, 2017, O'Reilly.

[13] M.A. Guillén-Navarro, R. Martínez-España, A. Llanes, A. Bueno-Crespo and J.M. Cecilia, A deep learning model to predict lower temperatures in agriculture, *Journal of Ambient Intelligence and Smart Environments* **12**(1) (2020), 21–34. doi:10.3233/AIS-200546.

[14] T.K. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8) (1998), 832–844. doi:10.1109/34.709601.

[15] J. Lu, R. Ehsani, Y. Shi, A.I. de Castro and S. Wang, Detection of multi-tomato leaf diseases (late blight, target and bacterial spots) in different stages by using a spectral-based sensor, *Scientific Reports* **8**(1) (2018), 2793. doi:10.1038/s41598-018-21191-6.

[16] P. Moghadam, D. Ward, E. Goan, S. Jayawardena, P. Sikka and E. Hernandez, Plant disease detection using hyperspectral imaging, in: *2017 International Conference on Digital Image Computing: Techniques and Applications DICTA 2017*, Vol. 2017, 2017, pp. 1–8.

[17] A.O. Ok, O. Akar and O. Gungor, Evaluation of random forest method for agricultural crop classification, *European Journal of Remote Sensing* **45**(1) (2012), 421–432. doi:10.5721/EuJRS20124535.

[18] E.S. Rao, A.D. Munshi, P. Sinha and Rajkumar, Genetics of rate limiting disease reaction to *Alternaria solani* in tomato, *Euphytica* **159**(1–2) (2008), 123–134.

[19] P.P. Ray, Internet of things for smart agriculture: Technologies, practices and future direction, *Journal of Ambient Intelligence and Smart Environments* **9**(4) (2017), 395–420. doi:10.3233/AIS-170440.

[20] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk and D. Stefanovic, Deep neural networks based recognition of plant diseases by leaf image classification, *Computational Intelligence and Neuroscience* **2016** (2016).

[21] W. Song, X. Ma, H. Tan and J. Zhou, Abscisic acid enhances resistance to *Alternaria solani* in tomato seedlings, *Plant Physiology and Biochemistry* **49**(7) (2011), 693–700. doi:10.1016/j.plaphy.2011.03.018.

[22] R. Van De Vijver, K. Mertens, K. Heungens, B. Somers, D. Nuyttens, I. Borra-Serrano, P. Lootens, I. Roldán-Ruiz, J. Vangeyte and W. Saeys, In-field detection of *Alternaria solani* in potato crops using hyperspectral imaging, *Computers and Electronics in Agriculture* **168** (2020).

[23] R. Van De Vijver, K. Mertens, K. Heungens, B. Somers, D. Nuyttens, T. De Swaef, I. Roldán-Ruiz, I. Borra-Serrano, P. Lootens, J. Vangeyte and W. Saeys, Towards automated field detection of *Alternaria solani* using hyperspectral imaging, no. July, 2018.

[24] X. Wang, M. Zhang, J. Zhu and S. Geng, Spectral prediction of Phytophthora infestans infection on tomatoes using artificial neural network (ANN), *International Journal*

AUTHOR COPY

of Remote Sensing **29**(6) (2008), 1693–1706. doi:10.1080/01431160701281007.

[25] C. Xie, Y. Shao, X. Li and Y. He, Detection of early blight and late blight diseases on tomato leaves using hyperspectral imaging, *Scientific Reports, Nat. Publ. Gr.* **5**(16564) (2015), 1–11.

[26] C. Xie, C. Yang and Y. He, Hyperspectral imaging for classification of healthy and gray mold diseased tomato leaves with different infection severities, *Computers and Electronics in Agriculture* (2017).

[27] M. Zhang and Z. Qin, Spectral analysis of tomato late blight infections for remote sensing of tomato disease stress in California IGARSS 2004, *2004 IEEE International Geoscience and Remote Sensing Symposium* **6** (2004), 4091–4094.

[28] H. Zhu, H. Cen, C. Zhang and Y. He, Early detection and classification of tobacco leaves inoculated with tobacco mosaic virus based on hyperspectral imaging technique, in: *2016 American Society of Agricultural and Biological Engineers*, ASABE, 2016, pp. 1–7.

AUTHOR COPY

Załącznik 2 – [A2]

Signal, Image and Video Processing (2024) 18:5461–5476
https://doi.org/10.1007/s11760-024-03247-5

ORIGINAL PAPER

Various tomato infection discrimination using spectroscopy

Bogdan Ruszczak[1] · Krzysztof Smykała[1,2] · Michał Tomaszewski[1] · Pedro Javier Navarro Lorente[3]

Received: 6 December 2023 / Revised: 8 April 2024 / Accepted: 26 April 2024 / Published online: 17 May 2024
© The Author(s) 2024

Abstract
Diagnosing plant diseases is a difficult task, but it could be made easier with the use of advanced instrumentation and the latest machine learning techniques. This paper is a further development of a previous authors study by the authors, which has been extended to provide the classification method for tomato diseases and to indicate the spectral ranges of greatest importance for this process. As tomatoes are one of the most popular and consumed vegetables, and diseases of this crop even reduce yields by up to 80% every year, their detection is a vague topic. This manuscript describes research in which spectroscopy was used to develop methods for discriminating between selected tomato diseases. The following, frequently occurring diseases were investigated for this research: *anthracnose, bacterial speck, early blight, late blight*, and *Septoria Leaf Spot*. The study used a dataset consisting of 3877 measurements taken with the ASD FieldSpec 4 Hi-Res spectroradiometer in the 350–2500 nm range from 2019/09/10 to 2019/12/20. The highest classification efficiency (F_1 score) of 0.896 was obtained for the logistic regression based model which was evaluated on *Septoria Leaf Spot* disease records.

Keywords Plant diseases classification · Machine learning · Spectroscopy · Hyperspectral measurements

1 Introduction

Tomato (*Solanum lycopersicum*) is one of the most grown vegetables in the world. The yearly global yield of this plant reaches 160 million tons, and it is the second most-consumed vegetable in the world [1]. However, tomatoes are very vulnerable to fungal, bacterial, and viral diseases such as late blight, powdery mildew, or early blight [2]. Some of these diseases, like *Alternaria solani*, could reduce yield by 80% under favorable conditions for the pathogen [3, 4]. Farmers counteract the effects of the diseases in various ways e.g., by fungicide control or planting plants partially resistant to diseases. However, a key element of the disease counteract process is a correct diagnosis. Currently, the standard diagnosis of these diseases is based on a microscopic assessment of the plant tissue damage and assessing the number of pathogen spores. Therefore, the non-invasive, reliable, and fast method of disease recognition is indispensable in modern agriculture [5].

The use of spectroscopy, along with digital signal processing algorithms and artificial intelligence, seems to be increasingly popular among researchers [5–7]. Spectral data are used on a large scale in remote sensing thanks to space agencies, such as NASA or ESA. The data are useful in, among others, crisis management, climate change monitoring, and agriculture. In the last one, the useful applications are indicators [8] that enable the determination of plant vegetation [8, 9], water stress [10, 75], and soil moisture [11, 12] or another soil properties [13].

The present study is a continuation of the authors' previous work [14] in which they discussed the possibility of early detection of several tomato pathogens using spectroscopy. As an extension, we further developed the machine learning

✉ Bogdan Ruszczak
b.ruszczak@po.edu.pl

Krzysztof Smykała
k.smykala@gmail.com

Michał Tomaszewski
m.tomaszewski@po.edu.pl

Pedro Javier Navarro Lorente
pedroj.navarro@upct.es

1 Department of Computer Science, Faculty of Electrical Engineering, Automatic Control and Informatics, Opole University of Technology, Prószkowska Street 76, 45-758 Opole, Opolskie, Poland

2 Research and Development Department, QZ Solutions Sp. z o.o., Technologiczna 2 Street, 45-839 Opole, Opolskie, Poland

3 Escuela Técnica Superior de Ingeniería de Telecomunicación (DSIE), Universidad Politécnica de Cartagena, Campus Muralla del Mar, 30202 Cartagena, Region of Murcia, Spain

models aiming to increase method precision. To be specific, in contrast to the previous work, three main goals were set for the presented work, which should be considered:

- *Goal 1*: finding the best classifier allowing one to distinguish diseased plants from healthy ones based on spectroscopy in 6 experiments variants:
  - CS (control samples) versus objects infected by AN—*Anthracnose* (*Colletotrichum coccodes*),
  - CS versus objects infected by BS - *Bacterial speck* (*Pseudomonas syringae*),
  - CS versus objects infected by EB - *early blight* (*Alternaria solani*),
  - CS versus objects infected by LB - *late blight* (*Phytophthora infestans*),
  - CS versus objects infected by SL - *Septoria Leaf Spot* (*Septoria lycopersici*),
  - CS versus randomly selected measurements of infected objects (AN, BS, EB, LB, SL).
- *Goal 2*: data processing method development, i.e., creating a data processing procedure that achieves the highest improvement in binary classification results,
- *Goal 3*: hypothesis verification that selected spectral bands are especially important in the disease analysis process.

Data used for this study consists of 3877 spectra of leaf reflectance taken by *ASD FieldSpec 4 Hi-Res* spectroradiometer in the spectral range 350–2500 nm and metadata for each measurement such as disease, date of measurement, date of infection, etc. The dataset contained control samples (CS) and infected samples (AN, BS, EB, LB, SL) at various stages of infection, from the early stage where symptoms are not visible by the naked eye to the late stage, which is a novelty compared to the researches conducted so far.

There are a number of mature machine learning algorithms developed that could be employed for investigating the spectral information for the purpose of classification. In Sect. 4 we enumerate the methods that are currently used for a number of pathogen detection applications, including linear models, tree-based methods, ensemble machine learning, and neural networks.

2 Diagnostic methods for solanum lycopersicum diseases

The issue of detecting plant diseases attracts the interest of many researchers, which is reflected in numerous publications [5, 9, 15–18]. Among other studies, one could find a research targeted to pathogens on *Asian soybean*[19], *wheat* infected with *Fusarium head blight* [20], *sugar beet*[21] or even for phenotyping disease resistance of crops[22]. A significant part of the researches concerns tomatoes crops [14, 23–40].

Conventional disease diagnosis methods involve the observation of disease symptoms. However, the symptoms of the various diseases can be markedly similar to each other, which makes it difficult even for experienced farmers to correctly diagnose the disease. The detailed examination requires, therefore, the collection of plant material and laboratory tests [41]. In the case of fungal diseases, the pathogen has to be isolated and cultured in vitro. The identification is confirmed using microscopic techniques by a taxonomist based on spore morphology and conidiogenesis [41]. Biochemical tests can determine bacterial diseases, while viruses are identified based on genetic material, transmission assays, and their host range [42].

Invasive methods, including enzyme-linked immunosorbent assay, polymerase chain reaction, and immunofluorescent assay, are imperfect in terms of cost, speed, and accuracy matter, while near-infrared spectroscopy and hyperspectral imaging allow for an equally good, cost-effective, and non-invasive plant diagnostics [43], not limited to tomato cultivation, with reported applications to, i.e. potato [44] or rice [45, 46], as well as other plant related phenomena, such as chlorophyll modeling [47, 48] or plant phenotyping [49, 50].

Some authors note that the strongly developing agrotechnical and food processing industries require the development of non-invasive diagnostic methods [43, 51]. Digital cameras [31–34, 52–54], multispectral imaging [30, 55, 56] and spectroscopy [26–28, 35, 37, 57] were used for the researches on that issue.

Attempts to diagnose crops and identify *Solanum lycopersicum* diseases on the basis of spectral measurements were made by, among others:

- P1 - Zhang et al. [24];
- P2 - Wang et al. [36];
- P3 - Jones et al. [38];
- P4 - Moghadam P. et al. [26];
- P5 - Xie C. et al. [35];
- P6 - Lu J. et al. [37].
- P7 - Pereira J. et al. [40]

In the mentioned publications, various spectral measurements were used, from ultraviolet (UV) to visual light range (VIS) and near-infrared (NIR) to short-wave-infrared (SWIR). In publications P1 and P2 a spectroradiometer *GER* that works in the 400–2500 nm spectral range was used. In P3 a spectroradiometer *Cary 500* measuring within 200–2500 nm spectral range was employed. A set consisting of 2

hyperspectral *Headwall* cameras, imaging in 400–1000 nm and 900–2500 nm ranges, was used in P4. In P5 a hyperspectral camera, imaging within 380–1023 nm spectral range, was used. Results presented in P6 were based on spectroradiometer measurements in the 400–1050 nm spectral range.

A synthetic summary of the characteristics of the conducted measurements, research, and results is presented in Table 1.

Dataset used in publications P1 and P2 consisted of spectral data of 60 plants and 1074 measurement points. Additionally, 195 spectroscopic images from *the AVIRIS programme*, covering 700–1130nm wavelengths, were used in P2. For P1, the average values for six spectral ranges were calculated (600–690 nm, 750–930 nm, 950–1030 nm, 1040–1130 nm, 1450–1850 nm, 2000–2400 nm), and novel spectral disease indices were proposed. The indices were quotients of two selected ranges. The result of the P1 publication was the discovery of the most relevant wavelengths in a matter of *late blight* detection on tomatoes, while P2 focused on the creation of a backpropagation artificial neural network for resolving the same issue. Authors of the P1 indicate the most important spectral ranges (750–930 nm in the first place, then 950–1030 nm and 1040–1130 nm) for the process of *Late Blight* detection in tomato crops. The researchers noted five specified, narrow wavelengths characterized by a high reflectance difference for diseased and healthy plants (625 nm, 850 nm, 1050 nm, 1500 nm and 2100 nm). Based on that observation, they developed novel vegetation indices (VI). Authors of the P2 indicate spectral ranges to distinguish healthy and infected objects—750–1350 nm for field measurements and 700–1105nm for *AVIRIS* imaging spectroscopy. The study was conducted based on the plants with symptoms visible by the bare eye.

In the P3 study analysis of a correlation coefficient spectrum, analysis of a B-matrix from partial least squares regression (PLS) and stepwise multiple linear regression procedure (SMLR) was used to diagnose bacterial leaf spot of tomato. The studied dataset consisted of 156 spectra. Authors of the P3 publication noted a significant correlation between disease severity and spectral absorbance in several wavelengths: 384nm, 626nm, 691nm, and 761nm. A spectral area of 750–760nm was identified as meaningful in all used approaches mentioned above. Additionally, researchers indicated wavelengths 395nm, 400nm, 630nm, and 633–635nm as significant.

The authors of the P4 publication implemented three various approaches to the issue of detecting diseased and healthy plants. A goal set by researchers in this particular study was to distinguish control plants from those infected by *tomato spotted wilt virus* (TSWV) by using hyperspectral imaging of 400–2500nm spectral range. The very first approach applied in the study was a feature extraction method based on existing vegetation indices (12 various indices). In the second approach, the whole available spectral range was used. In the third approach, probabilistic topic modeling (PTM) was used. PTM is a method usually used in natural language processing (NLP) for determining the topic of the document. The researchers treated the hyperspectral image as a document, and measurement values for each wavelength were treated as a word in the document.

The dataset used in P4 consisted of imaging of 30 diseased and 30 control plants of 1 variety of tomato. Plants were observed for 21 days. During this time, six measurement sessions were taken from the 3rd to the 21st day after inoculation. During the sessions, 133 diseased leaves and 103 control leaves were measured. The Kullback-Leibler divergence, also known as relative entropy, was used in the study to estimate the distance between the measurement distribution of diseased and control objects.

The best results were achieved using the full wavelength range. The method based on full SWIR wavelength attained an F_1 score equal to 0.92. For VNIR, the F_1 score reached 0.94. However, all analyzed approaches obtained an F_1 score equal to or higher than 0.8. The authors pointed out that no specified wavelength is the most significant for disease classification issues.

The study described in P5 was focused on using spectroradiometer measurements in the 380–1023 nm spectral range to train a multiclassification model for *Early Blight, Late Blight* and control plants. The study's main concentration was on wavelengths 442, 508, 573, 696, and 715 nm. To achieve this, 310 measurements were used—120 measurements of healthy plants, 120–*early blight*, and 70 infected by *late blight*. The obtained accuracy of the classification ranged from 97.1 to 100%.

The P6 publication covers the study of 57 spectral vegetation indices (SVIs). Based on the conducted studies, the researchers indicated the four most significant SVIs, which refer to 6 considerable wavelengths, i.e., 445, 450, 690, 707, 1070, and 1200nm. The 445 measurements were used in the study, and the aim of the work was the classification model to classify four various classes—control (74 measurements), infected but asymptotic plants (77 measurements), diseased plants with visible symptoms in an early stage of disease (148 measurements) and diseased plants with visible symptoms in a late stage of disease (146 measurements). The obtained accuracy reached 100% for control, asymptotic and late-stage classes.

The authors of the P7 report 15 additional bands between 455 and 666 nm as the most relevant compared to the presented results. However, all of the wavelengths mentioned in the P7 are in the visible spectrum 434.9−680.02 nm, as the authors focused on this specific spectral range and attribute it to the absorption of chlorophylls (430 to 480 nm and 640 to 700 nm), carotenoid pigments (450 to 480 nm and 600 to 650 nm), and xanthophylls (520 to 580 nm). For the diseases

Table 1 A synthetic summary of the characteristics of the conducted measurements, research and results

	Publication	Data source	Disease	Significant wavelengths	Number of samples	Number of plants	Results
P1	Spectral analysis of tomato late blight infections for remote sensing of tomato disease stress in California	SP (400–2500 nm)	Late blight	5 bands: 625 nm, 850 nm, 1050 nm, 1500nm and 2100 nm	1074	60	N/A
P2	Spectral prediction of phytophthora infestans infection on tomatoes using artificial neural network (ANN)	SP (400–2500 nm)	Late Blight	Field measurement: range 750–1350 nm; AVIRIS: range 700–1105 nm	1074	60	R=0.99
P3	Diagnosis of bacterial spot of tomato using spectral signatures	SP (200–2500 nm)	Bacterial speck	5 bands/ranges - 395 nm, 400nm, 630nm, range 633–635nm and 750–760nm	156 (104 train, 52 test)	N/A	R^2=0.82
P4	Plant disease detection using hyperspectral imaging	HSI (400–1000 nm and 900–2500 nm)	Tomato spotted wilt virus	N/A	236 (133 infected + 103 control)	60 (30 infected + 30 control)	F_1=0.800−0.939
P5	Detection of early blight and late blight diseases on tomato leaves using hyperspectral imaging	–	Early blight, late blight	5 bands - 442, 508, 573, 696 and 715 nm	310 (120 control, 120 early blight, 70 late blight	N/A	Acc=85.71–100[%]
P6	Detection of multi-tomato leaf diseases (late blight, target and bacterial spots) in different stages by using a spectral-based sensor	SP (400–1050 nm)	Late blight, target spot, bacterial spots	4 most significant spectral vegetation indices (SVI) based on bands - 445, 450, 690, 707, 750, 800, 1070, 1200 nm, but 750 and 800 was not selected by principal component analysis	445 (74 control, 77 asymptotic late blight, 148 symptomatic early stage, 146 symptomatic late-stage)	N/A (collected under real-life conditions)	Acc=100[%]
ER	Presented work	SP (350–2500 nm)	Control +5 diseases: AN, BS, EB, LB, SL	Described in the Goal 1—on page 2	11,634 (for 3 different measurement heights) 3877 (for height 5 cm)	Laboratory conditions (3 separated phytotrons)	F_1, described in the results section on the page 16

Table 1 continued

	Publication	Data source	Disease	Significant wavelengths	Number of samples	Number of plants	Results
P7	Enhancing host-pathogen phenotyping dynamics: early detection of tomato bacterial diseases using hyperspectral point measurement and predictive modeling	SP (200–1100 nm)	Bacterial speck, bacterial spot	Bands mostly in the blue-green and red VIS spectra regions, B: 434.9, 435.7, 438.2, 438.6, 440.2, 441.4, 442.7, 443.1, 445.5, 445.9, 448.4, 448.8, 494.6; G: 503.7, 508.7, 527.5; R: 556.1, 562.0, 562.8, 590.4, 607.8, 609.1, 611.2, 618.5, 643.4, 650.2, 673.9, 680 nm	3478 measuring points	9 plants (3 for control, 3 for bacterial spec, 3 for bacterial spots)	ACC=100% (bacterial speck) ACC=74% (bacterial spot)

where: *SP* spectroscopy, *HSI* hyperspectral imaging, *Acc* accuracy

studied (bacterial speck, bacterial spot) the most important wavelengths were found to be in the blue-green and red VIS regions of the electromagnetic spectrum.

3 Materials and methods

The following part of the manuscript aims to detail the investigation procedure. The whole research involved four main components: the plant experiments, the spectroscopy acquisitions, the development of the machine learning models and the drawing of the necessary conclusions. This chapter provides the required information on how this procedure was carried out and how it could be reproduced.

3.1 The experiment description

The *ASD FieldSpec 4 Hi-Res* spectroradiometer was used to perform the acquisitions. The device measures reflectance, transmission, radiance, or irradiance of the tested sample in the spectral range of 350–2500 nm with a spectral resolution of 3 nm and 8 nm for 700 nm, 1400 nm, and 2100 nm, respectively. The spectral sampling width was 1.4 nm for the UV-VNIR range (350–1000 nm) and 1.1 nm for the VNIR-SWIR range (1001–2500 nm). This device is built of 3 detectors: one 512-element silicon near infrared sensor (VNIR: 350–1000 [nm]) and two InGaAs photodiode-based, 2 Stage TE Cooled Graded detectors ($SWIR_1$: 1000–1800 [nm] and $SWIR_2$: 1800–2500 [nm]). This device is characterized by its decent performance, including the wavelength reproducibility of 0.1 [nm], the wavelength accuracy of 0.5 [nm] for the average error of wavelength calibration fit, the wavelength accuracy of ± 1 [nm] for any one line, and the Noise Equivalent Radiance of: 1.0×10^{-9}[W/cm^2/nm/sr] 1.4×10^{-9}[W/cm^2/nm/sr] 2.2×10^{-9}[W/cm^2/nm/sr], for detectors: VNIR at 700[nm], $SWIR_1$ at 1400[nm] and $SWIR_2$ at 2100[nm] respectively. In this study, reflectance measurements were used.

The measurements, the basis for the analyzed data set, were made from September 10, 2019 until December 20, 2019. Leaves were first removed from the investigated plants, which at that time had between 30 and 40 leaves, and immediately taken to the measuring station to avoid deterioration. This also means that the following measurements on different days were taken from different leaves, but were taken from the same plants, previously inoculated, and treated in the same way in separate vegetative chambers for each pathogen. Measurements were performed in laboratory conditions in vegetation chambers (phytotrons). Figure 1 shows the time of sowing (light green), planting (green), and infecting plants (purple), as well as the period of taking measurements on individual phytotrons (blue).

Two varieties of tomatoes were used in the experiment: *Benito* and *Polfast*. Six test plants were prepared for each of the varieties. Plants were infected with five different pathogens identified in the introduction to this article. The additional surplus was a control plant that was not treated with any pathogen. The test was carried out in 3 measurement cycles. The inoculation (infection) in the first cycle took place on September 10, 2019, in the second cycle—on November 12, 2019, and in the third cycle—on December

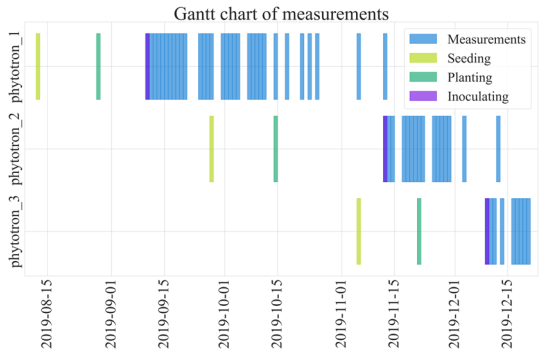


Fig. 1 Experiment Gantt chart presenting: sowing, planting and infecting days on a timeline and taking samples; every strip represents one measurement day

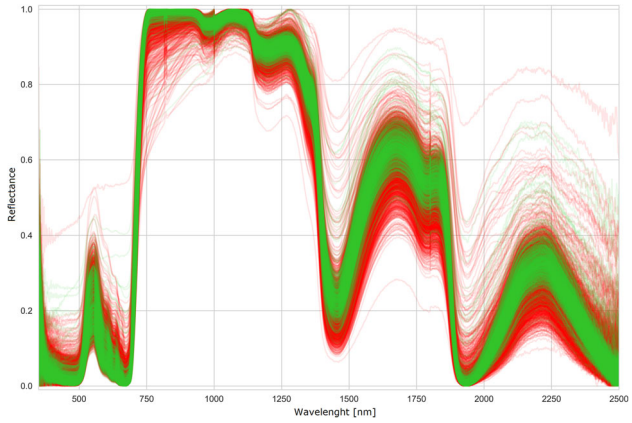


Fig. 2 Spectrum of all measurements normalized for a height of 5 cm (control measurements—green, infected objects—red)

9, 2019. The first symptoms of infection, visible to humans, were visible 3–5 days after the infection.

An artificial light source in the form of two identical halogen lamps was used to illuminate the tested objects properly. On two sides we set up two *Ushio Eurostar Reflekto MR16* bulbs of the color temperature of 3000 [K], with narrow 12°, spot beam spread, and 11,000 [cd] luminous intensity.

A spectroradiometer calibration is required before every measurement session. The calibration consists of measuring white reference and dark current, where white reference refers to an object with nearly 100% reflectance and dark current refers to the current generated within a detector in the absence of any external photons. Spectralon reference panel made of polytetrafluoroethylene (PTFE) and sintered halon was used as the reference white. Its reflectance in the 400–1500 nm range is over 99% and over 95% in the range of 250–2500 nm [58].

A single reflectance measurement is a result of averaging five successive spectral curves. The measurements were made from three distances of the measuring instrument from the infected leaves: 5, 30, and 60 cm. The analysis of the obtained measurements made at different heights showed that the reflectance differs depending on the height of the measurement. The light scattering and the spectrometer optical fiber parameters cause the reflectance measurements of the same *Solanum lycopersicum* leaves taken from greater distances (30 and 60 cm) to be subject to significant external disturbances. We found at the later stage of the project, that the accuracy of leaf targeting and the repeatability of scans were higher when the instrument was placed 5 cm above the sample. The resulting observed area diameter is 2.2 cm and that aids precise leaf targeting. The characteristics of various scan heights were also highly distinct and could hinder machine learning modeling. Therefore measurements taken at the height of 5 cm from the tested object were used (Fig. 2).

3.2 Structure of the dataset

The initial data processing consisted of isolating incorrect data, such as incorrect calibration measurements and objects that were not the subject of the project (e.g., background of the tested object) from the correctly performed measurements. Then the metadata encoded in the measurement reference numbers was extracted and added to the data set, creating a model matrix (tidy dataset) [59].

As a result, the number of primary measurements was reduced from 72,156 to 58,186 by screening out the measurement errors, then to 11,634 by determining the medians of the measurement series. Error detection was based on a straightforward visual inspection of the spectroscopy scans performed, filtering out the obvious outliers and out-of-range spectral curves from the dataset. Finally, the measurements made from a height of 5 cm were selected as the least noisy with the background image and at the same time, the most reliable in the research process, which resulted in obtaining 3877 reliable, fully described with metadata, unique measurements. The whole process of the experiment and data collection has been depicted in Fig. 3.

This study's calculations and data processing have been done using Python-based open-source software (licensed under GPLv3), applying Sci-Kit Learn, NumPy, SciPy, Pandas, Seaborn, and Matplotlib scientific libraries.

4 Classification using machine learning algorithms and their evaluation

Several studies reported promising results for plant disease symptoms identification, mainly for images, and the situation when the symptoms are rather visible, working on images using convolutional neural networks for tomato [60–64] as well as for other plants such as potato [65] cucumber [66]

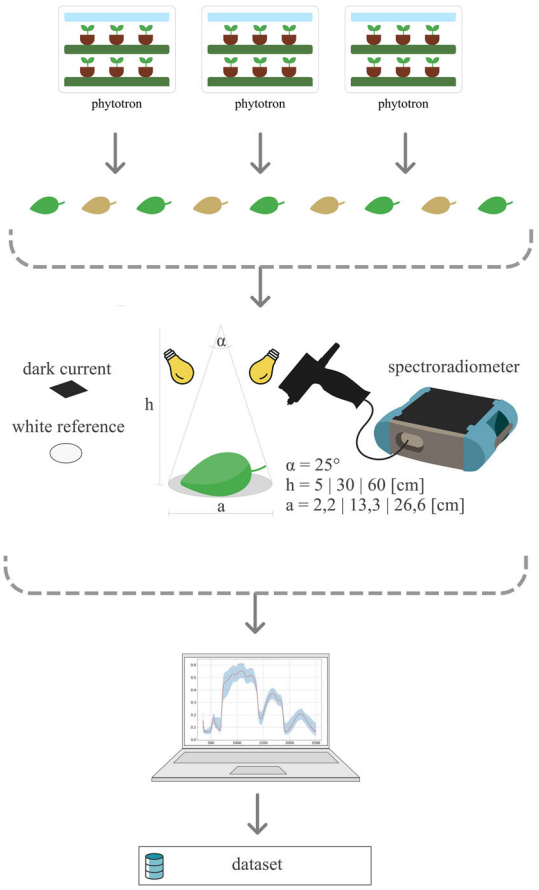


Fig. 3 Diagram of the measurement data recording process (from measurements to a tidy dataset)

or rice [67] (this study exploited also thermal imagery to improve modeling performance, for detecting any visual disease symptoms that cannot be detected externally). Some researchers, similarly to the presented study, explored the potential of hyperspectral information towards the detection of both symptomatic and non-symptomatic cases, of single pathogen detection in this case [68]. Some extended the investigation testing also the possibility of delivering the segmentation algorithms for pathogens on the visually monitored tomato fields [69, 70].

There are numerous different machine learning techniques that have been tested useful for plant disease classification. Even linear methods, like ridge classifier (RC) [71], logistic regression (LR) and linear support vector classification (linear SVC) could provide a sufficient model for some reported cases [28]. The more complex ensemble machine learning techniques, with the ability to combine several weaker submodels that are advised in the number of papers [72–74]. Those are, among others random forest (RF) [28, 73, 74], modified RF [75] light gradient boosting machine (LGBM) [72, 76] and extreme gradient boosting classifier (XGB) [74]. Some of the classifiers were tested on data in a similar fashion, collected from tomato samples, and after applying the ensemble machine learning models, concluded with auspicious results [28, 73]. Therefore, the set of the recalled models has been selected for further investigation in this study.

The range of machine learning methods for detecting plant diseases is quite extensive. Researchers have employed a variety of methods, including disease indices (P1[24]), linear modeling (P3[38]), probabilistic approaches (P4[26]), and nearest neighbor classifiers (P5[35]). These methods provide a useful set of tools for interpreting results later on. For certain investigative purposes, it may be necessary to employ more complex architectures when the searched patterns are not easily identifiable. These architectures may include neural networks (P2[36]), support vector machines[77] and convolutional neural networks[78] that support the discovery of more abstract pattern representations.

To efficiently conduct the experiment and evaluate the resulting models' performance the cross-validation procedure should be applied [79, 80]. In order to compare the performance of the prepared models, we decided to investigate the F_1 metric, which allows us to indirectly check the precision and recall results, and is denoted in the Eq. 3. For results calculated using the validation set, let TP be a number of true positives, or correctly indicated infected samples, TN be a number of true negatives, or correctly indicated not infected samples, FP be a number of false positives, or incorrectly classified not infected samples, and FN be a number of false negatives, that is infected samples missed the examined classifier. And let:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

and let:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

than the F_1 metric that gathers all the necessary information of both those metrics could be denoted as:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2 \cdot TP + FP + FN} \tag{3}$$

Such an approach provides more detailed information about model performance, than simple accuracy that takes into account only the number of correct predictions divided by the total number of predictions and does not reveal the whole case complexity. The usage of such a metric for pathogen classification evaluation was proposed also in [81].

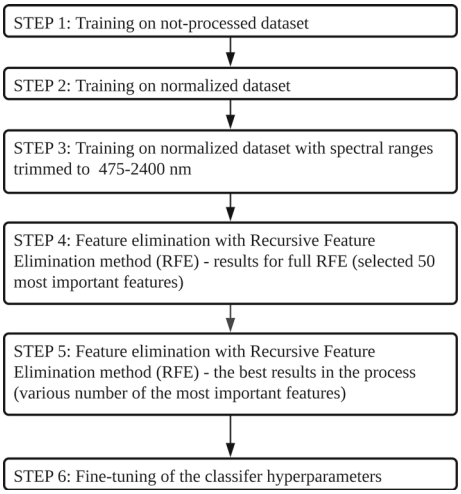


Fig. 4 Diagram of the subsequent stages of the research experiment

5 Results

The investigation described had two main objectives, therefore the results section provides two separate parts to support both. The first is concerned with disease classification and the second focuses on the search for spectral bands that are relevant to specific diseases.

Initial classification trials were performed in order to select machine learning algorithms for further research. The best results were obtained by the following classifiers: LGBM—light gradient boosting machine, linear SVC—linear support vector classification, LR—logistic regression, RF—random forest, RC—ridge classifier, XGB—extreme gradient boosting. These algorithms were used in the actual in-depth study.

In order to increase the effectiveness of the classification of Solanum lycopersicum cultivation diseases, a step-by-step approach to the problem under consideration was chosen. The first classification [STEP 1, Fig. 4] was performed for the base set (measurements obtained with a spectroradiometer without additional processing). In the next step, the set was normalized (to achieve this we performed the min-max normalization, and stretched the data in the range of [0, 1]), and the classification was performed again using previously selected algorithms [STEP 2, Fig. 4].

Then, because the distributor of measuring equipment indicated a high, difficult to determine measurement uncertainty in the initial (350–475 nm) and final (2400–2500 nm) range of the measurement spectrum, it was decided to remove this range from the data set, and the classification process was repeated [STEP 3, Fig. 4]. We have chosen not to examine these two narrow spectral regions at either end of the device spectra obtained with our spectrometer. We identify these two parts of the measurements as weaker. The readings in the first spectral region reflect the characteristics of the halogen light source used for this study, which has lower emission in this UV to blue part of the spectrum. At the other end of the measurement, the process is weaker due to the characteristic of the spectrometer employed, which uses an InGaAs sensor. Therefore, in order to avoid providing data with a lower signal-to-noise ratio, we omit these two extreme parts of the measurements. As a consequence, we consider the omission of these spectra as a limitation of this study. Nevertheless, in the following investigation we plan to address them, including additional blue LED array illuminations that have been reported helpful[82] as well as additional calibration targets included in the process[83].

The next step was to limit the feature space in the classification process. For this purpose, the Recursive Feature Elimination (RFE) method was used with the overall goal of reducing the feature space to 50 spectral bands at the end of STEP 4 and, if necessary, with a deeper reduction in STEP 5. We therefore applied two stages of the elimination process. The first feature elimination stage [STEP 4, Fig. 4] exploited the RFE to limit the number of features to 50 bands with recursive elimination of 50 features in each subsequent iteration. As a result, 50 spectral features (bands) were selected for later inclusion in the final classifier, for which the F_1 measure was calculated.

The second stage [STEP 5, Fig. 4] applied the dynamic elimination of the features, with the difference from STEP 4, that the F_1 score measure was determined after the elimination of each data feature. Based on the F_1 score measure, the best elimination stage was determined for each of the classification cases considered. The step of 50 features for each iteration of the process was chosen arbitrarily at first, as it both provided the expected performance improvements and did not significantly slow down the entire process. We later also tested other options for this parameter, but for shorter steps, the number of required computations increased substantially, slowing down the computations substantially, and for larger step sizes, we noted the accuracy loss. Therefore, we stayed with the 50 features step.

During this process, the cross-validation method was used. We performed data splitting for cross-validation with respect to the plant diseases studied. We aimed to distribute the samples for each disease and control group evenly between the subsets. This also meant that the vegetative chambers were equally represented, as the plants inoculated with different diseases were planted separately. Secondly, we also stratified the data samples collected in the following days across the subsets. However, we restricted the placement of scans of the same leaf in the same fold to avoid later processing very similar scans in the same algorithm training phase.

The last stage [STEP 6, Fig. 4] was to fine-tune the hyperparameters of the individual classifiers for the best step from STEP 5 by applying the method of randomly searching for optimal hyperparameters using cross-validation. To

address the disease detection task, we applied several well-established machine learning algorithms that fulfill two main criteria: they allow achieving high performance and support model interpretability by exploring the importance of features. An essential step in the development of these detectors was the search for their advantageous hyperparameter configuration, therefore we tested the following parameters for the following models.

- LGBM: with learning rates of the range from 10^{-6} to 10^{-1}, minimal number of required samples per trained tree leaf between 5 and 100 with step 5, and with the maximum allowed depth of each trained tree from 2 to 52 with step 5,
- Linear SVC: with regularisation parameter C between 1 and 1000 and with L_1 or L_2 regularisation,
- LR: with regularisation parameter C between 1 and 1000 and with L_1 or L_2 regularisation,
- RF: with the maximum number of submodels included from 200 to 2000 with step 100, with the maximum allowed depth of each trained tree from 2 to 52 with step 5, with the minimum required number of samples at each tree leaf node between 1 and 16, and minimum number of samples needed to perform a node split between 2 and 10,
- RC: with learning rates of the range from 10^{-4} to 10^{0},
- XGB: with learning rates of the range from 10^{-4} to $10^{-}1$, with allowed percentage of samples to be used per each tree between $1/5$ and $4/5$ with step $1/5$, with the maximum allowed depth of each trained tree from 2 to 52 with step 5, and with the maximum number of submodels included from 200 to 2000 with step 100.

The results presented below correspond to the activities performed at each stage. Results for all the evaluated models are presented in Table 2) and are respectively:

- result 0—tidy dataset,
- result 1—tidy dataset and min-max normalization,
- result 2—tidy dataset and elimination of marginal ranges of the reflectance spectrum,
- result 3—tidy dataset and min-max normalization and result for the full recursive feature elimination process,
- result 4—tidy dataset and min-max normalization and the best result obtained in the process of recursive feature elimination,
- result 5—tidy dataset and min-max normalization and the best result obtained in the process of recursive feature elimination and fine-tuning of the classifier hyperparameters.

The ALL column presents the binary classification results (F_1 score) distinguishing healthy plants from sick plants, regardless of the disease (CS vs AN + BS + EB + LB + SL). The following columns are the results of the classification of individual diseases in relation to the control sample (CS vs AN, CS vs BS, CS vs EB, CS vs LB, CS vs SL, respectively).

All used classifiers were trained on the same training dataset (2/3 of all measurements; n = 854 samples) and tested on a separate test dataset (1/3 of collected measurements; n = 420 samples) that did not participate in the learning process. Training and test sets for all experiment variants have been balanced, in that sense they were built of 50% of samples of infected plants and 50% of the control group.

For the part of the process where we are working on submodels for individual infections (i.e. CS vs. AN, and so on), the detection is based only on data for that pathogen and the control set. Therefore, at this stage, a smaller subset of samples is used sequentially, using only data for specific infections and control data at the same time.

In almost each of the experiment stages, the classification performance improved. Only STEP 3 (elimination of the marginal ranges of the reflectance spectrum) resulted in worse scores. For the ALL classification (CS vs. AN + BS + EB + LB + SL) the outcome of the classification was 0.879.

The detection of BS, EB, and LB diseases was characterized by slightly lower effects (0.872, 0.877, and 0.866, respectively), while the remaining diseases were detected with greater efficiency (AN: 0.894, SL: 0.896). After performing the entire procedure described, including the hyperparameter tuning, the best result improvements were observed for logistic regression (in 4 cases) and linear SVC (in 2 cases). The smallest improvement in the classification efficiency was obtained in the case of AN, and it was 0.084. On the other hand, the greatest improvement of 0.144 was achieved in the case of BS.

6 Discusion

Another goal of the study was to determine the participation of individual spectral bands (model features) in the decision-making process. Depending on the used classifier, feature weights ("*coef_*" attribute) or feature significance coefficients ("*feature_importances_*" attribute) were defined. The coefficients mentioned above can be treated as indicators of the significance of features for such classifiers as inter alia, linear regression, or ridge classifier. The determined feature weights were used for the RidgeClassifier, linear SVC and logistic regression classifiers, while the feature significance coefficients were used for the Random Forest, XGBClassifier, and LGBMClassifier classifiers.

The recursive feature elimination (RFE) method was used in the conducted study, using the levels mentioned earlier

Table 2 Training results for selected classifiers at various study steps for individual or all disease data combined (ALL). The resulting F_1 scores for the following steps are listed in columns ($F_1^{res_0}$, …, $F_1^{res_5}$)

Disease	The best classifier	$F_1^{res_0}$	$F_1^{res_1}$	$F_1^{res_2}$	$F_1^{res_3}$	$F_1^{res_4}$	$F_1^{res_5}$
ALL	LGBM	0.71	0.80	0.80	0.81	0.81	0.83
	Linear SVC	0.78	0.78	0.72	0.85	0.85	0.86
	LR	0.75	0.77	0.76	0.80	0.80	0.88
	RF	0.69	0.79	0.79	0.79	0.79	0.79
	RC	0.78	0.81	0.80	0.83	0.83	0.86
	XGB	0.69	0.80	0.81	0.83	0.83	0.79
AN	LGBM	0.75	0.81	0.80	0.84	0.85	0.83
	Linear SVC	0.81	0.81	0.79	0.78	0.88	0.89
	LR	0.78	0.82	0.81	0.83	0.84	0.88
	RF	0.71	0.78	0.79	0.80	0.82	0.82
	RC	0.82	0.84	0.83	0.85	0.86	0.88
	XGB	0.73	0.81	0.79	0.83	0.83	0.81
BS	LGBM	0.74	0.80	0.79	0.80	0.81	0.80
	Linear SVC	0.78	0.77	0.78	0.81	0.83	0.87
	LR	0.73	0.77	0.77	0.78	0.79	0.87
	RF	0.72	0.76	0.77	0.78	0.81	0.80
	RC	0.78	0.81	0.78	0.80	0.80	0.86
	XGB	0.74	0.78	0.78	0.80	0.81	0.77
EB	LGBM	0.70	0.82	0.80	0.84	0.85	0.85
	Linear SVC	0.79	0.82	0.78	0.86	0.86	0.88
	LR	0.74	0.78	0.77	0.81	0.82	0.87
	RF	0.69	0.79	0.78	0.80	0.84	0.82
	RC	0.78	0.81	0.79	0.82	0.83	0.85
	XGB	0.70	0.81	0.80	0.85	0.86	0.85
LB	LGBM	0.75	0.82	0.80	0.83	0.83	0.83
	Linear SVC	0.79	0.79	0.71	0.83	0.84	0.85
	LR	0.74	0.77	0.77	0.79	0.80	0.87
	RF	0.72	0.80	0.80	0.83	0.83	0.82
	RC	0.78	0.80	0.79	0.81	0.81	0.86
	XGB	0.73	0.81	0.80	0.83	0.84	0.82
SL	LGBM	0.72	0.82	0.81	0.86	0.86	0.84
	Linear SVC	0.83	0.84	0.83	0.86	0.87	0.89
	LR	0.76	0.81	0.79	0.82	0.83	0.90
	RF	0.71	0.80	0.80	0.80	0.83	0.81
	RC	0.83	0.85	0.85	0.86	0.86	0.89
	XGB	0.72	0.82	0.81	0.84	0.84	0.82

of the significance of features and feature coefficients. The method is based on the iterative elimination of successive, least significant features of the data set. In the case of the considered data set, these were data columns representing successive lengths of the reflectance spectrum. Fifty features that showed the least significance in the decision-making process in each subsequent iteration were removed from the set. Then the classifier was retrained on a smaller data set. The elimination process ended when the set was limited to 50 traits. Depending on the data set under consideration, the whole process consisted of 39 or 44 elimination steps. Each feature was assigned a ranking index that allowed it to determine the feature's elimination stage. For each of the stages, the value of the F_1 score measure was also determined for the test data set.

Based on the F_1 score measure, the step for which the highest value of the indicator was obtained was determined. In this way, all the bands involved in the decision-making process were listed. Then it was determined how often particular features were taken into account by a given classifier. On this basis, a graph of the share of spectral features in the decision-making process was created.

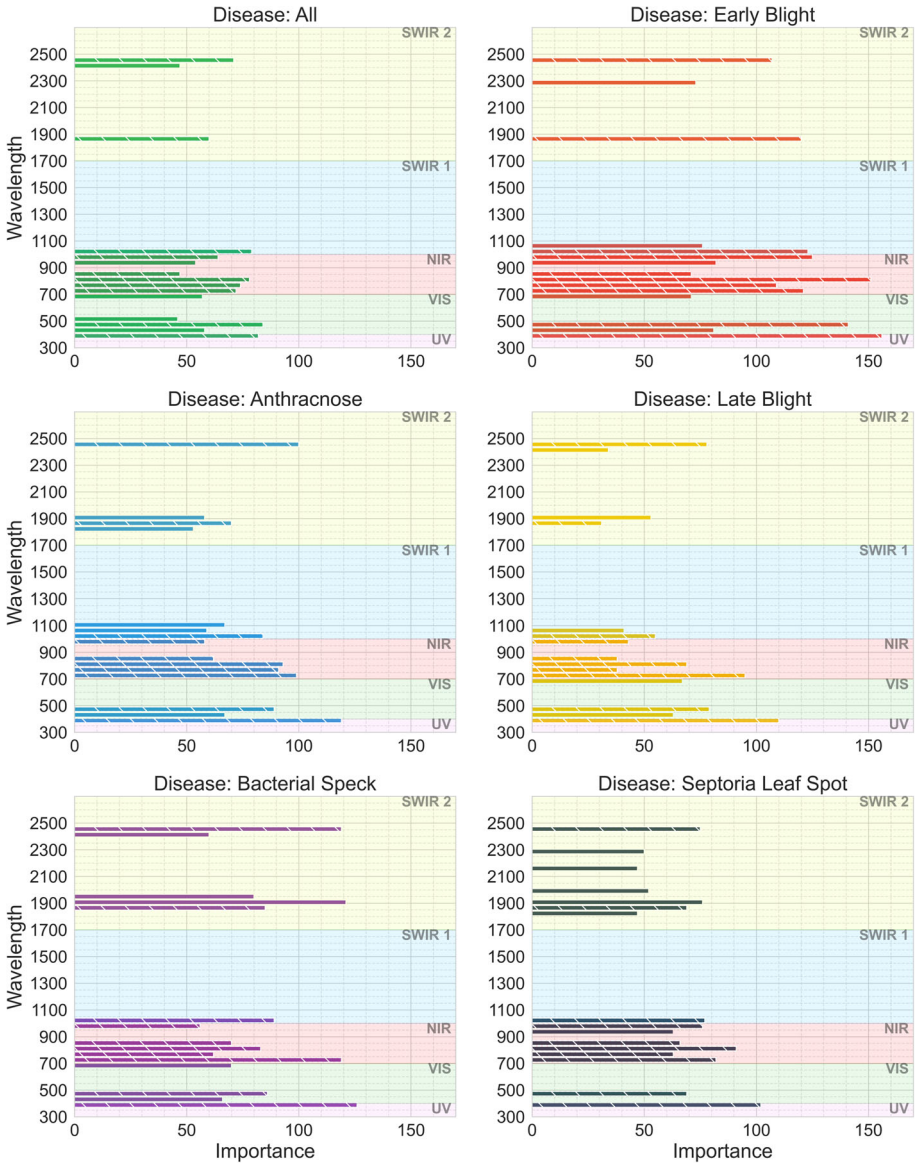


Fig. 5 The importance of particular ranges of the reflectance spectrum in the process of disease classification. The white dashed line marks the overlapping spectral ranges for all the analyzed diseases

As part of an in-depth analysis, it was decided to group the adjacent spectral ranges (50 bands) and to define the most frequently repeated spectral ranges (70% of the most rarely occurring bands were removed).

On the basis of the conducted research, the following spectrum ranges for the analyzed diseases were obtained (the numbers provided indicate the measures of the determined ranges of the reflectance spectrum with a width of around 43 [nm]):

- AN: 371, 413, 455, 708, 751, 793, 835, 961, 1004, 1046, 1088, 1805, 1847, 1889, 2438 [nm],
- BS: 371, 413, 455, 666, 708, 751, 793, 835, 961, 1004, 1847, 1889, 1932, 2395, 2438 [nm],
- EB: 371, 413, 455, 666, 708, 751, 793, 835, 919, 961, 1004, 1046, 1847, 2269, 2438 [nm],
- LB: 371, 413, 455, 666, 708, 751, 793, 835, 961, 1004, 1046, 1847, 1889, 2395, 2438 [nm],
- SL: 371, 455, 708, 751, 793, 835, 919, 961, 1004, 805, 1847, 1889, 1974, 2142, 2269, 2438 [nm],
- ALL: 371, 413, 455, 498, 666, 708, 751, 793, 835, 919, 961, 1004, 1847, 2395, 2438 [nm].

The above-mentioned spectral reflectance ranges are illustrated in Fig. 5.

As shown in Fig. 5, several ranges of the reflectance spectrum cover all analyzed diseases and the case where the control was compared with a pooled set of samples of all diseases. These are the following ranges: 371, 455, 708, 751, 793, 835, 961, 1004, 1847, 2438 [nm]. The conducted study also indicated spectral ranges specific for each of the ana-

Table 3 The specific identified reflectance spectrum ranges for the classification of studied diseases

The reflectance spectrum range [nm]	Disease
413	AN, BS, EB, LB
666	BS, EB, LB
919	EB, SL
1046	AN, EB, LB
1088	AN
1805	AN, SL
1889	AN, BS, LB, SL
1932	BS
1974	SL
2142	SL
2269	EB, SL
2395	BS, LB

lyzed diseases, and the relevant summary is presented in Tab. 3.

The conducted study indicated several specific ranges of the reflectance spectrum, including:

- Spectral bands influential in the classification of three or four examined diseases (413, 666, 1046, 1889 [nm]),
- Bands characteristic of two examined diseases: EB, SL (919, 2269 [nm]), AN, SL (1805 [nm]), BS, LB (2395 [nm]),
- Spectral ranges specific for one analyzed disease: 1088 [nm] (AN), 1932 [nm] (BS), 1974 and 2142 [nm] (SL).

Additionally, the spectral ranges mentioned in the literature review were compared with the results obtained during the research for LB, BS, and EB diseases. The summary of the results is visualized in Fig. 6.

The spectral band ranges indicated in the literature review were compared with ranges obtained during the described research. The range of visible and near-infrared light appears in the cited publications and the authors' results. The conducted research also indicated other bands of the reflectance spectrum, significant in the classification of the studied diseases of *Solanum lycopersicum*, located outside the areas mentioned above. These are, among others:

- for LB: 1847, 1889, 2395, 2438 [nm] (the P1 publication indicates: 2100 [nm]),
- for BS: 1847, 1889, 1932, 2395, 2438 [nm],
- for EB: 1847, 2269, 2438 [nm].

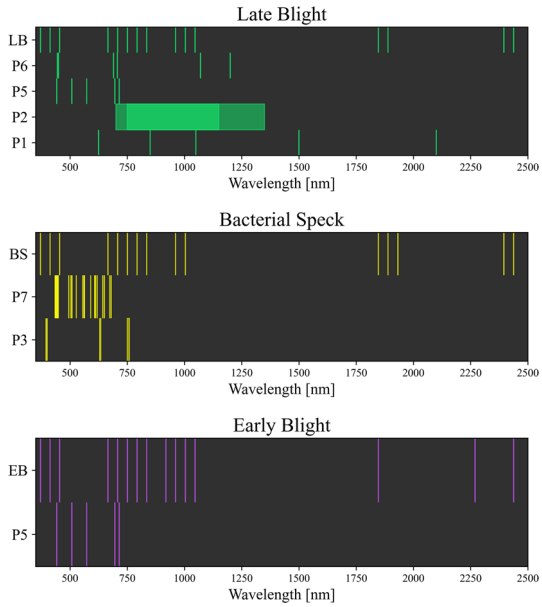


Fig. 6 Comparative report concerning the spectral bands indicated in the literature review (P1-P6—publications described in the introduction to the article, LB, BS, EB—results of the experiment described in this article)

7 Conclusion

The research confirmed that spectroscopy reflectance measurements could be used to detect some *Solanum lycopersicum* diseases. Regarding objective 1, the best results were obtained for the Logistic Regression and Linear SVC classifiers. The Logistic Regression classifier achieved the highest F_1 score for *Septoria Leaf Spot* disease, amounting to 0.896 after hyperparameter adjustment.

The procedure proposed in the article (Stage 1–6), the analysis of which was the second main objective of the study, resulted in an improvement in the classification results in almost every analyzed case—for the *Bacterial Speck disease*, the most significant improvement of the F_1 score metric was obtained (from 0.728 to 0.877).

The performed analysis of the significance of the spectral bands in the classification of diseases (objective no. 3) indicated overlapping reflectance ranges for various diseases (Fig. 6). Additionally, different and specific reflectance ranges were indicated, which were particularly important in classifying specific diseases (Table 2). The indicated ranges can be used to develop new methods and measuring instruments dedicated to diagnosing appropriate *Solanum lycopersicum* diseases and after further research into other crops' diagnoses.

The method described in the article can also be applied to other crops and their diseases. However, this requires the development of appropriate, dedicated data sets consisting of measurements made in laboratory or field conditions.

Based on spectroscopy, the presented diagnostic method can be attempted to be transferred to aviation platforms (i.e., unmanned aerial vehicles, manned aircraft, satellites) [24]. The conducted literature review and the research described in the article show that collecting the most extensive possible set of measurement data, covering various disease cases, may allow for the creation of quick and possibly reliable methods of plant disease diagnostics. In subsequent scientific and research works, the authors plan to attempt to classify diseased plants regardless of the stage of disease development, i.e., at the stage when symptoms are not visible (analysis of the classification effectiveness in the days following infection). It also seems advisable to use the constructed dataset to analyze the effectiveness of disease classification using various artificial neural network (ANN) models, including Deep Learning.

Author Contributions BR: Conceptualization, methodology, software, validation, investigation, writing—original draft, Writing—review and editing, supervision, funding acquisition. KS: conceptualization, methodology, software, validation, formal analysis, literature review, investigation, writing—original draft, visualization, funding acquisition. MT: conceptualization, methodology, validation, formal analysis, literature review, investigation, writing—original draft, supervision. PJN: validation, writing—original draft, writing - review and editing, funding acquisition.

Funding This work was partially supported by The National Centre of Research and Development of Poland under project POIR.01.01.01−00. 1317/17 and QZ Solutions sp. z o.o.

Availability of data and materials The data that support the findings of this study are available from the authors but restrictions apply to the availability of these data. Data are, however, available from the corresponding author upon reasonable request and with permission from the QZ Solutions.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Foolad, M.R.: Genome mapping and molecular breeding of tomato. Int. J. Plant Genom. (2007). https://doi.org/10.1155/2007/64358
2. Nawrocka, B., Robak, J., Ślusarski, C., Macias, W.: Choroby i Szkodniki Pomidora W Polu i Pod Osłonami. Wydawnictwo Plantpress Sp. z o.o, Kraków (2001)
3. Datar, V.V., Mayee, C.D.: Conidial dispersal of *Alternaria solani* in tomato. Indian Phytopathol. **35**, 68–70 (1982)
4. Grigolli, J.F.J., Kubota, M.M., Alves, D.P., Rodrigues, G.B., Cardoso, C.R., Silva, D.J.H., Mizubuti, E.S.G.: Characterization of tomato accessions for resistance to early blight. Crop Breed. Appl. Biotechnol. **11**(2), 174–180 (2011). https://doi.org/10.1590/S1984-70332011000200010
5. Jin, X., Jie, L., Wang, S., Qi, H., Li, S.: Classifying wheat hyperspectral pixels of healthy heads and *Fusarium* head blight disease using a deep neural network in the wild field. Remote Sens. **10**(3), 395 (2018). https://doi.org/10.3390/rs10030395
6. Khan, A., Vibhute, A.D., Mali, S., Patil, C.H.: A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications. Ecol. Inf. **69**, 101678 (2022). https://doi.org/10.1016/j.ecoinf.2022.101678
7. Polder, G., Gowen, A.: The hype in spectral imaging. J. Spectr. Imaging **9**(1), 4 (2020). https://doi.org/10.1255/jsi.2020.a4
8. Tomaszewski, M., Gasz, R., Smykała, K.: Monitoring vegetation changes using satellite imaging - NDVI and RVI4S1 indicators. In: Paszkiel, S. (ed.) Control, Computer Engineering and Neuroscience, pp. 268–278. Springer, Cham (2021)
9. Jones, H.G., Vaughan, R.A.: Remote Sensing of Vegetation: Principles. Techniques and Applications, Oxford University Press, Oxford (2010)
10. Ihuoma, S.O., Madramootoo, C.A.: Recent advances in crop water stress detection. Comput. Electron. Agric. **141**, 267–275 (2017). https://doi.org/10.1016/j.compag.2017.07.026
11. Finn, M.P., Lewis, M.D., Bosch, D.D., Giraldo, M., Yamamoto, K., Sullivan, D.G., Kincaid, R., Luna, R., Allam, G.K., Kvien, C., Williams, M.S.: Remote sensing of soil moisture using airborne hyperspectral data. GIScience Remote Sens. **48**(4), 522–540 (2011). https://doi.org/10.2747/1548-1603.48.4.522
12. Ruszczak, B., Boguszewska-Mańkowska, D.: Deep potato - the hyperspectral imagery of potato cultivation with reference agronomic measurements dataset: towards potato physiological features modeling. Data Brief **42**, 108087 (2022). https://doi.org/10.1016/j.dib.2022.108087
13. Singh, S., Kasana, S.S.: Estimation of soil properties from the EU spectral library using long short-term memory networks. Geoderma Reg. **18**, 00233 (2019). https://doi.org/10.1016/j.geodrs.2019.e00233
14. Tomaszewski, M., Nalepa, J., Moliszewska, E., Ruszczak, B., Smykała, K.: Early detection of *Solanum lycopersicum* diseases from temporally-aggregated hyperspectral measurements using machine learning. Sci. Rep. **13**(1), 7671 (2023). https://doi.org/10.1038/s41598-023-34079-x
15. Marino, S., Beauseroy, P., Smolarz, A.: Weakly-supervised learning approach for potato defects segmentation. Eng. Appl. Artific. Intell. **85**, 337–346 (2019). https://doi.org/10.1016/j.engappai.2019.06.024
16. Reis-Pereira, M., Tosin, R., Martins, R., Santos, F., Tavares, F., Cunha, M.: Kiwi plant canker diagnosis using hyperspectral signal processing and machine learning: detecting symptoms caused by *Pseudomonas syringae* pv. *Actinidiae*. Plants. (2022) https://doi.org/10.3390/plants11162154
17. Naik, B.N., Malmathanraj, R., Palanisamy, P.: Detection and classification of chilli leaf disease using a squeeze-and-excitation-based

CNN model. Ecol. Inf. **69**, 101663 (2022). https://doi.org/10.1016/j.ecoinf.2022.101663
18. Keceli, A.S., Kaya, A., Catal, C., Tekinerdogan, B.: Deep learning-based multi-task prediction system for plant disease and species detection. Ecol. Inf. **69**, 101679 (2022). https://doi.org/10.1016/j.ecoinf.2022.101679
19. Furlanetto, R.H., Nanni, M.R., Mizuno, M.S., Crusiol, L.G.T., da Silva, C.R.: Identification and classification of Asian soybean rust using leaf-based hyperspectral reflectance. Int. J. Remote Sens. **42**(11), 4177–4198 (2021). https://doi.org/10.1080/01431161.2021.1890855
20. Alisaac, E., Behmann, J., Kuska, M.T., Dehne, H.-W., Mahlein, A.-K.: Hyperspectral quantification of wheat resistance to *Fusarium* head blight: comparison of two *Fusarium* species. Eur. J. Plant Pathol. **152**(4), 869–884 (2018). https://doi.org/10.1007/s10658-018-1505-9
21. Mahlein, A.-K.: Detection, identification, and quantification of fungal diseases of sugar beet leaves using imaging and non-imaging hyperspectral techniques. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn (2011). https://hdl.handle.net/20.500.11811/4713
22. Mahlein, A.-K., Kuska, M.T., Thomas, S., Wahabzada, M., Behmann, J., Rascher, U., Kersting, K.: Quantitative and qualitative phenotyping of disease resistance of crops by hyperspectral sensors: seamless interlocking of phytopathology, sensors, and machine learning is needed!. Curr. Opin. Plant Biol. **50**, 156–162 (2019) https://doi.org/10.1016/j.pbi.2019.06.007 . Biotic interactions
23. Kaur, P., Harnal, S., Gautam, V., Singh, M.P., Singh, S.P.: An approach for characterization of infected area in tomato leaf disease based on deep learning and object detection technique. Eng. Appl. Artific. Intell. **115**, 105210 (2022). https://doi.org/10.1016/j.engappai.2022.105210
24. Zhang, M., Qin, Z.: Spectral analysis of tomato late blight infections for remote sensing of tomato disease stress in California. In: IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004, vol. 6, pp. 4091–4094. IEEE, Anchorage, AK, USA (2004). https://doi.org/10.1109/IGARSS.2004.1370031 . http://ieeexplore.ieee.org/document/1370031/
25. Xie, C., Yang, C., He, Y.: Hyperspectral imaging for classification of healthy and gray mold diseased tomato leaves with different infection severities. Comput. Electron. Agric. **135**, 154–162 (2017). https://doi.org/10.1016/j.compag.2016.12.015
26. Moghadam, P., et al.: Plant disease detection using hyperspectral imaging. In: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8. IEEE, Sydney (2017). https://doi.org/10.1109/DICTA.2017.8227476 . http://ieeexplore.ieee.org/document/8227476/
27. Smykała, K., Ruszczak, B., Dziubański, K.: Application of ensemble learning to detect *Alternaria solani* infection on tomatoes cultivated under foil tunnels. Intell. Environ. (2020). https://doi.org/10.3233/AISE200033
28. Ruszczak, B., Smykała, K., Dziubański, K.: The detection of *Alternaria solani* infection on tomatoes using ensemble learning. J. Ambient Intell. Smart Environ. **12**(5), 407–418 (2020). https://doi.org/10.3233/AIS-200573
29. Tasrif Anubhove, M.S., Ashrafi, N., Saleque, A.M., Akter, M., Saif, S.U.: Machine learning algorithm based disease detection in tomato with automated image telemetry for vertical farming. In: 2020 International Conference on Computational Performance Evaluation (ComPE), pp. 250–254. IEEE, Shillong, India (2020). https://doi.org/10.1109/ComPE49325.2020.9200129 . https://ieeexplore.ieee.org/document/9200129
30. Fahrentrapp, J., Ria, F., Geilhausen, M., Panassiti, B.: Detection of gray mold leaf infections prior to visual symptom appearance using a five-band multispectral sensor. Front. Plant Sci. **10**(May), 1–14 (2019). https://doi.org/10.3389/fpls.2019.00628
31. Ferentinos, K.P.: Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. **145**(February), 311–318 (2018). https://doi.org/10.1016/j.compag.2018.01.009
32. Brahimi, M., Boukhalfa, K., Moussaoui, A.: Deep learning for tomato diseases: classification and symptoms visualization. Appl. Artific. Intell. **31**(4), 299–315 (2017). https://doi.org/10.1080/08839514.2017.1315516
33. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Front. Plant Sci. (2016). https://doi.org/10.3389/fpls.2016.01419. arXiv:1604.03169
34. Adhikari, S., KC, E., Balkumari, L., Shrestha, B., Baiju, B.: Tomato plant diseases detection system using image processing. In: Kantipur Engineering College Conference, pp. 81–86 (2018)
35. Xie, C., Shao, Y., Li, X., He, Y.: Detection of early blight and late blight diseases on tomato leaves using hyperspectral imaging. Sci. Rep. **5**(1), 16564 (2015). https://doi.org/10.1038/srep16564
36. Wang, X., Zhang, M., Zhu, J., Geng, S.: Spectral prediction of Phytophthora infestans infection on tomatoes using artificial neural network (ANN). Int. J. Remote Sens. **29**(6), 1693–1706 (2008). https://doi.org/10.1080/01431160701281007
37. Lu, J., Ehsani, R., Shi, Y., Castro, A.I., Wang, S.: Detection of multi-tomato leaf diseases (late blight, target and bacterial spots) in different stages by using a spectral-based sensor. Sci. Rep. **8**(1), 2793 (2018). https://doi.org/10.1038/s41598-018-21191-6
38. Jones, C.D., Jones, J.B., Lee, W.S.: Diagnosis of bacterial spot of tomato using spectral signatures. Comput. Electron. Agric. **74**(2), 329–335 (2010). https://doi.org/10.1016/j.compag.2010.09.008
39. Patil, M.A., Manur, M.: Enhanced radial basis function neural network for tomato plant disease leaf image segmentation. Ecol. Inf. **70**, 101752 (2022). https://doi.org/10.1016/j.ecoinf.2022.101752
40. Reis Pereira, M., Santos, F.N., Tavares, F., Cunha, M.: Enhancing host-pathogen phenotyping dynamics: early detection of tomato bacterial diseases using hyperspectral point measurement and predictive modeling. Front. Plant. Sci. (2023). https://doi.org/10.3389/fpls.2023.1242201
41. Sharma, P., Sharma, S.: Paradigm shift in plant disease diagnostics: a journey from conventional diagnostics to nano-diagnostics. In: Fungal biology, pp. 237–264. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27312-9_11 . http://link.springer.com/10.1007/978-3-319-27312-9_11
42. Golhani, K., Balasundram, S.K., Vadamalai, G., Pradhan, B.: A review of neural networks in plant disease detection using hyperspectral data. Inf. Process. Agric. **5**(3), 354–371 (2018). https://doi.org/10.1016/j.inpa.2018.05.002
43. Sakudo, A., Suganuma, Y., et al.: Near-infrared spectroscopy: promising diagnostic tool for viral infections. Biochem. Biophys. Res. Commun. **341**(2), 279–284 (2006). https://doi.org/10.1016/j.bbrc.2005.12.153
44. Garhwal, A.S., Pullanagari, R.R., Li, M., Reis, M.M., Archer, R.: Hyperspectral imaging for identification of zebra chip disease in potatoes. Biosyst. Eng. **197**, 306–317 (2020). https://doi.org/10.1016/j.biosystemseng.2020.07.005
45. Zhang, J., Tian, Y., Yan, L., Wang, B., Wang, L., Xu, J., Wu, K.: Diagnosing the symptoms of sheath blight disease on rice stalk with an in-situ hyperspectral imaging technique. Biosyst. Eng. **209**, 94–105 (2021). https://doi.org/10.1016/j.biosystemseng.2021.06.020
46. Singh, A., Kaur, J., Singh, K., Singh, M.L.: Deep transfer learning-based automated detection of blast disease in paddy crop. Signal Image Video Process. (2023). https://doi.org/10.1007/s11760-023-02735-4
47. Ruszczak, B., Wijata, A.M., Nalepa, J.: Unbiasing the estimation of chlorophyll from hyperspectral images: a benchmark dataset, validation procedure and baseline results. Remote Sens. (2022). https://doi.org/10.3390/rs14215526

48. Ruszczak, B., Wijata, A.M., Nalepa, J.: Estimating chlorophyll content from hyperspectral data using gradient features. In: Computational Science - ICCS 2023 - 23nd International Conference, Prague, Czech Republic, 3-5 July, 2023, Proceedings. Lecture Notes in Computer Science. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36021-3_18
49. Ruszczak, B.: Reducing high-dimensional feature set of hyperspectral measurements for plant phenotype classification. GECCO '23. ACM, Lisbon (2023).https://doi.org/10.1145/3583133.3596941
50. Navarro, P.J., Miller, L., Díaz-Galián, M.V., Gila-Navarro, A., Aguila, D.J., Egea-Cortines, M.: A novel ground truth multispectral image dataset with weight, anthocyanins, and brix index measures of grape berries tested for its utility in machine learning pipelines. GigaScience (2022). https://doi.org/10.1093/gigascience/giac052
51. Desai, M., Kumar Jain, A., Jain, N.K., Jethwa, K.: Detection and classification of fruit disease : a review. Int. Res. J. Eng. Technol. (2016)
52. Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., Stefanovic, D.: Deep neural networks based recognition of plant diseases by leaf image classification. Comput. Intell. Neurosci. (2016). https://doi.org/10.1155/2016/3289801
53. Islam, M., Anh Dinh, Wahid, K., Bhowmik, P.: Detection of potato diseases using image segmentation and multiclass support vector machine. In: 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1–4. IEEE, London, Canada (2017). https://doi.org/10.1109/CCECE.2017.7946594 . http://ieeexplore.ieee.org/document/7946594/
54. Wang, Q., Qi, F., Sun, M., Qu, J., Xue, J.: Identification of tomato disease types and detection of infected areas based on deep convolutional neural networks and object detection techniques. Comput. Intell. Neurosci. (2019). https://doi.org/10.1155/2019/9142753
55. Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., Ramon, H.: Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. Comput. Electron. Agric. **44**(3), 173–188 (2004). https://doi.org/10.1016/j.compag.2004.04.003
56. Navarro, P.J., Miller, L., Gila-Navarro, A., Díaz-Galián, M.V., Aguila, D.J., Egea-Cortines, M.: 3deepm: an ad hoc architecture based on deep learning methods for multispectral image classification. Remote Sens. (2021). https://doi.org/10.3390/rs13040729
57. Van De Vijver, R., et al.: In-field detection of Alternaria solani in potato crops using hyperspectral imaging. Comput. Electron. Agric. **168**, 105106 (2020). https://doi.org/10.1016/j.compag.2019.105106
58. Georgiev, G.T., Butler, J.J.: Long-term calibration monitoring of Spectralon diffusers BRDF in the air-ultraviolet. Appl. Opt. **46**(32), 7892 (2007). https://doi.org/10.1364/AO.46.007892
59. Wickham, H.: Tidy data. J. Stat. Softw. **59**(10), 1–23 (2014)
60. Wang, B., et al.: An ultra-lightweight efficient network for image-based plant disease and pest infection detection. Precis. Agric. **24**(5), 1836–1861 (2023). https://doi.org/10.1007/s11119-023-10020-0
61. Thangaraj, R., et al.: A deep convolution neural network model based on feature concatenation approach for classification of tomato leaf disease. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-16347-0
62. Kumar, A., Patel, V.K.: Classification and identification of disease in potato leaf using hierarchical based deep learning convolutional neural network. Multimed. Tools Appl. **82**(20), 31101–31127 (2023). https://doi.org/10.1007/s11042-023-14663-z
63. Kurmi, Y., Gangwar, S., Agrawal, D., Kumar, S., Srivastava, H.S.: Leaf image analysis-based crop diseases classification. Signal Image Video Process **15**(3), 589–597 (2021). https://doi.org/10.1007/s11760-020-01780-7
64. Bora, R., Parasar, D., Charhate, S.: A detection of tomato plant diseases using deep learning MNDLNN classifier. Signal Image Video Process **17**(7), 3255–3263 (2023). https://doi.org/10.1007/s11760-023-02498-y
65. Appeltans, S., Pieters, J.G., Mouazen, A.M.: Potential of laboratory hyperspectral data for in-field detection of phytophthora infestans on potato. Precis. Agric. **23**(3), 876–893 (2022). https://doi.org/10.1007/s11119-021-09865-0
66. Omer, S.M., Ghafoor, K.Z., Askar, S.K.: Lightweight improved yolov5 model for cucumber leaf disease and pest detection based on deep learning. Signal Image Video Process (2023). https://doi.org/10.1007/s11760-023-02865-9
67. Bhakta, I., et al.: A novel plant disease prediction model based on thermal images using modified deep convolutional neural network. Precis. Agric. **24**(1), 23–39 (2023). https://doi.org/10.1007/s11119-022-09927-x
68. Haagsma, M., Hagerty, C.H., Kroese, D.R., Selker, J.S.: Detection of soil-borne wheat mosaic virus using hyperspectral imaging: from lab to field scans and from hyperspectral to multispectral data. Precision Agric. **24**(3), 1030–1048 (2023). https://doi.org/10.1007/s11119-022-09986-0
69. Kaur, P., Harnal, S., Gautam, V., Singh, M.P., Singh, S.P.: Performance analysis of segmentation models to detect leaf diseases in tomato plant. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-16238-4
70. Kaur, P., Harnal, S., Gautam, V., Singh, M.P., Singh, S.P.: Hybrid deep learning model for multi biotic lesions detection in solanum lycopersicum leaves. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-15940-7
71. Kang, J., Jin, R., Li, X., Zhang, Y., Zhu, Z.: Spatial upscaling of sparse soil moisture observations based on ridge regression. Remote Sens. (2018). https://doi.org/10.3390/rs10020192
72. Xu, C., Ding, J., Qiao, Y., Zhang, L.: Tomato disease and pest diagnosis method based on the Stacking of prescription data. Comput. Electron. Agric. **197**, 106997 (2022). https://doi.org/10.1016/j.compag.2022.106997
73. Panchal, P., Raman, V.C., Mantri, S.: Plant diseases detection and classification using machine learning models. In: CSITSS 2019 - 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution, Proceedings. Institute of Electrical and Electronics Engineers Inc. (2019). https://doi.org/10.1109/CSITSS47250.2019.9031029
74. Boguszewska-Mańkowska, D., Ruszczak, B., Zarzyńska, K.: Classification of potato varieties drought stress tolerance using supervised learning. Appl. Sci. (2022). https://doi.org/10.3390/app12041939
75. Srinivas, L.N.B., Bharathy, A.M.V., Ramakuri, S.K., Sethy, A., Kumar, R.: An optimized machine learning framework for crop disease detection. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-15446-2
76. Ruszczak, B., Boguszewska-Mańkowska, D.: Soil moisture a posteriori measurements enhancement using ensemble learning. Sensors (2022). https://doi.org/10.3390/s22124591
77. Xie, Y., Plett, D., Liu, H.: The promise of hyperspectral imaging for the early detection of crown rot in wheat. AgriEngineering **3**(4), 924–941 (2021). https://doi.org/10.3390/agriengineering3040058
78. Nguyen, C., Sagan, V., Maimaitiyiming, M., Maimaitijiang, M., Bhadra, S., Kwasniewski, M.T.: Early detection of plant viral disease using hyperspectral imaging and deep learning. Sensors (Basel) **21**(3), 742 (2021)
79. Mourtzinis, S., Esker, P.D., Specht, J.E., Conley, S.P.: Advancing agricultural research using machine learning algorithms. Sci. Rep. **11**(1), 17879 (2021). https://doi.org/10.1038/s41598-021-97380-7
80. Wicaksono, P., Aryaguna, P.A., Lazuardi, W.: Benthic habitat mapping model and cross validation using machine-learning classification algorithms. Remote Sens. (2019). https://doi.org/10.3390/rs11111279

81. Nagasubramanian, K., et al.: Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. Plant Methods **14**(1) (2018). https://doi.org/10.1186/s13007-018-0349-9. arXiv:1710.04681
82. Mahlein, A.-K., Hammersley, S., Oerke, E.-C., Dehne, H.-W., Goldbach, H., Grieve, B.: Supplemental blue led lighting array to improve the signal quality in hyperspectral imaging of plants. Sensors **15**(6), 12834–12840 (2015). https://doi.org/10.3390/s150612834
83. Crusiol, L.G.T., Nanni, M.R., Silva, G.F.C., et al.: Semi professional digital camera calibration techniques for VIS/NIR spectral data acquisition from an unmanned aerial vehicle. Int J Remote Sens **38**(8–10), 2717–2736 (2017). https://doi.org/10.1080/01431161.2016.1264032

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Załącznik 3 – [A3]



Reducing high-dimensional feature set of hyperspectral measurements for plant phenotype classification

Bogdan Ruszczak
b.ruszczak@po.edu.pl
Faculty of Electrical Engineering, Automatic Control and Informatics, Opole University of Technology
Opole, Opolskie, Poland

ABSTRACT

Quantifying the drought resistance of potato cultivars plays a key role in precision agriculture, and it may lead to the development of new varieties of plants that are more resistant to harsh environmental conditions. In this work, we tackle the issue of extracting such information in a non-invasive way by acquiring in-field hyperspectral measurements of the potato leaves. Then, we exploit an array of machine learning models to classify plants into three wilting classes based on such data, with those classes corresponding to their drought resistance. We show that evolutionary band selection can dramatically reduce the dimensionality of hyperspectral data while improving classification accuracy. Our experimental study revealed that the evolutionarily-optimized models offer high-quality performance with the impartial r_ϕ reaching 0.784, accuracy: 0.867, and a 30% improvement over the baseline models which do not benefit from band selection.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; **Genetic algorithms**; • **Applied computing** → Bioinformatics.

KEYWORDS

genetic algorithm, machine learning, spectrophotometer, hyperspectral measurements, feature selection, plants' phenotyping

ACM Reference Format:
Bogdan Ruszczak. 2023. Reducing high-dimensional feature set of hyperspectral measurements for plant phenotype classification. In *Proceedings of The Genetic and Evolutionary Computation Conference 2023 (GECCO '23).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3583133.3596941

1 INTRODUCTION

In the field of agronomy, many experts aim to develop new varieties of plants. Their goal is to provide such new phenotypes of plants that would be more resistant to specific environmental conditions, i.e. chilling, drought, or biotic stress like insects and fungi-related diseases. Such a process has been described in [1], where, in addition, for the new candidate varieties of potato, a machine learning (ML) model has been proposed that allows one to classify the group of drought resistance of each variety. Although a fairly good accuracy of 96.7% has been reported for the 50 potato varieties studied (new candidate varieties), the approach presented there has a serious limitation. The described classifier makes it possible to accurately mimic the responses of plants to drought stress but on the basis of measurements taken after the agronomic season (post-harvest). This means that the whole process of obtaining information about the developed cultivar is relatively long, and the feedback for the next-generation preparation comes late.

To overcome this challenge, another methodology has been proposed. With the hope that each of the candidate varieties would show some drought tolerance symptoms during their growth, it was decided to monitor their leaf system. For this purpose, the high-resolution wide-range radio hyperspectrometer was used, which provides the reflectance information on the leaves of the plants. Later that year, the acquired spectral scans were linked to the reference information on the plants' actual drought resistance. The resulting data set consisted of the collection of scans related to one of 3 wilting classes, representing each of the potato cultivars. Those were ranging from least resistant to drought (class $\Delta = 0$; significant wilting) to moderately resistant ($\Delta = 1$; moderate wilt.) to most drought tolerant ($\Delta = 2$; low wilt.). The challenge in processing this data set was the high-dimensional nature of this data collection. For a noticeably higher number of the model training features, of which there were 2151, than the number of candidate cultivars scans, of which there were only 216 training samples.

2 OVERVIEW OF THE PROCESS

The data collection that was studied in this research was the result of the scans performed with the ASD FieldSpec 4 Hi-Res radiospectrophotometer, which provided leaf reflectance in the range of 350 to 2500 [nm] and spectral resolution of 1[nm], with redundant calibration between each scanning round of the 50 different cultivars. The plants were prepared at the Plant Breeding and Acclimatization Institute-National Research Institute, Serock, Poland. The resulting spectral curves are depicted in Figure 1, where the median value for each class is also shown. The exact values for the selected spectral bands (including the visible light spectra) are given in Table 1, with their differentiation for the studied classes, with the number of samples per class. The number of measurements performed was 2160, and each of the examined samples in the set was an aggregation of the median value of the 10 component scans.

The task of class separation for the spectral scans turned out to be non-trivial. As can be seen in Figure 1 for most of the spectral range the measurements overlap, although, for some subranges and some classes, their separation is possible using classical thresholding (1b). For the classification of more difficult data samples, it is advisable

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
GECCO '23 Companion, July 15–19, 2023, Lisbon, Portugal
© 2023 Copyright is held by the owner/author(s).
ACM ISBN 979-8-4007-0120-7/23/07…$15.00
https://doi.org/10.1145/3583133.3596941

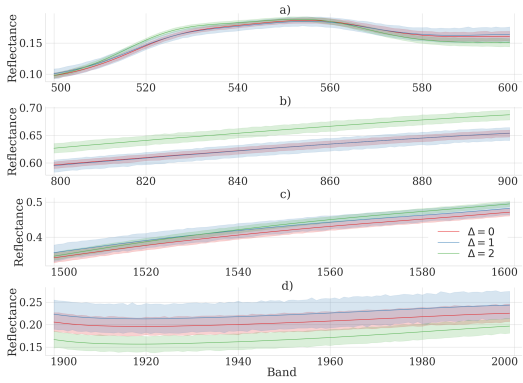


Figure 1: Measured reflectance for selected spectral ranges and 3 different wilting classes with their median (solid line).

Table 1: Wilting class distribution and the median values of measured leaf reflectance for several selected bands.

Class	n	$\overline{\mu}_{435}$	$\overline{\mu}_{546}$	$\overline{\mu}_{700}$	$\overline{\mu}_{850}$	$\overline{\mu}_{1550}$	$\overline{\mu}_{1950}$
$\Delta = 0$	108	0.068	0.182	0.282	0.628	0.419	0.204
$\Delta = 1$	24	0.070	0.181	0.284	0.628	0.431	0.221
$\Delta = 2$	84	0.069	0.186	0.267	0.660	0.436	0.166

to apply a method capable of selecting subsets of data, and for example, in [5] the application of decision trees and RF was reported with accuracies between 0.78 and 0.98 for HS measurements. The paper [4] suggests the use of ensemble learning for HS signals, and reports, for instance, the experiments on tomatoes inoculated with the pathogen Alternaria solani and studied with a similar methodology, for which the training process and the adjustment of the hyperparameters resulted in a classification accuracy of 0.987, suggesting also to avoid too high data dimensionality in the case of a lower number of examples, as it could lead to poor performance.

3 EXPERIMENTS

The experiments comprised of the following steps: first, two baseline models were trained to provide a reference for later comparison, then, for various configurations, the evolutionary computation was evaluated for reducing the feature space. All ML training rounds were evaluated with the same initial shuffle order, using 5-fold cross-validation and the r_ϕ as the main performance metric, as it employs all outcomes of the confusion matrix for the assessed model. The very first ML model tested was built upon an ensemble model (RF). Although the main parameters of the model were tuned, the classification accuracy was quite low (see Table 2. A slightly better outcome was noted for the linear model, with r_ϕ=0.520, which was improved using L_2 regularisation, but still, probably due to the opposite of the widely recommended ratio of the number of features to the number of rows in the input training data set.

To improve the performance, the evolutionary approach was used to reduce the number of features (i.e., to perform band selection [3]). This was done with the following initial configurations: *(i)* with no restriction to the number of used parameters: $N_F \in 100\%$,

Table 2: Figures for baseline models and the evolutionary approach *EA*, with the best setting marked as bold, and returned number of features N_F^* and α for L_2 regularisation.

Experiment	α	r_ϕ	F_1	Acc.	N_F^*
Baseline RF:max_{dep}=6	—	0.412	0.615	0.662	2150
Baseline Linear + L_2	10^{-1}	0.457	0.653	0.694	2150
EA: $(\mu + \lambda)$, $N_F \in 100\%$	10^{-4}	0.760	0.825	0.857	810
EA: $(\mu + \lambda)$, $N_F < 50\%$	10^{-4}	0.777	0.830	0.866	789
EA: (μ, λ), $N_F < 37.5\%$	10^{-8}	**0.784**	**0.864**	**0.867**	**783**
EA: $(\mu + \lambda)$, $N_F < 25\%$	10^{-4}	0.755	0.813	0.852	519

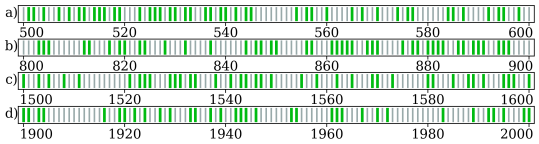


Figure 2: A preview of the bands returned by the EA algorithm for the selected set of spectral ranges, marked in green.

(ii) with a strict limit to a maximum of half of the features: $N_F < 50\%$, and *(iii)* an even more aggressive limit to a maximum of 37.5% and *(iv)* 25% of available data set columns. Each training was run twice, with different evolution strategy: $(\mu + \lambda)$ or (μ, λ) [2]. The results of the experiments are shown in Table 2. It is worth noting that for all the experiments a better efficiency has been noted compared to the baselines, and all the models remained the tendency to finish with a similar number of features, regardless of the setting, and the best classifier for this research reached a decent value r_ϕ of 0.784. Bands selected by this model are illustrated in Figure 2.

4 CONCLUSIONS

The performance statistics of the final model based on the reduced feature space are promising. The next step for its development should be a deeper study of the found subset of spectral bands and their relation to wilting drought stress response of plants. Other activities could include the investigation of data pre-processing, such as Savitzky-Golay filtering or signal differentiation, and the application of more sophisticated ML methods, such as extreme gradient boosting or deep learning approaches to the reduced data set, and the evolutionary optimization for the entire future setup.

REFERENCES

[1] Boguszewska-Mańkowska Dominika, Ruszczak Bogdan, and Zarzyńska Krystyna. 2022. Classification of Potato Varieties Drought Stress Tolerance Using Supervised Learning. Applied Sciences 12, 4 (2022). https://doi.org/10.3390/app12041939
[2] Sean Luke. 2013. Essentials of Metaheuristics (second ed.). Lulu. Available at http://cs.gmu.edu/~sean/book/metaheuristics/.
[3] Pablo Ribalta Lorenzo et al. 2020. Hyperspectral Band Selection Using Attention-Based Convolutional Neural Networks. IEEE Access 8 (2020), 42384–42403. https://doi.org/10.1109/ACCESS.2020.2977454
[4] Ruszczak Bogdan, Smykała Krzysztof, and Dziubański Karol. 2020. The detection of Alternaria solani infection on tomatoes using ensemble learning. J Ambient Intell Smart Environ 12, 5 (2020), 407–418. https://doi.org/10.3390/app12041939
[5] Smykała Krzysztof, Ruszczak Bogdan, and Dziubański Karol. 2020. Application of Ensemble Learning to Detect Alternaria Solani Infection on Tomatoes Cultivated Under Foil Tunnels. Ambient Intelligence and Smart Environments 28 (2020), 127–132. https://doi.org/10.3233/AISE200033

Załącznik 4 – [A4]

remote sensing

MDPI

Article

Unbiasing the Estimation of Chlorophyll from Hyperspectral Images: A Benchmark Dataset, Validation Procedure and Baseline Results

Bogdan Ruszczak [1,2], Agata M. Wijata [2,3] and Jakub Nalepa [2,4,*]

[1] Faculty of Electrical Engineering, Automatic Control and Informatics, Department of Informatics, Opole University of Technology, Prószkowska 76, 45-758 Opole, Poland
[2] KP Labs, Konarskiego 18C, 44-100 Gliwice, Poland
[3] Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 41-800 Zabrze, Poland
[4] Faculty of Automatic Control, Electronics and Computer Science, Department of Algorithmics and Software, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
* Correspondence: jnalepa@ieee.org

Abstract: Recent advancements in hyperspectral remote sensing bring exciting opportunities for various domains. Precision agriculture is one of the most widely-researched examples here, as it can benefit from the non-invasiveness and enormous scalability of the Earth observation solutions. In this paper, we focus on estimating the chlorophyll level in leaves using hyperspectral images—capturing this information may help farmers optimize their agricultural practices and is pivotal in planning the plants' treatment procedures. Although there are machine learning algorithms for this task, they are often validated over private datasets; therefore, their performance and generalization capabilities are virtually impossible to compare. We tackle this issue and introduce an open dataset including the hyperspectral and in situ ground-truth data, together with a validation procedure which is suggested to follow while investigating the emerging approaches for chlorophyll analysis with the use of our dataset. The experiments not only provided the solid baseline results obtained using 15 machine learning models over the introduced training-test dataset splits but also showed that it is possible to substantially improve the capabilities of the basic data-driven models. We believe that our work can become an important step toward standardizing the way the community validates algorithms for estimating chlorophyll-related parameters, and may be pivotal in consolidating the state of the art in the field by providing a clear and fair way of comparing new techniques over real data.

Dataset: DOI:10.1016/j.dib.2022.108087.

Dataset License: The license under which the data set is made available is (CC-BY).

Keywords: hyperspectral imaging; machine learning; chlorophyll estimation; benchmark; validation

check for updates

Citation: Ruszczak, B.; Wijata, A.M.; Nalepa, J. Unbiasing the Estimation of Chlorophyll from Hyperspectral Images: A Benchmark Dataset, Validation Procedure and Baseline Results. *Remote Sens.* **2022**, *14*, 5526. https://doi.org/10.3390/rs14215526

Academic Editors: Yuhong He and Anita Simic-Milas

Received: 9 September 2022
Accepted: 31 October 2022
Published: 2 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent advancements in sensor technology bring new possibilities in hyperspectral image (HSI) analysis—such data effectively captures hundreds of spectral bands in the electromagnetic spectrum. In precision agriculture, acquiring detailed information concerning the chlorophyll saturation lets the plant breeders optimize their operation and plan the plants' treatment. Because the chlorophyll fluorescence, which is induced by solar radiation, is a direct representative of the actual vegetation photosynthesis, it is also the main vegetation performance indicator [1]. Therefore, monitoring the chlorophyll fluorescence parameters could bring important information on plants' stress or help us to detect the moment of the crop photosynthesis termination [2]. Furthermore, such measurements are valuable because they could make us understand the plant response to herbicide treatments and enable us to react quickly to possible plant conditions' changes. Additionally,

the exposition to toxicants can be inferred from HSI [3]. Finally, the decreasing amount of chlorophyll—as a result of combustion—causes changes in the characteristics of the spectral signature of hyperspectral measurements [4]. It allows us to extract important insights concerning the scanned area, e.g., to assess whether it is under an active fire (because there are continuous changes over time), the area is burnt (because the chlorophyll content is low) or if there is a risk of the active area re-developing (partial burnout). It is worth mentioning that the process of monitoring and prevention of fires in Europe is carried out by the European Forest Fire Information System, which is part of the Earth Observation Programme. As part of it, monitoring of fires is carried out using multispectral data from Sentinel-2—the key information source is the red-edge band, which is one of the best descriptors of chlorophyll [4]. Furthermore, the chlorophyll-a index allows for the preparation of the pigmentation map [5], which can be the basis for detecting harmful algae blooms using the remotely sensed data [6,7]. Thus, determining the concentration of chlorophyll is important in monitoring water quality as well. Overall, although we focus on estimating the chlorophyll in leaves, non-invasive determination of its level is of paramount importance in an array of applications that could be potentially targeted at an enormous scale thanks to airborne and satellite imaging [8].

There are several ways to determine the level of chlorophyll in leaves, but most of them require direct and invasive access to leaves. Therefore, due to the high costs of such economically-infeasible, time-consuming and non-scalable procedures, the non-invasive techniques have become an important yet still under-developed research venue. The approaches which are focused on exploiting the multi- and hyperspectral images for this task span across those that use the in-field [9], airborne [10] and satellite [11,12] imaging, with the latter offering immediate scalability over large areas. Although there are works that reported promising results for the multispectral data [13], hyperspectral imaging is the current main focus, as it can allow for precise chlorophyll estimation thanks to the very detailed spectral information available in such data [14].

The state-of-the-art techniques for estimating the chlorophyll level exploit classic and deep supervised machine learning [15], with the latter benefiting from automated representation learning [16]. Such algorithms, however, require representative and large training sets capturing both image data and in situ measurements to generalize over new data. Unfortunately, albeit chlorophyll estimation is an important topic, there are no publicly available and established datasets that could provide an unbiased way of comparing the approaches for this task. Therefore, we are currently facing the reproducibility crisis [17]. Additionally, collecting high-quality ground truth is time- and cost-inefficient, hence such datasets are often synthesized [18]. Haboudane et al. focused on estimating the chlorophyll level from HSI and used a private set containing the images (72 VIS-NIR bands with 2 m GSD) and 12 reference measurements [10]. Similarly, the HSI band selection targeting the chlorophyll estimation was tackled over the in-house data in [19]. Although it is possible to obtain a rough chlorophyll level approximation using the spectral indices [12,20], its quality is questionable [21].

Acquiring precise in situ measurements is a pivotal step in building datasets that could be used to train and validate machine learning chlorophyll estimation algorithms. The majority of approaches that measure the actual level of chlorophyll utilize the soil-plant analysis development (SPAD) parameter [22]. There are, however, techniques exploiting the photosynthesis efficiency parameter and the chemical reflectance index [23]. Interestingly, some works presented the methodology to model the relationship between the chlorophyll content and the plant stress that could be investigated using the maximal photochemical efficiency of PSII (Fv/Fm) [24] or the fluctuations in the light intensity. Overall, the algorithms for non-invasive monitoring of chlorophyll parameters from multi/hyperspectral image data have been intensively researched due to their practical applicability and potential scalability (e.g., if deployed on board a satellite [8]) in precision agriculture, but there are no standardized procedures to validate them in an unbiased way. Furthermore, there are no publicly available and adopted datasets that could be used in such validation pipelines.

1.1. Contribution

In this paper, we address both research gaps (the lack of standardized procedures to validate the algorithms for chlorophyll estimation from HSI, and the lack of public and adopted datasets that could be utilized to investigate the approaches for this task) and introduce an end-to-end and reproducible validation procedure coupled with the real-life dataset of hyperspectral imagery and in situ measurements. We captured the in situ chlorophyll measurements together with the high-resolution HSI data and introduced a standardized approach for using this dataset to validate the emerging chlorophyll estimation techniques. This validation procedure will help us avoid any experimental flaws—we discussed such flaws concerned with the training/test dataset splits in our previous work in the context of the HSI segmentation [25].

Our contributions are therefore threefold:

- We introduce a publicly available set of (i) chlorophyll content measurements with some complementary information, including the soil moisture, weather parameters collected during the measurements or the relative water content, being the amount of water in a leaf at the time of sampling relative to the maximal water a leaf can hold and (ii) the corresponding high-resolution hyperspectral imagery (2.2 cm GSD). The dataset encompasses the orthophotomaps with the marked plots where the chlorophyll sampling has been completed, as well as the extracted images of separate plots. We performed the on-the-ground chlorophyll measurements, which resulted in four ground-truth parameters (Section 3.1):
  1. The SPAD index [22];
  2. The maximum quantum yield of the PSII photochemistry (Fv/Fm) [24];
  3. The performance index for energy conservation from photons absorbed by PSII to the final PSI electron acceptors (PI) [26];
  4. Relative water content (RWC) measurements for the sampled canopy, for capturing additional derivative information on the nutrition of the plants.
- We introduce a procedure for the unbiased validation of machine learning algorithms for estimating the chlorophyll-related parameters from HSI, and we ensure the full reproducibility of the experiments over our dataset (Section 3.2).
- We deliver the baseline results obtained for the introduced dataset (for four ground-truth parameters) using 15 machine learning techniques which can constitute the reference for any future studies emerging from our work (Section 4). Additionally, we show that the performance of a selected model can be further improved through regularization.

1.2. Structure of the Paper

In Section 2, we contextualize our work within state of the art by providing a review of an array of applications which can benefit from machine learning techniques operating on multi- and hyperspectral data, with a special emphasis put on chlorophyll estimation and on the way such approaches are validated. Section 3 introduces our chlorophyll estimation dataset, together with the validation procedure which is suggested to follow while exploiting it for experimentation. Our experimental results, obtained for various machine learning techniques over the suggested training-test dataset split, are gathered and discussed in Section 4. Finally, Section 5 concludes the paper.

2. Related Literature

The nature of activities in the agricultural sector has changed over the years as a result of a broadly-understood human activity, which encompasses—among other factors—rapidly growing population, environmental pollution, climate change and depletion of natural resources. The premise of precision agriculture is an effective food production process with a reduced impact on the environment. To achieve this goal, however, it is required to assess the soil quality, its irrigation, fertilizer content and seasonal changes

that occur in the ecosystem. Estimating the yield volume planned for a given region may also constitute the important information related to the effectiveness of the implemented agricultural practices [27]. Remote sensing may easily become a tool enabling the identification of soil and crop parameters due to the possibility of assessing a large area in subsequent time points. For agriculture, it is carried out using both passive and active methods. In the former case, multi- and hyperspectral remote sensing is used. The approaches using multispectral images (MSIs) are mainly based on the content of chlorophyll and its related parameters [28,29]. Nevertheless, the wide bandwidth that characterizes multispectral imaging results in limited accuracy in the early detection of negative symptoms such as nutrient deficiency or plant diseases [30]. The use of hyperspectral imaging, on the other hand, which is characterized by high spectral resolution (the bands are narrow and continuous), allows for the detection of more subtle details in the spectral response of a given area [31]. HSI-based methods can detect potential abnormalities, such as plant diseases, much faster than the MSI-based ones because the spectral signature contains more detailed characteristics derived from significantly narrower bands [32]. Additionally, satellites equipped with multispectral sensors (e.g., WorldView, QuickBird, Sentinel-2, Landsat) are still more popular than those with hyperspectral sensors (see, e.g., the EO-1 Hyperion mission and various emerging missions, including Intuition-1). Furthermore, there are some practical challenges that need to be faced for HSI missions, as such data may be extremely large, hence should be processed on board a satellite to downlink the "information" instead of raw image data. However, hyperspectral analysis in agriculture is popularly carried out by field-point methods using a spectroradiometer. The limitation to a few selected places makes spatial estimation impossible; therefore, research is conducted to determine the correlation between data collected by the field methods and data recorded by satellites [29,33,34] or by manned or unmanned airplanes [27,35–37]. In Figure 1, we show that the popularity of the topics (quantified as the number of papers published yearly) related to the MSI/HSI analysis in agriculture and including chlorophyll estimation, has been steadily growing over the last ten years. It also confirms the importance of introducing standardized validation procedures, which can be easily used to confront the emerging approaches for a given task in an unbiased and reproducible way.

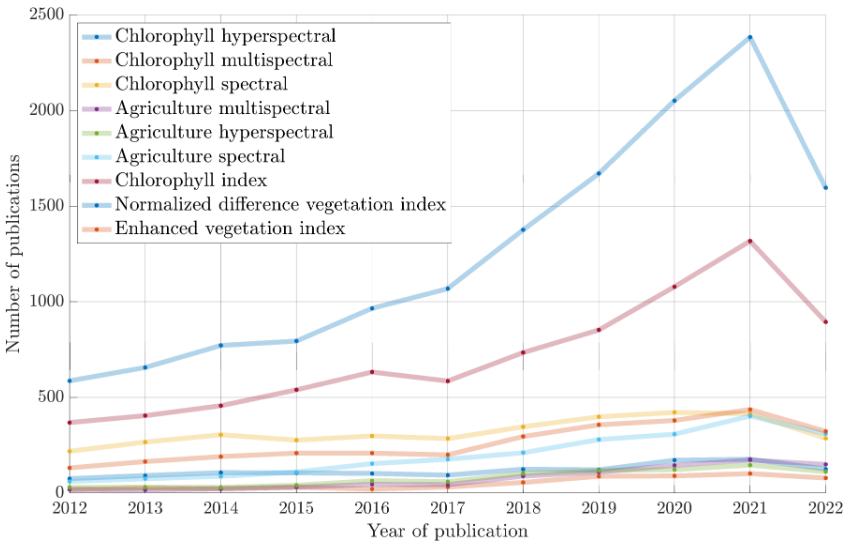


Figure 1. The popularity of topics related to the HSI analysis in various agricultural applications, including the chlorophyll estimation, quantified as the number of papers on such topics published between 2012 and 2022 (this analysis is based on https://app.dimensions.ai/discover/publication (accessed on 1 November 2022), the analysis was performed on 8 September 2022; in the legend, we present the keyphrase which was used). We can appreciate that the number of articles tackling the automated chlorophyll determination from hyperspectral imagery is increasing at a steady pace.

HSIs provide a tremendous amount of spectral-spatial information. On its basis, target agricultural parameters can be estimated—the narrow infrared and near-infrared bands can be used to accurately calculate the leaf area index (LAI) [38]. Similarly, the analysis of spectral data allowed us to define a vegetation index based on the assessment of chlorophyll content [35,39]. Assessing the content of chlorophyll in crops is one of the most important approaches, as chlorophyll is a reliable indicator of the crop health. The reason for the high usefulness of this biophysical pigment is that it enables us to evaluate the biochemical processes, which reflects the productivity of plants [31]. Vegetation coefficients are the basis for the estimation and monitoring of biomass, as well as for the assessment of soil composition and its moisture. There have been an array of machine learning techniques proposed for such tasks—the biomass estimation was performed using random forests [35,36,40,41], support vector machines [36,40,41] and multivariate regression modeling [28,36,41]. To tackle the problem of the high dimensionality of hyperspectral data, band selection and feature extraction, using, e.g., principal component analysis [42], are commonly deployed [36,43]. They elaborate a subset of the most discriminative bands or features [44]—the experiments using HSI obtained by unmanned aerial vehicles suggest that limiting the spectral range in the context of monitoring the plant growth to 454–950 nm [35] or 454–882 nm [27] is enough to achieve this goal.

In Table 1, we gather a set of the selected works focusing on MSI/HSI analysis in various agricultural applications (the papers tackling the chlorophyll estimation are presented in green), whereas Table 2 presents the corresponding experimental results reported in those publications. We can appreciate that both classic and deep machine learning models have been extensively developed throughout the years, but their direct comparison is virtually impossible, as the authors (i) utilize different datasets (of different cardinality and underlying characteristics, as they may have been captured using various sensors, in different acquisition conditions and following different acquisition procedures) and validation scenarios (e.g., different cross-validation approaches), and (ii) commonly report different metrics quantifying the capabilities of the investigated techniques. In this work, we tackle this issue and introduce a dataset capturing high-resolution hyperspectral imagery coupled with the in situ chlorophyll measurements, alongside the suggested validation procedure (its training-test split and a set of metrics, which should be used to quantify the prediction performance of the algorithms) and our baseline results obtained in this experimental scenario. We believe that it may be an important step toward unbiasing the way the community verifies the emerging approaches for estimating chlorophyll-related parameters from hyperspectral imagery, and it can help us effectively tackle the reproducibility crisis in the machine learning-based HSI analysis [17].

Table 1. Selected works focusing on the machine learning-powered analysis in agricultural applications, together with the additional feature extraction step performed in the corresponding method. The papers focusing on the chlorophyll estimation are in green.

Ref.	Year	Goal	Feature Extraction	Algorithm
[31]	2022	prediction of chlorophyll content in maize	selected bands; vegetation indices	random forest
[35]	2021	monitoring of above-ground biomass of maize	vegetation indices	stepwise regression; random forest; extreme gradient boosting
[45]	2021	monitoring of wheat yellow rust	vegetation indices; meteorological features	linear discriminant analysis; support vector machine; artificial neural network
[46]	2020	mapping of live fuel moisture content	NDVI; NDWI; NIRv	recurrent neural network
[28]	2020	estimation of biomass	vegetation indices	deep neural network

Table 1. *Cont.*

Ref.	Year	Goal	Feature Extraction	Algorithm
[27]	2020	monitoring of crop growth	high-frequency IWD information; continuous wavelet transform (CWT)	multiple linear stepwise regression
[40]	2020	dry biomass estimation of paddy rice	VH; VV	random forest; support vector machine; *k*-nearest neighbor; gradient boosting decision tree
[47]	2020	LAI detection	LAI	Gaussian process regression (GPR); variational heteroscedastic GPR
[33]	2020	estimation of soil organic carbon	principal component analysis; NDI; RI; DI	discrete wavelet transform at different scales; random forest; support vector machine; back-propagation neural network
[48]	2020	estimation of soil organic carbon	vegetation indices: NDVI, SAVI, NBSI, NDWI, NDBI, FI	random forest
[36]	2019	monitoring of above-ground biomass of maize	recursive-feature elimination	multiple linear regression; support vector machine; artificial neural network; random forest
[41]	2019	pasture conditions, seasonal dynamics of LAI and AGB	NDVI; EVI; LSWI	multiple linear regression; support vector machine; random forest
[38]	2018	canopy green leaf area	GBVI; NDVI; CI	empirical vegetation index regression (NDVIa-b and CIa-b); physically-based inversion, support vector regression
[49]	2016	maize biomass estimation, the seasonal variation	specific leaf area	simple algorithm for yield estimates
[43]	2015	detection of favorable wavelengths	singular value decomposition	stepwise regression

Table 2. The results reported in the selected works on machine learning-powered analysis in various agricultural applications. The papers focusing on the chlorophyll estimation are in green.

Ref.	Date Source	Type	Wavelength	Amount of Data	Measure	Value
[31]	DJI S1000 UAV; MicaSense Altum; Downwelling Light Sensor 2 (DLS-2)	MSI	465, 532, 630 nm, 680–730 nm, 1200–1600 nm	3576	R^2 RMSE RRMSE	results reported for the seasons

Table 2. *Cont.*

Ref.	Date Source	Type	Wavelength	Amount of Data	Measure	Value
[35]	DJI S1000 UAV; Cubert UHD 185	HSI	454–950 nm	1809	R^2 RMSE RRMSE	0.85 0.27 $\frac{t}{ha}$ 0.84%
[45]	Sentinel-2; National Meteorological Information Center	MSI	490, 560, 665, 842, 705 nm	58	accuracy	84.20%
[46]	Sentinel-1; Landsat-8	SAR MSI	C-band; 450–510 nm, 530–590 nm, 640–670 nm, 850–880 nm, 1570–1650 nm, 2110–2290 nm	not specified	R^2 RMSE bias	0.63 25.00% 1.90%
[28]	Sentinel-2	MSI	443–2190 nm	209	R^2 RMSE RRMSE	0.87 1.84 $\frac{t}{ha}$ 24.76%
[27]	DJI S1000 UAV; Cubert UHD 185; Sony DSC QX100	HSI	454–882 nm	144	R^2 RMSE MAE	0.85 0.79 $\frac{t}{ha}$ 1.01 $\frac{t}{ha}$
[40]	Sentinel-1A	SAR	C-band	175	R^2 RMSE	0.72 362.40 $\frac{g}{m^2}$
[47]	Sentinel-2	MSI	400–2400 nm	114	R^2 RMSE R^2 RMSE	0.78 (GPR) 0.70 (GPR) 0.80 (VHGPR) 0.63 (VHGPR)
[33]	Gaofen-5	HSI	433–1342 nm, 1460–1763 nm, 1990–2445 nm	14	R^2 RMSE	0.83 2.89 $\frac{g}{kg}$
[48]	ALOS PALSAR; Landsat TM	SAR MSI	L-band; 530–590 nm, 640–670 nm, 850–880 nm, 1570–1650 nm, 2110–2290 nm	not specified	R^2 RMSE RDP	0.59 9.27 $\frac{g}{kg}$ 1.98
[36]	DJI S1000 UAV; 1.2 megapixel Parrot Sequoia camera	MSI	550–790 nm	120 185	R^2 RMSE MAE	0.94 0.50 0.36
[41]	Sentinel-1A; Landsat-8; Sentinel-2	SAR MSI	C-band; 452–512 nm, 636–673 nm, 851–879 nm, 1566–1651 nm	not specified	R^2 RMSE	0.78 119.40 $\frac{g}{m^2}$
[38]	HyMap; CHRIS/PROBA	HSI MSI	677–707 nm	118	R^2	0.79
[49]	Formosat-2; SPOT4-Take5; Landsat-8; Deimos-1	MSI	specific to the sensor	195	R^2 RRMSE	0.96 4.6%
[43]	Landsat ETM+; KONOS; GeoEye-1; WorldView-2; Hyperion	HSI MSI	772, 539, 758, 914, 1130, 1320 nm	9, 23, 23, 24, 10	R^2 RMSE	0.12–0.97 1.15–2.47 $\frac{g}{m^2}$

3. Materials and Methods

In this section, we present our dataset containing high-resolution hyperspectral imagery coupled with the in situ chlorophyll measurements (Section 3.1). The validation procedure which should be followed while utilizing this dataset to investigate the capabilities of emerging chlorophyll estimation techniques is discussed in detail in Section 3.2.

3.1. Chlorophyll Estimation Dataset (CHESS)

In Figure 2, we visualize the process of building the **CH**lorophyll **ES**timation Data**S**et (CHESS). The data was collected in 2020 in the Plant Breeding and Acclimatization Institute —National Research Institute (IHAR-PIB) facility located in the Central Poland region (Jadwisin, Masovian Voivodeship). For the selected 24 outdoor plots of two different soil profiles (12 plots for each soil profile, without any repetitions or overlaps), two popular (in the central Europe) potato varieties were planted: *Lady Claire* and *Markies* (split evenly). The acquisition was carried out in June and July 2020 (3 rounds of acquisition, 4 weeks apart: 3 flights over 2 sets of 12 plots resulting in 72 HSIs acquired in total) when the leaves were fully developed. The images were captured using an unmanned aerial vehicle with the push-broom imaging spectrometer that registers 150 continuous spectral bands (460–902 nm with the 2.2 cm GSD). The orthorectification procedure was executed using the collected image material of each spectral band—it was possible thanks to using four location targets (see the far left image in Figure 2) whose geographical positions have been collected with a precise GPS device. The spectral correction of those maps using four calibration targets of different spectral characteristics (selected to cover the spectrum range in which plants are perceived best) was performed. This allowed us to finalize the image acquisition process with low location error (less than 1 cm), high image resolution (2.2 cm GSD), and consistent spectral characteristics.

In parallel to the image acquisition campaign, the in situ on-the-ground measurements were performed on each plot (sampling was executed at the same time). To provide precise measurements, we captured (i) the readout of the photosynthesis efficiency quantified as the chlorophyll content using the SPAD index using the Minolta SPAD-502 device, (ii) the measurement of the maximum quantum yield of the PSII photochemistry (Fv/Fm) using the Multifunctional Plant Efficiency Analyzer (Handy-PEA fluorimeter, Hansatech Instruments Ltd. and Pea Plus software), (iii) the performance of the electron flux to the final PSI electron acceptors, as discussed in [50], and (iv) RWC which reflects the lab-measured degree of hydration of the leaf's tissue [51–53].

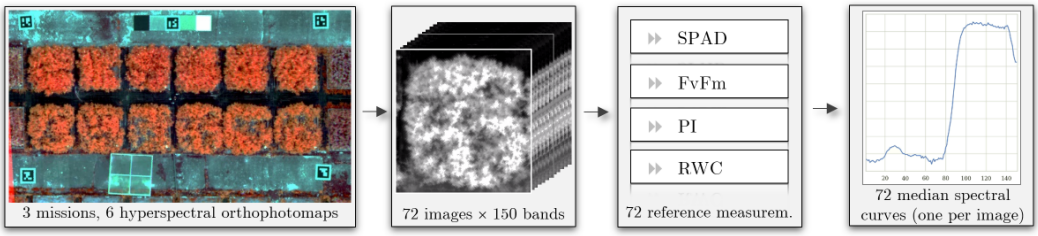


Figure 2. The dataset preparation procedure: 6 hyperspectral orthophotomaps, one for each flight for each set of 12 plots (**left**) were used to extract 72 hyperspectral 150-band images (**middle left**). The ground measurements of four parameters were performed for each plot (**middle right**). We extracted the spectral curves, individually for each pixel and aggregated across all pixels—see, e.g., the median spectral curve in the far **right** image.

The detailed agronomic setup and the dataset [54] with the training-test split constituting the validation procedure suggested in this paper, hence ensuring full reproducibility of the study, are available at https://data.mendeley.com/datasets/xn2wy75f8m (accessed on 1 November 2022). Since the measurements were collected in three independent rounds of data acquisition performed in the outdoor environment, CHESS reflects different plant characteristics and is intrinsically heterogeneous. We believe that exploiting a standardized

and non-biased validation procedure built upon such real-life data is of utmost importance to ensure full reproducibility and to avoid the "illusion of progress" in the field [17].

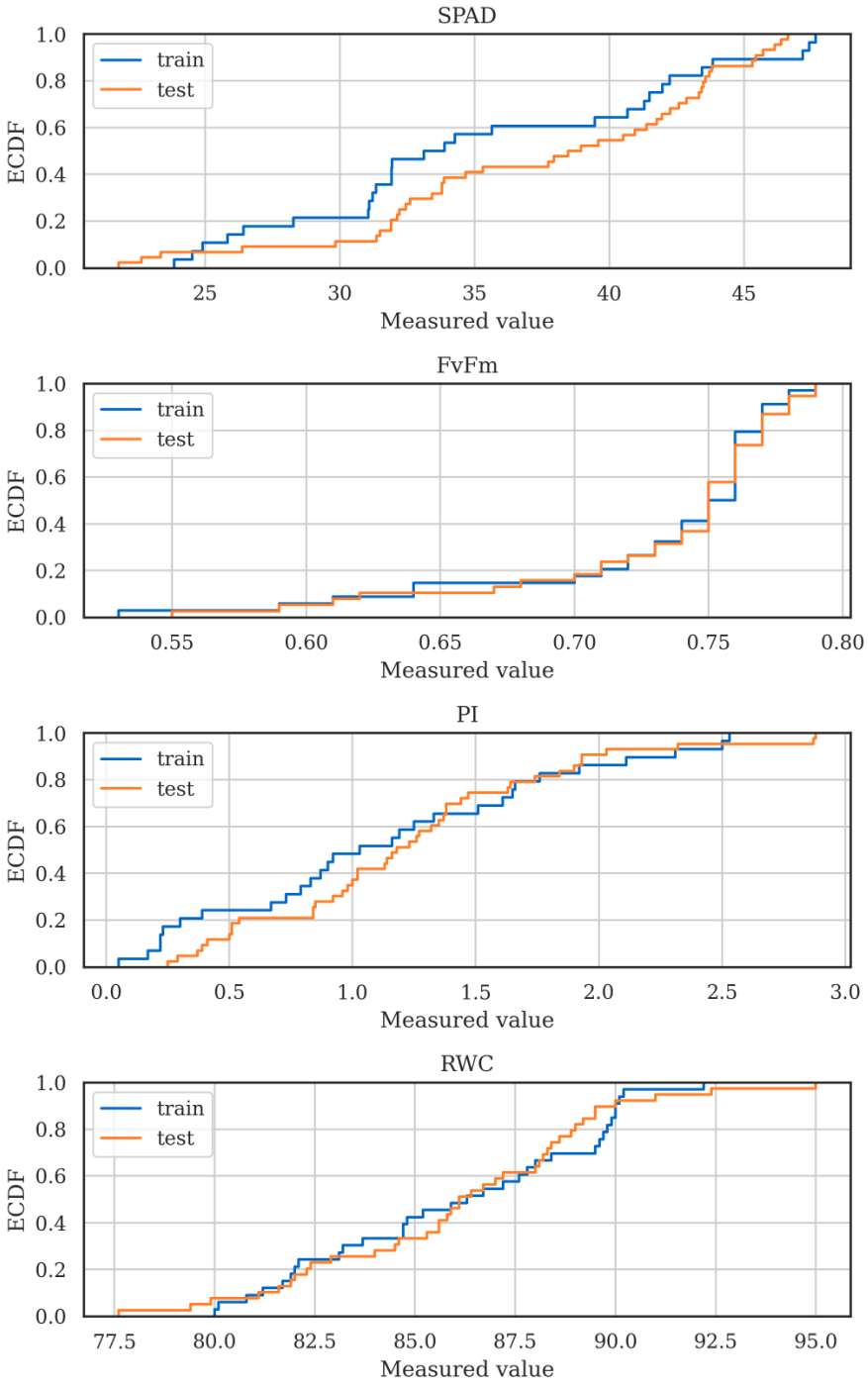


Figure 3. Empirical cumulative distribution (ECDF) of all parameters (SPAD, FvFm, PI and RWC) in the training and test sets.

3.2. Unbiased Validation of Chlorophyll Estimation

Unbiased and fair validation of the emerging algorithms for the non-invasive estimation of the chlorophyll-related parameters from HSI is critical to allow the community to track the progress in the field, and to accelerate the practical adoption of such approaches. Since there are differences across the measurement methodology followed for each in situ parameter (the SPAD index, Fv/Fm, PI and RWC), we provide four separate training-test dataset splits, independently for each ground-truth chlorophyll-related parameter. Each

split (i.e., for each parameter) is equinumerous, meaning that the number of plots is equal in the training and test subsets (36 HSIs with the ground-truth measurements captured for 36 separate plots of interest in both training and test sets). To be able to effectively quantify the generalization capabilities of the machine learning models trained and validated over such dataset splits, we stratified them according to the corresponding parameter's distribution to maintain similar distributions in both training and test sets (Figure 3). The reflectance characteristics of all HSIs across all folds are rendered in Figure 4. They show that there is a high agreement in the spectral features captured for the training and test images. Hence, the test set indeed resembles the characteristics of the training data and may be used to quantify the generalizability of the data-driven algorithms.

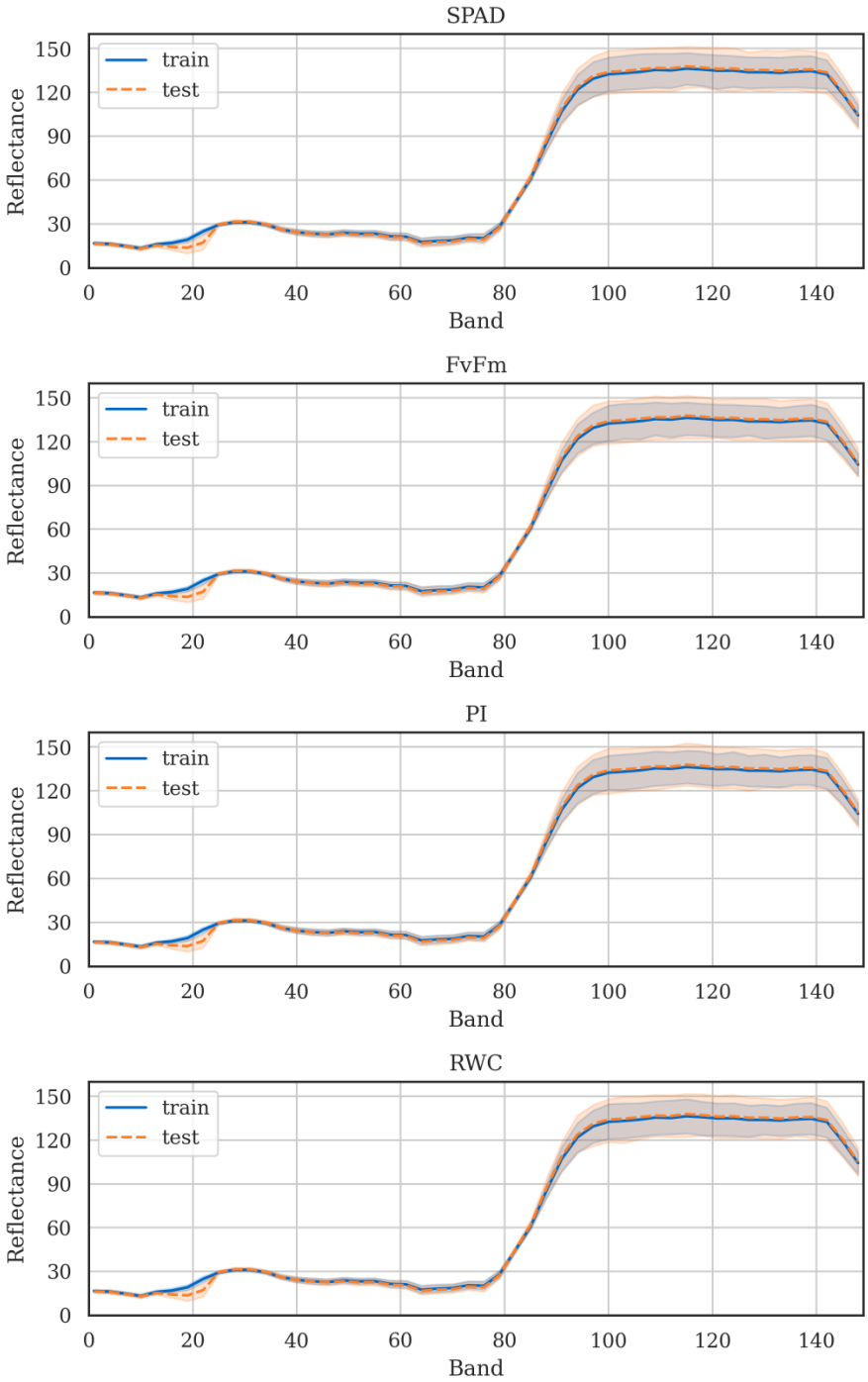


Figure 4. Spectral characteristics of all HSIs in the training and test sets. The mean spectral curves are rendered as blue and orange (dashed) lines for the training and test set, respectively.

Although the captured ground-truth parameters are commonly utilized in agronomy to assess the condition of the plants, they are measured differently, hence their characteristics are inherently varying. In Table 3, we gather the correlation coefficients across all of the parameters, indicating that, albeit some of them are indeed correlated (e.g., SPAD and PI, with the Pearson's and Spearman's coefficients amounting to 0.675 and 0.674, respectively), RWC is unrelated to other ones.

Table 3. Inter-parameter correlation between the measured ground-truth parameters (the Pearson's correlation coefficient values are reported over the main diagonal, whereas the Spearman's correlation coefficient values are given below the main diagonal; the corresponding *p*-values for the presented correlation coefficients are shown in brackets).

Parameter	SPAD	FvFm	PI	RWC
SPAD	1.000 (1.000)	0.668 (0.000)	0.675 (0.000)	0.109 (0.363)
FvFm	0.574 (0.000)	1.000 (1.000)	0.743 (0.000)	0.352 (0.002)
PI	0.674 (0.000)	0.877 (0.000)	1.000 (1.000)	−0.094 (0.434)
RWC	0.697 (0.550)	0.253 (0.161)	−0.047 (0.692)	1.000 (1.000)

4. Experimental Results

The objectives of our experiments are twofold: (i) to present the baseline results, obtained using a variety of machine learning algorithms (15 in total), over the introduced chlorophyll estimation dataset (CHESS) using the proposed training-test dataset splits (independently for SPAD, Fv/Fm, PI, and RWC), and to (ii) show that the predictive power of a selected model can be improved through additional regularization. To quantify the prediction performance of the algorithms (over the test sets), we exploit the classic metrics, including the coefficient of determination R^2 (upper bounded by the value of one indicating the perfect score, and all negative values of R^2 indicating a worse fit than the average fit [55]), the mean absolute percentage error (MAPE), the mean squared error (MSE) and the mean absolute error (MAE)—all errors should be minimized. For all models, we utilized their default parameters (Table 4), as suggested by Pedregosa et al. [56]—we intentionally have not executed any additional hyperparameter optimization to present the baseline solutions elaborated using the machine learning techniques with default parameterization. The algorithms are fed with the median spectral curves (hence, the feature vectors contain 150 values corresponding to the median value of each band within the image), and each model predicts a single chlorophyll-related parameter (SPAD, Fv/Fm, PI or RWC). Therefore, we do not perform any additional feature extraction or band selection (they may easily improve the performance of data-driven HSI analysis algorithms [44]).

Table 4. The most important hyperparameter values, as suggested by Pedregosa et al. [56], of all parameterized machine learning algorithms investigated in this study.

Algorithm	Hyperparameters
Ada Boost	**Maximum number of estimators**: 50, **learning rate**: 1, linear loss.
Decision Tree	**Loss function**: squared error, **maximum depth of the tree**: not set, **minimum number of samples required to split an internal node**: 2, **minimum number of samples required to be at a leaf node**: 1, allowing all features to be considered for the best split with unlimited number of leaf nodes.
Extra Trees	**Maximum number of estimators**: 10^2, **loss function**: squared error, **maximum depth of the tree**: not set, **minimum number of samples required to split an internal node**: 2, **minimum number of samples required to be at a leaf node**: 1, allowing all features to be considered for the best split with samples' bootstrapping while building trees, unlimited number of leaf nodes.
Extreme Gradient Boosting	**Learning rate**: 0.3, **maximum depth of an individual estimator**: 6, **minimum sum of instance weight**: 1, **maximum delta step**: 0, **regularization terms** λ: 1 and α: 0.

Table 4. *Cont.*

Algorithm	Hyperparameters
Gradient Boosting	**Learning rate**: 0.1, **maximum number of estimators**: 10^2, **maximum depth of an individual estimator**: 3, **loss function**: squared error, **percentage of samples required to split an internal node**: 10%.
Kernel Ridge	α: 1, **the degree of the linear polynomial kernel**: 3, **zero coefficient for polynomial and sigmoid kernels**: 1.
Lasso	α: 1, **maximum number of iterations**: 10^3, **tolerance stopping criteria**: 10^{-4}.
Light Gradient Boosting Machine	**Maximum tree leaves for base learners**: 31 without any limit for the tree depth for base learners, **boosting learning rate**: 10^{-1}, **number of boosted trees to fit**: 10^2, **number of samples for constructing bins**: $2 \cdot 10^4$, no minimum loss reduction required to make a further partition on a leaf node, **minimum sum of instance weight in a leaf**: 10^{-3}, **minimum samples in a child**: 20.
Linear Support Vector Machine	**Regularization parameter** C: 1, L1 loss, **maximum number of iterations**: 10^3, **the tolerance stopping criteria**: 10^{-4}.
Nu Support Vector Machine	**Kernel**: Radial Basis Function, **upper bound on the fraction of training errors and a lower bound of the fraction of support vectors**: 0.5, **regularization parameter** C: 1, γ: $1 \,/\, \left(\lvert\mathcal{F}\rvert \cdot \sigma^2(\boldsymbol{T})\right)$, where $\mathcal{F}$ is the number of features, and $\sigma^2(\boldsymbol{T})$ is the variance of the training set.
Random Forest	**Maximum number of estimators**: 10^2, **function measuring the quality of split**: squared error, **minimum number of samples required to split an internal node**: 2, **minimum number of samples required to be at a leaf**: 1, allowing all features to be considered for the best split with unlimited number of leaf nodes, no maximum number of samples for bootstrapping.
Ridge	α: 1, **tolerance stopping criteria**: 10^{-3}.
Stochastic Gradient Descent	**Loss function**: squared error with L2 regularization, $\alpha = 10^{-4}$, **L1 ratio**: 0.15, **maximum number of passes over the training data**: 10^3, **the stopping criterion for loss**: 10^{-3}, data shuffling after each epoch, **the initial learning rate**: 10^{-2}, **the exponent for inverse scaling learning rate**: 0.25, 10% of training data is the validation set for early stopping with 5 iterations with no improvement termination.
Support Vector Machine	**Kernel:** Radial Basis Function, γ: $1 \,/\, \left(\lvert\mathcal{F}\rvert \cdot \sigma^2(\boldsymbol{T})\right)$, where $\mathcal{F}$ denotes the number of features, and $\sigma^2(\boldsymbol{T})$ is the variance of the training set, **tolerance for stopping criterion**: 10^{-3}, **regularization parameter** C: 1, ϵ: 10^{-1}, where ϵ specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance ϵ from the actual value.

In Table 5, we gather the results obtained for all investigated parameters (SPAD, FvFm, PI and RWC) using top-3 machine learning models (with default parameterization), according to the R^2 metric, being the most widely utilized quality measure in precision agriculture. We can appreciate that Linear Regression allows for obtaining the best coefficients of determination for SPAD, FvFm and PI, which are significantly larger than those achieved by the second-best algorithm (R^2 smaller by 0.099, 0.110, and 0.288 than for Extreme Gradient Boosting, Gradient Boosting and Extra Trees for SPAD, FvFm and PI, respectively). For RWC, this linear model resulted in R^2 of 0.720 (it was ranked fourth), and it was outperformed by the non-linear regression techniques. The results indicate that building heterogeneous regression ensembles, capturing both linear and non-linear models [57], may further improve the overall quality of estimating the parameters of interest.

To show that a selected model can be further improved, we employed a classic L2 regularization to the Ridge regression model, as it outperformed the other techniques for RWC (note that it failed to deliver high-quality prediction for other parameters, as presented later). A similar application of a regularized model was utilized to estimate the chlorophyll concentration by Lin and Lin [58], and it was shown to be effective in enhancing the algorithm's generalization capabilities. In Figure 5, we present the R^2 values over the

test sets (for each parameter) obtained for a range of the α hyperparameter values which controls the regularization strength (the larger α becomes, the stronger regularization is). We can observe that for a fine-tuned α parameter, we can significantly improve the model's performance for all chlorophyll parameters. Ridge regression with L2 regularization (with α of 2.5×10^{-5}, 5×10^{-5}, 10^{-11} and 10^{-3} for SPAD, FvFm, PI and RWC, respectively) not only did provide statistically-significant improvements over the baseline Ridge model for all parameters, but also outperformed the other investigated models in 3/4 chlorophyll parameters (Table 6). Here, only for PI, the results were the same as those obtained by Linear Regression (R^2 of 0.667). Thus, further improvements of the machine learning models, including the optimization of their hyperparameters or selection of appropriate training and/or feature sets, can easily lead to better regressors which may be confronted with the other techniques using our validation procedure in an unbiased and fair way.

Table 5. Three best machine learning models (according to R^2) with default parameterization for each ground-truth parameter. We indicate if the metric should be minimized (↓) or maximized (↑).

Param.	Model	R^2 ↑	MAPE ↓	MSE ↓	MAE ↓
SPAD	Linear Regression	0.818	0.072	9.583	1.569
	Extreme Gradient Boosting	0.719	0.092	14.808	2.784
	AdaBoost	0.698	0.080	15.935	1.814
FvFm	Linear Regression	0.718	0.037	0.001	0.021
	Gradient Boosting	0.608	0.037	0.001	0.017
	AdaBoost	0.600	0.037	0.001	0.016
PI	Linear Regression	0.667	0.532	0.169	0.280
	Extra Trees	0.379	1.251	0.315	0.368
	Random Forest	0.213	1.594	0.400	0.470
RWC	Ridge	0.817	0.014	2.249	1.207
	Extreme Gradient Boosting	0.793	0.014	2.541	1.021
	Support Vector Machine	0.745	0.017	3.127	1.541

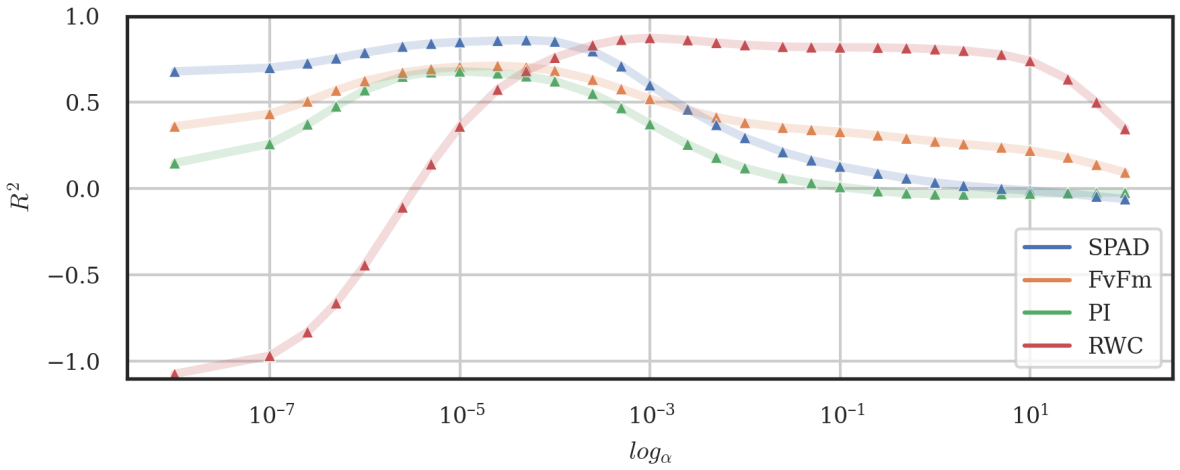


Figure 5. The Ridge regression results (R^2 over the test sets for each parameter) obtained using a range of the α hyperparameter values.

Table 6. The Ridge regression model with the L2 regularization.

Param.	α	$R^2 \uparrow$	MAPE ↓	MSE ↓	MAE ↓
SPAD	2.5×10^{-5}	0.827	0.072	9.095	1.625
FvFm	5×10^{-5}	0.727	0.036	0.001	0.021
PI	10^{-11}	0.667	0.532	0.169	0.280
RWC	10^{-3}	0.859	0.013	1.731	0.941

Finally, in Table 7, we present the experimental results obtained for all investigated parameters using all regression models. Here, the best results obtained using the models with default parameterization are **boldfaced in black**, whereas the Ridge regression model with additional L2 regularization is rendered in green (we discuss the process of improving a selected baseline model to enhance its capabilities later in this section). Finally, if the Ridge regression model with regularization led to obtaining the globally-best metric value (when compared with other techniques), we **boldfaced and underlined** the corresponding entry. The results indeed show that it is possible to enhance the generalization capabilities of the "default" machine learning models.

Table 7. Performance of the investigated machine learning models for each chlorophyll-related parameter of interest. The best results obtained using the baseline machine learning models are **boldfaced in black**, whereas the model with further regularization is indicated in green (if it resulted in the best metric value across all models, the values are boldfaced and underlined). We indicate if the metric should be minimized (↓) or maximized (↑).

Parameter	Model	$R^2 \uparrow$	MAPE ↓	MSE ↓	MAE ↓
SPAD	AdaBoost	0.698	0.080	15.935	1.814
	Decision Tree	0.178	0.132	43.324	2.640
	ExtraTrees	0.649	0.099	18.494	2.307
	Extreme Gradient Boosting	0.719	0.092	14.808	2.784
	Gradient Boosting	0.639	0.095	19.025	2.059
	Kernel Ridge	0.132	0.179	45.768	4.787
	Lasso	−0.091	0.213	57.509	6.045
	Light Gradient Boosting Machine	0.180	0.178	43.255	5.016
	Linear Regression	**0.818**	**0.072**	**9.583**	**1.569**
	Linear Support Vector Machine	0.216	0.131	41.321	3.242
	Nu Support Vector Machine	−0.142	0.219	60.211	6.065
	Random Forest	0.587	0.123	21.760	3.029
	Ridge	0.037	0.202	50.789	6.153
	Ridge with L2	**0.827**	**0.072**	**9.095**	1.625
	Stochastic Gradient Descent	0.119	0.187	46.465	5.794
	Support Vector Machine	−0.289	0.231	67.984	6.831
FvFm	AdaBoost	0.600	**0.037**	**0.001**	0.016
	Decision Tree	−0.005	0.055	0.003	0.020
	ExtraTrees	0.136	0.049	0.003	**0.016**
	Extreme Gradient Boosting	0.407	0.044	0.002	0.019
	Gradient Boosting	0.608	0.037	**0.001**	0.017
	Kernel Ridge	−1.609	0.100	0.009	0.054
	Lasso	−0.002	0.061	0.003	0.028
	Light Gradient Boosting Machine	−0.002	0.061	0.003	0.028
	Linear Regression	**0.718**	0.037	**0.001**	0.021
	Linear Support Vector Machine	0.363	0.048	0.002	0.022
	Nu Support Vector Machine	0.316	0.047	0.002	0.022
	Random Forest	0.510	0.043	0.002	0.019
	Ridge	0.272	0.051	0.003	0.025
	Ridge with L2	**0.727**	**0.036**	**0.001**	0.021
	Stochastic Gradient Descent	−4.424	0.155	0.019	0.102
	Support Vector Machine	−0.112	0.070	0.004	0.034

Table 7. *Cont.*

Parameter	Model	$R^2 \uparrow$	MAPE ↓	MSE ↓	MAE ↓
PI	AdaBoost	−0.067	1.825	0.542	**0.262**
	Decision Tree	−0.150	0.955	0.584	0.350
	Extra Tree	0.379	1.251	0.315	0.368
	Extreme Gradient Boosting	0.168	1.407	0.423	0.388
	Gradient Boosting	0.114	1.508	0.450	0.296
	Kernel Ridge	−0.147	1.834	0.583	0.588
	Lasso	−0.020	1.941	0.518	0.502
	Light Gradient Boosting Machine	−0.010	1.623	0.513	0.456
	Linear Regression	**0.667**	**0.532**	**0.169**	0.280
	Linear Support Vector Machine	−0.133	1.334	0.576	0.465
	Nu Support Vector Machine	−0.010	1.529	0.513	0.634
	Random Forest	0.213	1.594	0.400	0.470
	Ridge	−0.036	1.993	0.526	0.442
	Ridge with L2	**0.667**	**0.532**	**0.169**	0.280
	Stochastic Gradient Descent	−0.108	1.550	0.563	0.601
	Support Vector Machine	0.019	1.818	0.499	0.506
RWC	AdaBoost	0.657	0.018	4.213	1.233
	Decision Tree	0.432	0.025	6.977	1.900
	Extra Tree	0.704	0.018	3.641	1.090
	Extreme Gradient Boosting	0.793	**0.014**	2.541	**1.021**
	Gradient Boosting	0.646	0.020	4.350	1.250
	Kernel Ridge	−4.435	0.074	66.776	4.045
	Lasso	0.000	0.036	12.288	3.549
	Light Gradient Boosting Machine	0.000	0.036	12.288	3.549
	Linear Regression	0.720	0.018	3.440	1.385
	Linear Support Vector Machine	−14.497	0.125	190.396	7.235
	Nu Support Vector Machine	0.695	0.019	3.745	1.610
	Random Forest	0.709	0.018	3.581	1.499
	Ridge	**0.817**	**0.014**	**2.249**	1.207
	Ridge with L2	**0.859**	**0.013**	**1.731**	**0.941**
	Stochastic Gradient Descent	−1.491	0.051	30.598	3.334
	Support Vector Machine	0.745	0.017	3.127	1.541

5. Conclusions and Future Work

Capturing the information concerning the chlorophyll level in leaves is an important practical issue in precision agriculture, as it helps practitioners optimize their operation and appropriately plan and monitor the treatment process of various plants. Although there exist in-field methods that allow us to recognize the actual chlorophyll level through elaborating an array of indicators, they are invasive, time-inefficient and lack scalability. Therefore, developing non-invasive approaches benefiting from the detailed information available in hyperspectral imaging has attracted research attention. However, the emerging data-driven algorithms for this task are commonly evaluated using private datasets without any standardized validation procedure, which makes their comparison with the current state of the art virtually impossible. In this paper, we tackled this issue and proposed an open dataset (coupling HSI with in situ ground-truth measurements), together with its training-test splits and quality metrics that can be used to confront the emerging and existing techniques in a fully-unbiased and fair way. Our experimental study not only provided a solid baseline obtained using 15 classic machine learning predictors, but we also showed that it is possible to enhance such models to improve their generalizability. We believe that our work may constitute an important step toward standardizing the way we compare the chlorophyll-analysis algorithms and may help consolidate state of the art in the field by providing a clear way of comparing new approaches over real data.

Our work is an interesting point of departure for further research. In this work, we did not intend to introduce a new, "ground-breaking" algorithm for estimating the chlorophyll level from HSI. There are, however, immediate next steps which should be

performed to improve the performance capabilities of machine learning models, with the hyperparameter optimization being one of them. In Figure 6, we show how two most important hyperparameters of Support Vector Machines (C and γ) affect their performance. Observing the results of the models optimized for each target parameter separately and gathered in Table 8, we can appreciate that the grid-searched Support Vector Machines significantly outperformed their default parameterization (Table 7). Therefore, optimizing the most important hyperparameters of other techniques would likely lead to their noticeable improvements.

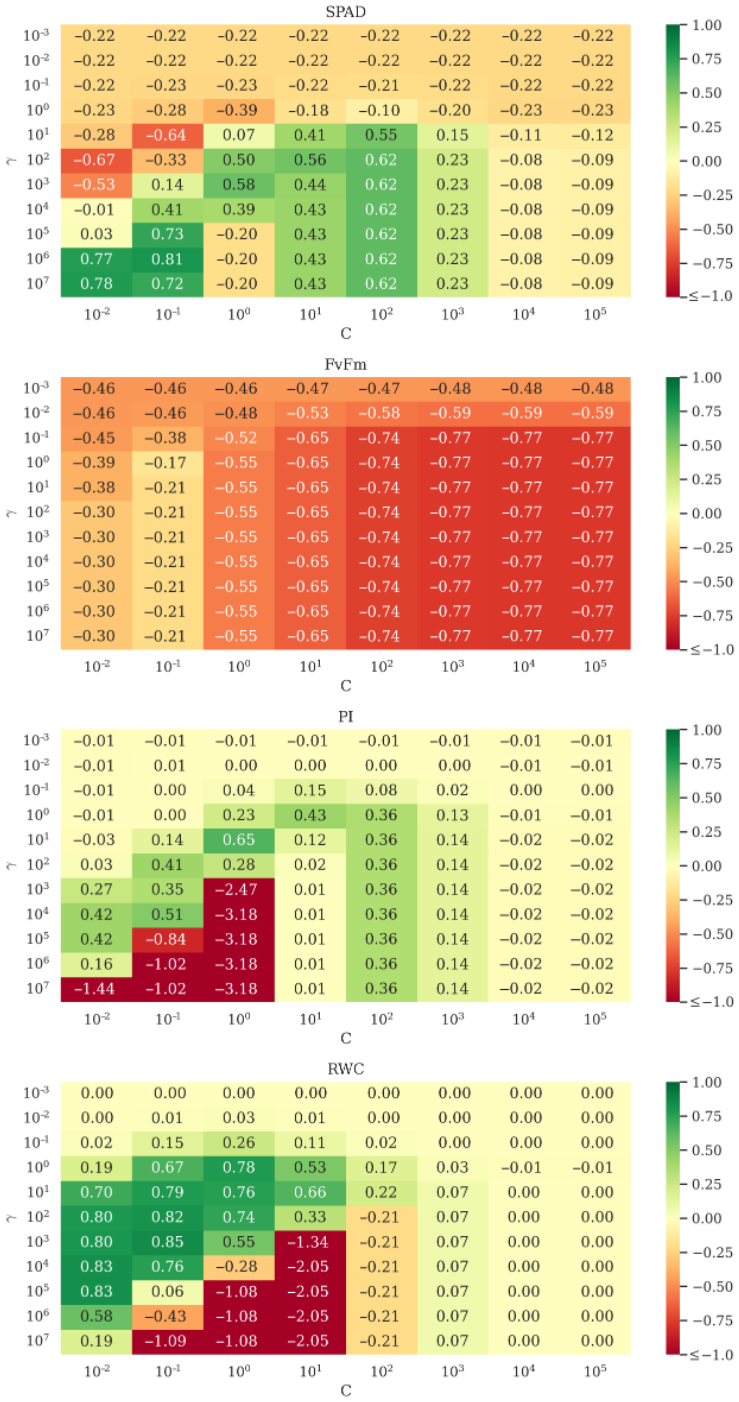


Figure 6. The results elaborated using a Support Vector Machine over the test sets for each parameter, and obtained using a range of the C and γ hyperparameter values.

Table 8. The results elaborated using a Support Vector Machine with the hyperparameters optimized for each parameter separately. We indicate if the metric should be minimized (↓) or maximized (↑).

Param.	γ	C	R^2 ↑	MAPE ↓	MSE ↓	MAE ↓
SPAD	10^6	10^{-1}	0.808	0.073	10.136	2.401
FvFm	10^0	10^{-1}	−0.169	0.076	0.004	0.052
PI	10^1	10^0	0.652	1.157	0.177	0.372
RWC	10^3	10^{-1}	0.845	0.016	1.904	1.006

We are observing an unprecedented success of deep learning in HSI analysis—such techniques may certainly improve the quality of the estimated chlorophyll-related parameters [15]. To unchain the full scalability potential of HSI, we can perform the analysis process on board a satellite to extract knowledge from raw pixels. However, the algorithms to be deployed in such hardware-constrained execution environments should be resource-frugal and robust against various noise affecting the in-orbit image acquisition [8] and must be thoroughly validated before they can run in space [59]. Developing the (deep) machine learning models for on-board processing is currently widely-explored due to a significant number of emerging Earth observation missions, including our Intuition-1 satellite.

Author Contributions: Conceptualization, B.R. and J.N.; Methodology, B.R., J.N. and A.M.W.; Software, B.R.; Validation, B.R. and J.N.; Formal Analysis, B.R.; Literature Review, A.M.W., B.R. and J.N.; Investigation, B.R.; Resources, B.R.; Data Curation, B.R.; Writing—Original Draft Preparation, B.R., J.N. and A.M.W.; Writing—Review and Editing, J.N., B.R. and A.M.W.; Visualization, B.R. and A.M.W.; Supervision, B.R. and J.N.; Funding Acquisition, J.N. All authors have read and agreed to the submitted version of the manuscript.

Funding: A.M.W. and J.N. were supported by the Silesian University of Technology grants for maintaining and developing research potential (A.M.W.: 07/010/BKM22/1017). This work was partially supported by The National Centre for Research and Development of Poland under project POIR.04.01.04-00-0009/19.

Data Availability Statement: Publicly available dataset was analyzed in this study. This data can be found here: https://doi.org/10.1016/j.dib.2022.108087 and here: https://data.mendeley.com/datasets/xn2wy75f8m/ (accessed on 1 November 2022).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AGB	Above-ground biomass
ALOS	Advanced Land Observing Satellite
CHESS	**CH**lorophyll **ES**timation Data**S**et
CHRIS	Compact High Resolution Imaging Spectrometer
CWT	Continuous wavelet transform
DNN	Deep neural network
EFFIS	The European Forest Fire Information System
EVI	Enhanced vegetation index
FvFm	The maximum quantum yield of the PSII photochemistry
GBDT	Gradient boosting decision tree
GBVI	Green brown vegetation index
GSD	Ground sample distance
GPR	Gaussian process regression
HSI	Hyperspectral image
k-NN	*k*-nearest neighbor

LAI	Leaf area index
LSWI	Land surface water index
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MLR	Multiple linear regression
MSE	Mean squared error
MSI	Multispectral image
NBSI	Non-binary snow index
NDVI	Normalized difference vegetation index
NIR	Near-infrared spectra
NIRv	Near-infrared reflectance of vegetation
PALSAR	Phased Array L-band Synthetic Aperture Radar
PCA	Principal component analysis
PI	The index for energy conservation from photons absorbed by PSII to PSI electron acceptors
R^2	Coefficient of determination
RDP	Ratio of the performance to deviation
RF	Random forest
RFE	Recursive-feature-elimination
RMSE	Root mean squared error
RRMSE	Relative root mean squared error
RWC	Relative water content measurements for the sampled canopy
SAVI	Soil-adjusted vegetation index
SAR	Synthetic-aperture radar
SOC	Soil organic carbon
SPAD	The actual level of chlorophyll utilizing the soil-plant analysis development
SVD	Singular value decomposition
SVM	Support vector machine
VHGPR	Variational heteroscedastic GPR
VIS-NIR	Visible–near-infrared spectra
VH	Vertical transmitted and horizontal received
VV	Vertical transmitted and vertical received polarization
XGB	Extreme gradient boosting

References

1. Shen, Q.; Lin, J.; Yang, J.; Zhao, W.; Wu, J. Exploring the Potential of Spatially Downscaled Solar-Induced Chlorophyll Fluorescence to Monitor Drought Effects on Gross Primary Production in Winter Wheat. *IEEE J-STARS* **2022**, *15*, 2012–2022. [CrossRef]
2. Long, Y.; Ma, M. Recognition of Drought Stress State of Tomato Seedling Based on Chlorophyll Fluorescence Imaging. *IEEE Access* **2022**, *10*, 48633–48642. [CrossRef]
3. Oláh, V.; Hepp, A.; Irfan, M.; Mészáros, I. Chlorophyll Fluorescence Imaging-Based Duckweed Phenotyping to Assess Acute Phytotoxic Effects. *Plants* **2021**, *10*, 2763. [CrossRef] [PubMed]
4. Lazzeri, G.; Frodella, W.; Rossi, G.; Moretti, S. Multitemporal Mapping of Post-Fire Land Cover Using Multiplatform PRISMA Hyperspectral and Sentinel-UAV Multispectral Data: Insights from Case Studies in Portugal and Italy. *Sensors* **2021**, *21*. [CrossRef] [PubMed]
5. Pyo, J.; Duan, H.; Baek, S.; Kim, M.S.; Jeon, T.; Kwon, Y.S.; Lee, H.; Cho, K.H. A Convolutional Neural Network Regression for Quantifying Cyanobacteria Using Hyperspectral Imagery. *Remote Sens. Environ.* **2019**, *233*, 111350. [CrossRef]
6. Hill, P.R.; Kumar, A.; Temimi, M.; Bull, D.R. HABNet: Machine Learning, Remote Sensing-Based Detection of Harmful Algal Blooms. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3229–3239. [CrossRef]
7. Torres Palenzuela, J.M.; Vilas, L.G.; Bellas Aláez, F.M.; Pazos, Y. Potential Application of the New Sentinel Satellites for Monitoring of Harmful Algal Blooms in the Galician Aquaculture. *Thalass. Int. J. Mar. Sci.* **2020**, *36*, 85–93. [CrossRef]
8. Nalepa, J.; Myller, M.; Cwiek, M.; Zak, L.; Lakota, T.; Tulczyjew, L.; Kawulok, M. Towards On-Board Hyperspectral Satellite Image Segmentation: Understanding Robustness of Deep Learning through Simulating Acquisition Conditions. *Remote Sens.* **2021**, *13*, 1532. [CrossRef]
9. Liu, N.; Liu, G.; Sun, H. Real-Time Detection on SPAD Value of Potato Plant Using an In-Field Spectral Imaging Sensor System. *Sensors* **2020**, *20*, 3430. [CrossRef]
10. Haboudane, D.; Miller, J.R.; Tremblay, N.; Zarco-Tejada, P.J.; Dextraze, L. Integrated Narrow-Band Vegetation Indices for Prediction of Crop Chlorophyll Content for Application to Precision Agriculture. *Remote Sens. Environ.* **2002**, *81*, 416–426. [CrossRef]

11. Hai-ling, J.; Li-fu, Z.; Hang, Y.; Xiao-ping, C.; Shu-dong, W.; Xue-ke, L.; Kai, L. Comparison of Accuracy and Stability of Estimating Winter Wheat Chlorophyll Content Based on Spectral Indices. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 2985–2988. [CrossRef]
12. Raya-Sereno, M.D.; Alonso-Ayuso, M.; Pancorbo, J.L.; Gabriel, J.L.; Camino, C.; Zarco-Tejada, P.J.; Quemada, M. Residual Effect and N Fertilizer Rate Detection by High-Resolution VNIR-SWIR Hyperspectral Imagery and Solar-Induced Chlorophyll Fluorescence in Wheat. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
13. Wang, J.; Zhou, Q.; Shang, J.; Liu, C.; Zhuang, T.; Ding, J.; Xian, Y.; Zhao, L.; Wang, W.; Zhou, G.; et al. UAV- and Machine Learning-Based Retrieval of Wheat SPAD Values at the Overwintering Stage for Variety Screening. *Remote Sens.* **2021**, *13*, 5166. [CrossRef]
14. Yuan, Z.; Ye, Y.; Wei, L.; Yang, X.; Huang, C. Study on the Optimization of Hyperspectral Characteristic Bands Combined with Monitoring and Visualization of Pepper Leaf SPAD Value. *Sensors* **2021**, *22*, 183. [CrossRef] [PubMed]
15. Ye, H.; Tang, S.; Yang, C. Deep Learning for Chlorophyll-a Concentration Retrieval: A Case Study for the Pearl River Estuary. *Remote Sens.* **2021**, *13*, 3717. [CrossRef]
16. Tulczyjew, L.; Kawulok, M.; Longépé, N.; Le Saux, B.; Nalepa, J. A Multibranch Convolutional Neural Network for Hyperspectral Unmixing. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
17. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv* **2022**, arXiv:2207.07048. [CrossRef].
18. Inoue, Y.; Guérif, M.; Baret, F.; Skidmore, A.; Gitelson, A.; Schlerf, M.; Darvishzadeh, R.; Olioso, A. Simple and Robust Methods for Remote Sensing of Canopy Chlorophyll Content: A Comparative Analysis of Hyperspectral Data for Different Types of Vegetation. *Plant Cell Environ.* **2016**, *39*, 2609–2623. [CrossRef]
19. Mayranti, F.P.; Saputro, A.H.; Handayani, W. Chlorophyll A and B Content Measurement System of Velvet Apple Leaf in Hyperspectral Imaging. In Proceedings of the ICICOS, Semarang, Indonesia, 29–30 October 2019; pp. 1–5. [CrossRef]
20. Tomaszewski, M.; Gasz, R.; Smykała, K. Monitoring Vegetation Changes Using Satellite Imaging—NDVI and RVI4S1 Indicators. In Proceedings of the Control, Computer Engineering and Neuroscience, Opole, Poland, 21 September 2021; Paszkiel, S., Ed.; Springer: Berlin/Heidelberg, Germany, 2021; pp. 268–278. [CrossRef]
21. Bannari, A.; Khurshid, K.S.; Staenz, K.; Schwarz, J.W. A Comparison of Hyperspectral Chlorophyll Indices for Wheat Crop Chlorophyll Content Estimation Using Laboratory Reflectance Measurements. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3063–3074. [CrossRef]
22. El-Hendawy, S.; Dewir, Y.H.; Elsayed, S.; Schmidhalter, U.; Al-Gaadi, K.; Tola, E.; Refay, Y.; Tahir, M.U.; Hassan, W.M. Combining Hyperspectral Reflectance Indices and Multivariate Analysis to Estimate Different Units of Chlorophyll Content of Spring Wheat under Salinity Conditions. *Plants* **2022**, *11*, 456. [CrossRef] [PubMed]
23. Middleton, E.M.; Julitta, T.; Campbell, P.E.; Huemmrich, K.F.; Schickling, A.; Rossini, M.; Cogliati, S.; Landis, D.R.; Alonso, L. Novel Leaf-Level Measurements of Chlorophyll Fluorescence for Photosynthetic Efficiency. In Proceedings of the IGARSS, Milan, Italy, 26–31 July 2015; pp. 3878–3881. [CrossRef]
24. Jia, M.; Zhou, C.; Cheng, T.; Tian, Y.; Zhu, Y.; Cao, W.; Yao, X. Inversion of Chlorophyll Fluorescence Parameters on Vegetation Indices at Leaf Scale. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 4359–4362. [CrossRef]
25. Nalepa, J.; Myller, M.; Kawulok, M. Validating Hyperspectral Image Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1264–1268. [CrossRef]
26. Singh, H.; Kumar, D.; Soni, V. Performance of Chlorophyll a Fluorescence Parameters in Lemna Minor under Heavy Metal Stress Induced by Various Concentration of Copper. *Sci. Rep.* **2022**, *12*, 10620. [CrossRef] [PubMed]
27. Yue, J.; Zhou, C.; Guo, W.; Feng, H.; Xu, K. Estimation of Winter-Wheat Above-Ground Biomass Using the Wavelet Analysis of Unmanned Aerial Vehicle-Based Digital Images and Hyperspectral Crop Canopy iImages. *Int. J. Remote Sens.* **2021**, *42*, 1602–1622. [CrossRef]
28. Jin, X.; Li, Z.; Feng, H.; Ren, Z.; Li, S. Deep Neural Network Algorithm for Estimating Maize Biomass Based on Simulated Sentinel 2A Vegetation Indices and Leaf Area Index. *Crop J.* **2020**, *8*, 87–97. [CrossRef]
29. Lu, B.; He, Y. Evaluating Empirical Regression, Machine Learning, and Radiative Transfer Modelling for Estimating Vegetation Chlorophyll Content Using Bi-Seasonal Hyperspectral Images. *Remote Sens.* **2019**, *11*, 1979. [CrossRef]
30. Adão, T.; Hruška, J.; Pádua, L.; Bessa, J.; Peres, E.; Morais, R.; Sousa, J.J. Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sens.* **2017**, *9*, 1110. [CrossRef]
31. Brewer, K.; Clulow, A.; Sibanda, M.; Gokool, S.; Naiken, V.; Mabhaudhi, T. Predicting the Chlorophyll Content of Maize over Phenotyping as a Proxy for Crop Health in Smallholder Farming Systems. *Remote Sens.* **2022**, *14*, 518. [CrossRef]
32. Lu, B.; Dao, P.D.; Liu, J.; He, Y.; Shang, J. Recent Advances of Hyperspectral Imaging Technology and Applications in Agriculture. *Remote Sens.* **2020**, *12*, 2659. [CrossRef]
33. Meng, X.; Bao, Y.; Liu, J.; Liu, H.; Zhang, X.; Zhang, Y.; Wang, P.; Tang, H.; Kong, F. Regional Soil Organic Carbon Prediction Model Based on a Discrete Wavelet Analysis of Hyperspectral Satellite Data. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *89*, 102111. [CrossRef]
34. Hong, Y.; Chen, S.; Chen, Y.; Linderman, M.; Mouazen, A.M.; Liu, Y.; Guo, L.; Yu, L.; Liu, Y.; Cheng, H.; et al. Comparing Laboratory and Airborne Hyperspectral Data for the Estimation and Mapping of Topsoil Organic Carbon: Feature Selection Coupled with Random Forest. *Soil Tillage Res.* **2020**, *199*, 104589. [CrossRef]

35. Zhang, Y.; Xia, C.; Zhang, X.; Cheng, X.; Feng, G.; Wang, Y.; Gao, Q. Estimating the Maize Biomass by Crop Height and Narrowband Vegetation Indices Derived from UAV-based Hyperspectral Images. *Ecol. Indic.* **2021**, *129*, 107985. [CrossRef]
36. Han, L.; Yang, G.; Dai, H.; Xu, B.; Yang, H.; Feng, H.; Li, Z.; Yang, X. Modeling Maize Above-Ground Biomass Based on Machine Learning Approaches Using UAV Remote-Sensing Data. *Plant Methods* **2019**, *15*, 1746–4811. [CrossRef] [PubMed]
37. Ji, S.; Zhang, C.; Xu, A.; Shi, Y.; Duan, Y. 3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images. *Remote Sens.* **2018**, *10*, 75. [CrossRef]
38. Cui, Z.; Kerekes, J.P. Potential of Red Edge Spectral Bands in Future Landsat Satellites on Agroecosystem Canopy Green Leaf Area Index Retrieval. *Remote Sens.* **2018**, *10*, 1458. [CrossRef]
39. Zhang, F.; Zhou, G. Estimation of Vegetation Water Content using Hyperspectral Vegetation Indices: A Comparison of Crop Water Indicators in Response to Water Stress Treatments for Summer Maize. *BMC Ecol.* **2019**, *19*, 1–12. [CrossRef]
40. Mansaray, L.R.; Zhang, K.; Kanu, A.S. Dry Biomass Estimation of Paddy Rice With Sentinel-1A Satellite Data Using Machine Learning Regression Algorithms. *Comput. Electron. Agric.* **2020**, *176*, 105674. [CrossRef]
41. Wang, J.; Xiao, X.; Bajgain, R.; Starks, P.; Steiner, J.; Doughty, R.B.; Chang, Q. Estimating Leaf Area Index and Aboveground Biomass of Grazing Pastures Using Sentinel-1, Sentinel-2 and Landsat Images. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 189–201. [CrossRef]
42. Guo, H.; Liu, J.; Xiao, Z.; Xiao, L. Deep CNN-based Hyperspectral Image Classification Using Discriminative Multiple Spatial-spectral Feature Fusion. *Remote Sens. Lett.* **2020**, *11*, 827–836. [CrossRef]
43. Marshall, M.; Thenkabail, P. Advantage of Hyperspectral EO-1 Hyperion over Multispectral IKOMOS, Geoeye-1, Worldview-2, Landsat ETM+, and MODIS Vegetation Indices in Crop Biomass Estimation. *ISPRS J. Photogramm. Remote Sens.* **2015**, *108*, 205–218. [CrossRef]
44. Ribalta Lorenzo, P.; Tulczyjew, L.; Marcinkiewicz, M.; Nalepa, J. Hyperspectral Band Selection Using Attention-Based Convolutional Neural Networks. *IEEE Access* **2020**, *8*, 42384–42403. [CrossRef]
45. Zheng, Q.; Ye, H.; Huang, W.; Dong, Y.; Jiang, H.; Wang, C.; Li, D.; Wang, L.; Chen, S. Integrating Spectral Information and Meteorological Data to Monitor Wheat Yellow Rust at a Regional Scale: A Case Study. *Remote Sens.* **2021**, *13*, 278. [CrossRef]
46. Rao, K.; Williams, A.P.; Flefil, J.F.; Konings, A.G. SAR-enhanced Mapping of Live Fuel Moisture Content. *Remote Sens. Environ.* **2020**, *245*, 111797. [CrossRef]
47. Estévez, J.; Vicent, J.; Rivera-Caicedo, J.P.; Morcillo-Pallarés, P.; Vuolo, F.; Sabater, N.; Camps-Valls, G.; Moreno, J.; Verrelst, J. Gaussian Processes Retrieval of LAI from Sentinel-2 Top-of-atmosphere Radiance Data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 289–304. [CrossRef] [PubMed]
48. Wang, X.; Zhang, Y.; Atkinson, P.M.; Yao, H. Predicting Soil Organic Carbon Content in Spain by Combining Landsat TM and ALOS PALSAR Images. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *92*, 102182. [CrossRef]
49. Battude, M.; Al Bitar, A.; Morin, D.; Cros, J.; Huc, M.; Marais Sicre, C.; Le Dantec, V.; Demarez, V. Estimating Maize Biomass and Yield over Large Areas Using High Spatial and Temporal Resolution Sentinel-2 like Remote Sensing Data. *Remote Sens. Environ.* **2016**, *184*, 668–681. [CrossRef]
50. Kalaji, H.M.; Račková, L.; Paganová, V.; Swoczyna, T.; Rusinowski, S.; Sitko, K. Can Chlorophyll-a Fluorescence Parameters Be Used as Bio-indicators to Distinguish Between Drought and Salinity Stress in Tilia Cordata Mill? *Environ. Exp. Bot.* **2018**, *152*, 149–157. [CrossRef]
51. Li, C.; Chen, P.; Ma, C.; Feng, H.; Wei, F.; Wang, Y.; Shi, J.; Cui, Y. Estimation of potato chlorophyll content using composite hyperspectral index parameters collected by an unmanned aerial vehicle. *Int. J. Remote Sens.* **2020**, *41*, 8176–8197. [CrossRef]
52. Liu, N.; Xing, Z.; Zhao, R.; Qiao, L.; Li, M.; Liu, G.; Sun, H. Analysis of Chlorophyll Concentration in Potato Crop by Coupling Continuous Wavelet Transform and Spectral Variable Optimization. *Remote Sens.* **2020**, *12*, 2826. [CrossRef]
53. Yang, H.; Hu, Y.; Zheng, Z.; Qiao, Y.; Zhang, K.; Guo, T.; Chen, J. Estimation of Potato Chlorophyll Content from UAV Multispectral Images with Stacking Ensemble Algorithm. *Agronomy* **2022**, *12*, 2318. [CrossRef]
54. Ruszczak, B.; Boguszewska-Mańkowska, D. Deep potato—The Hyperspectral Imagery of Potato Cultivation with Reference Agronomic Measurements Dataset: Towards Potato Physiological Features Modeling. *Data Brief* **2022**, *42*, 108087. [CrossRef]
55. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef]
56. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
57. Nalepa, J.; Myller, M.; Tulczyjew, L.; Kawulok, M. Deep Ensembles for Hyperspectral Image Data Classification and Unmixing. *Remote Sens.* **2021**, *13*, 4133. [CrossRef]
58. Lin, C.Y.; Lin, C. Using Ridge Regression Method to Reduce Estimation Uncertainty in Chlorophyll Models Based on Worldview Multispectral Data. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1777–1780. [CrossRef]
59. Ziaja, M.; Bosowski, P.; Myller, M.; Gajoch, G.; Gumiela, M.; Protich, J.; Borda, K.; Jayaraman, D.; Dividino, R.; Nalepa, J. Benchmarking Deep Learning for On-Board Space Applications. *Remote Sens.* **2021**, *13*, 3981. [CrossRef]

Załącznik 5 – [A5]

Jiří Mikyška · Clélia de Mulatier ·
Maciej Paszynski ·
Valeria V. Krzhizhanovskaya ·
Jack J. Dongarra · Peter M. A. Sloot (Eds.)

LNCS 14074

Computational Science – ICCS 2023

23rd International Conference
Prague, Czech Republic, July 3–5, 2023
Proceedings, Part II



Springer

MOREMEDIA

Lecture Notes in Computer Science **14074**

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*
Wen Gao, *Peking University, Beijing, China*
Bernhard Steffen, *TU Dortmund University, Dortmund, Germany*
Moti Yung, *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Jiří Mikyška · Clélia de Mulatier ·
Maciej Paszynski · Valeria V. Krzhizhanovskaya ·
Jack J. Dongarra · Peter M. A. Sloot
Editors

Computational Science – ICCS 2023

23rd International Conference
Prague, Czech Republic, July 3–5, 2023
Proceedings, Part II

Springer

Editors
Jiří Mikyška
Czech Technical University in Prague
Prague, Czech Republic

Maciej Paszynski
AGH University of Science and Technology
Krakow, Poland

Jack J. Dongarra
University of Tennessee at Knoxville
Knoxville, TN, USA

Clélia de Mulatier
University of Amsterdam
Amsterdam, The Netherlands

Valeria V. Krzhizhanovskaya
University of Amsterdam
Amsterdam, The Netherlands

Peter M. A. Sloot
University of Amsterdam
Amsterdam, The Netherlands

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-36020-6 ISBN 978-3-031-36021-3 (eBook)
https://doi.org/10.1007/978-3-031-36021-3

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Estimating Chlorophyll Content from Hyperspectral Data Using Gradient Features

Bogdan Ruszczak[1,4(✉)], Agata M. Wijata[2,4(✉)], and Jakub Nalepa[3,4(✉)]

[1] Faculty of Electrical Engineering, Automatic Control and Informatics, Department of Informatics, Opole University of Technology, Opole, Poland
b.ruszczak@po.edu.pl
[2] Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland
awijata@ieee.org
[3] Faculty of Automatic Control, Electronics and Computer Science, Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland
jnalepa@ieee.org
[4] KP Labs, Gliwice, Poland

Abstract. Non-invasive estimation of chlorophyll content in plants plays an important role in precision agriculture. This task may be tackled using hyperspectral imaging that acquires numerous narrow bands of the electromagnetic spectrum, which may reflect subtle features of the plant, and inherently offers spatial scalability. Such imagery is, however, high-dimensional, therefore it is challenging to transfer from the imaging device, store and investigate. We propose a machine learning pipeline for estimating chlorophyll content from hyperspectral data. It benefits from the Savitzky-Golay filtering to smooth the (potentially noisy) spectral curves, and from gradient-based features extracted from such a smoothed signal. The experiments revealed that our approach significantly outperforms the state of the art according to the widely-established estimation quality metrics obtained for four chlorophyll-related parameters.

Keywords: machine learning · chlorophyll content · feature engineering · hyperspectral image · regression

1 Introduction

The agricultural sector has evolved over the years, in response to a growing demand for food, fiber and fuel [13]. The limited availability of land requires targeted management of resource production and leads to the increasing adoption of

This work was partially supported by The National Centre for Research and Development of Poland (POIR.04.01.04-00-0009/19). AMW was supported by the Silesian University of Technology, Faculty of Biomedical Engineering grant (07/010/BK_23/1023). JN was supported by the Silesian University of Technology Rector's grant (02/080/RGJ22/0026).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
J. Mikyška et al. (Eds.): ICCS 2023, LNCS 14074, pp. 196–203, 2023.
https://doi.org/10.1007/978-3-031-36021-3_18

precision agriculture [17]. In this context, remote sensing can easily become a tool for identifying soil and crop parameters, due to its intrinsic scalability [9,17]. In the case of agriculture, methods using multi- and hyperspectral remote sensing, capturing multispectral and hyperspectral images (MSI and HSI) are used, and non-invasive extraction of the chlorophyll content plays an increasingly important role, as it can ultimately lead to improving agricultural practices [7].

The majority of vegetation indices (VIs) were designed for multispectral sensors [14]. However, the wide bands in such imagery result in limited accuracy in the early detection of negative plant symptoms [1]. The use of HSIs, which are characterized by high spectral resolution, allows for the extraction of more details in the spectral response of an object [10]. Here, estimating chlorophyll content from hyperspectral data is commonly carried out by calculating the value of narrow-band VIs [19]. They include the Normalized Difference Vegetation Index (NDVI), Optimal Soil Adjusted Vegetation Index, Ratio Index and Difference Index [19]. Another parameter is the maximum quantum yield of photochemistry (Fv/Fm) [15]. Also, the Soil and Plant Analyzer Development (SPAD) tool is exploited, which measures the relative level of chlorophyll content in a crop taking into account the level of chlorophyll in the canopy. Finally, the performance index (PI) makes it possible to estimate the level of chlorophyll too [10].

Some approaches operate directly on the HSI data to estimate the chlorophyll content, hence they omit the stage of determining VIs. In such techniques, selected hyperspectral bands are analyzed—the encompass the Red and Near-Infrared (NIR) combination [6], the Blue, Green, Red, and NIR combined range [16], the Blue, Green and Red channels [17], or just the NIR band [4]. The bands undergo feature extraction (often followed by feature selection) in classic machine learning algorithms, whereas deep learning models benefit from automated representation learning over such data. The former group of techniques span across a variety of feature extractors and regression models, including continuous wavelet transforms (CWTs) [17], partial least square regression (PLSR) [18], kernel ridge regression [12] or regression using random forests. On the other hand, convolutional neural networks (CNNs) [14] and generative adversarial nets (GAN) [16] have been utilized for chlorophyll estimation as well.

Unfortunately, the data-driven algorithms are commonly validated over the in-house (private) data, following different validation procedures. This ultimately leads to the reproducibility crisis, and to inability to confront the existing approaches in a fair way [8]. In our recent work, we addressed this research gap and introduced a benchmark dataset, together with the validation procedure and a set of suggested metrics which should be used to quantify the generalization capabilities of machine learning models for chlorophyll estimation [11]. Here, we exploit this dataset and validation procedure to understand the abilities of the proposed processing chain, and to compare its estimation performance with 15 baseline models (for clarity, we focus on the best algorithms from [11]).

We tackle the problem of automated analysis of hyperspectral data using machine learning algorithms in the context of estimating the chlorophyll content (Sect. 2). We show that appropriately designed feature extractors fed into well-established supervised regression models can dramatically enhance their

operational capabilities. Our experiments (Sect. 3), performed over the recent dataset and following the validation suggested in [11], indicated that smoothing the spectral curves using the Savitzky-Golay filtering and extracting gradient-based features lead to significant improvements in the estimation quality when compared to the current state of the art. Also, we executed extensive computational experiments to understand the impact of the hyperparameter selection on the regression engine, and to optimize the hyperparameters of the system.

2 Materials and Methods

In this section, we summarize the dataset and the quantitative metrics used to assess the investigated algorithms (Sect. 2.1), together with our machine learning pipeline for estimating chlorophyll content from HSI (Sect. 2.2).

2.1 Dataset

We exploit the CHESS (**CH**lorophyll **ES**timation Data**S**et) dataset introduced in our recent study [11]—it was collected in the Plant Breeding and Acclimatization Institute -National Research Institute (IHAR-PIB) facility located in Central Poland (Jadwisin, Masovian Voivodeship) during the 2020 campaign (June—July, with three rounds of data acquisition, 4 weeks apart from each other). There were three flights over two sets of 12 plots resulting in 72 HSIs (150 bands, 460–902 nm, with the 2.2 cm ground sampling distance). In the plots, there were two potato varieties planted: *Lady Claire* (12 plots) and *Markies* (12 plots). The image data is accompanied with the in-situ measurements for each plot: (*i*) the SPAD, (*ii*) the maximum quantum yield of the PSII photochemistry (Fv/Fm), (*iii*) the performance of the electron flux to the final PSI electron acceptors, and (*iv*) relative water content (RWC), reflecting the degree of hydration of the leaf's tissue. In [11], we introduced the training-test split in which both subsets are equinumerous, and they are stratified following the distribution of each ground-truth parameter independently, so that both training and test subsets (each containing 36 plots) maintain a similar parameter's distribution.

To quantify the regression performance, we use the metrics, as suggested in [11]: the coefficient of determination R^2 which should be maximized ($\uparrow$; R^2 with one being the perfect score; its negative values indicate a worse fit than the average), mean absolute percentage error (MAPE), mean squared error (MSE) and mean absolute error (MAE)—all those measures should be minimized ($\downarrow$).

2.2 Estimating Chlorophyll Content Using Machine Learning

We exploit a processing chain, in which the input HSI undergoes feature extraction, and the features are fed to a regression model to predict the value of each parameter (we train four independent models). The algorithms at each step of the pipeline can be conveniently replaced by other techniques. We build upon several insights concerning the shape of the median spectral curves extracted for

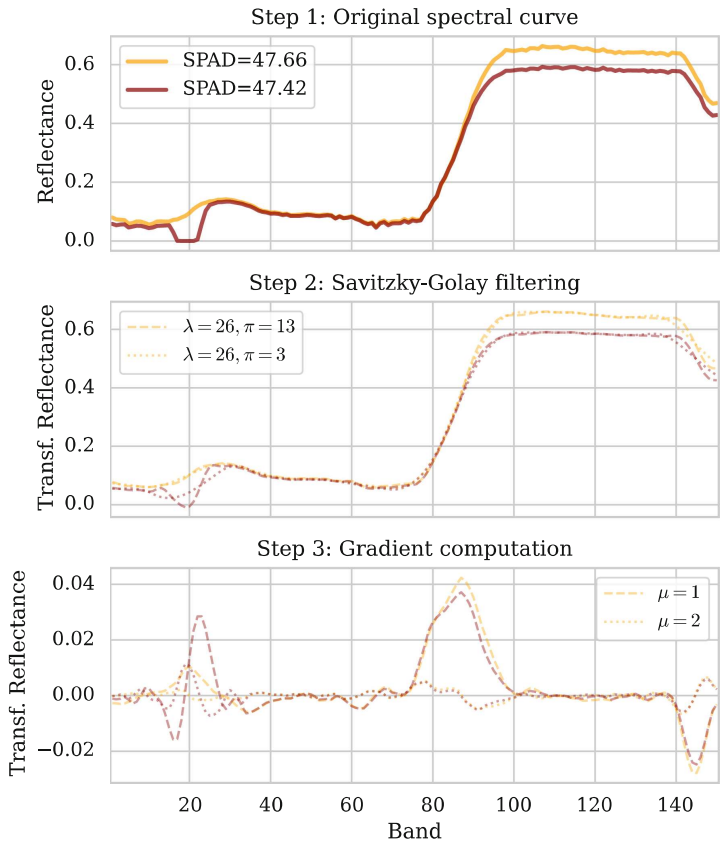


Fig. 1. The flowchart presenting the proposed feature engineering process.

the separate fields of interest (the spectral curves for all pixels are aggregated to generate a single median curve for the field). It is of note that some of the curves extracted for CHESS that should represent similar measurements (e.g., similar ground truth) do not look alike, and there exist noisy curves—this could be related to the difficulties in capturing enough light for selected spectral bands, which could easily lead to a narrower tonal range of the photosensitive camera.

To tackle the issue of the noisiness of spectral curves (and to increase its signal-to-noise ratio through removing high-frequency noise from the signal), the feature extraction stage is preceded by the filtering of spectral data using the Savitzky-Golay filter which may be considered as a generalized moving average filter. We aim at eliminating the influence of random noise and at reducing the drift phenomenon on the spectral reflection coefficient [5]. The Savitzky-Golay filter is given as the discrete convolution (h denotes the signal):

$$y[\pi] = \sum_{m=-\lambda}^{\lambda} h[m] * [\pi - m] = \sum_{m=\pi-\lambda}^{\pi+\lambda} h[\pi - m] * [m], \tag{1}$$

where $2\lambda + 1$ is the length of the approximation interval, and π is the polynomial order, both being the tunable hyperparameters. In Fig. 1, we present the example of the original spectral curves for two fields with similar SPAD values amounting to 47.66 and 47.42 (first row) which underwent Savitzky-Golay filtering for two

hyperparameter sets ($\lambda = 26$, $\pi = 13$ and $\lambda = 26$, $\pi = 3$), resulting in slightly different shapes of the smoothed signal (second row). In both cases, the small "noisy" variations of the original curve were removed across the entire spectrum.

Once the original spectral curve is filtered, we extract the gradients which constitute the feature vectors for each field. We build upon the observation that the detection of anomalies in the signal data can be supported by the use of a gradient elaborated over the spectral curve, which allows us to capture the subtle characteristics in such data [2]. The feature extraction may be performed μ times, with $\mu = 0$ denoting the original curve (i.e., the feature vector includes the original reflectance values). Such feature vectors of size $\mathcal{B}$, where $\mathcal{B}$ is the number of hyperspectral bands (here, $\mathcal{B} = 150$), are fed to the regression model.

3 Experimental Validation

We investigate the linear machine learning models: (*i*) linear regressors with L_1, and (*ii*) L_2 regularization, (*iii*) support vector machines with linear kernel, and (*iv*) the elastic net with regularization, all implemented in `Scikit-learn`. Savitzky-Golay filtering was implemented in `SciPy`, and feature extraction in `NumPy`. We focus on the linear regression models to avoid heavily parameterized techniques—due to this assumption, we were able to extensively evaluate thousands of models in a reasonable time (27 840 model's configurations). At the same time, we maintained the high interpretability of the study.

For each model, we optimized its hyperparameters (this fine-tuning was performed following the 5-fold cross-validation strategy over the training set):

- Linear regression with L1 regularization, for: $\alpha \in \{10^{-15}, 10^{-14}, \ldots, 10^{15}\}$,
- Linear regression with L2 regularization, for: $\alpha \in \{10^{-15}, 10^{-14}, \ldots, 10^{15}\}$,
- Support vector machine with linear kernel, for: $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{8}\}$, and the maximum number of iterations $\mathcal{I} \in \{500, 1000, 2500\}$,
- Elastic net with L_1 regularization, for $\alpha \in \{10^{-15}, 10^{-14}, \ldots, 10^{15}\}$, and $L_{1ratio} \in \{0.05, 0.1, \ldots, 0.9\}$.

Similarly, the feature extraction is a parameterized step, as it may be performed multiple times. We denoted the number of gradient runs as μ, and by $\mu = 0$ we report the results obtained for the regression models operating over the original curves. Therefore, the full configuration for the performed experimental search was as follows: $\mu \in \{0, 1, 2, 3\}, \lambda \in \{1, 2, \ldots 30\}, \pi \in \{1, 2, \ldots, 29\}$.

In Fig. 2, we depict the impact of π, λ, and μ on the R^2 coefficient for the elastic net models predicting the SPAD parameter. Here, we focused on a single machine learning model (with default parameterization) to verify the importance of signal filtering and feature extraction on the regression capabilities of the algorithm. Albeit the insights learned from this experiment may not be generalizable to other models, we anticipate that a similar trend would be observed, as feature engineering constitutes one of the most important aspects of building machine learning pipelines [3]. We can observe that the exhaustive traversal of the search spaces allows to indicate their most promising regions for the (λ, π)

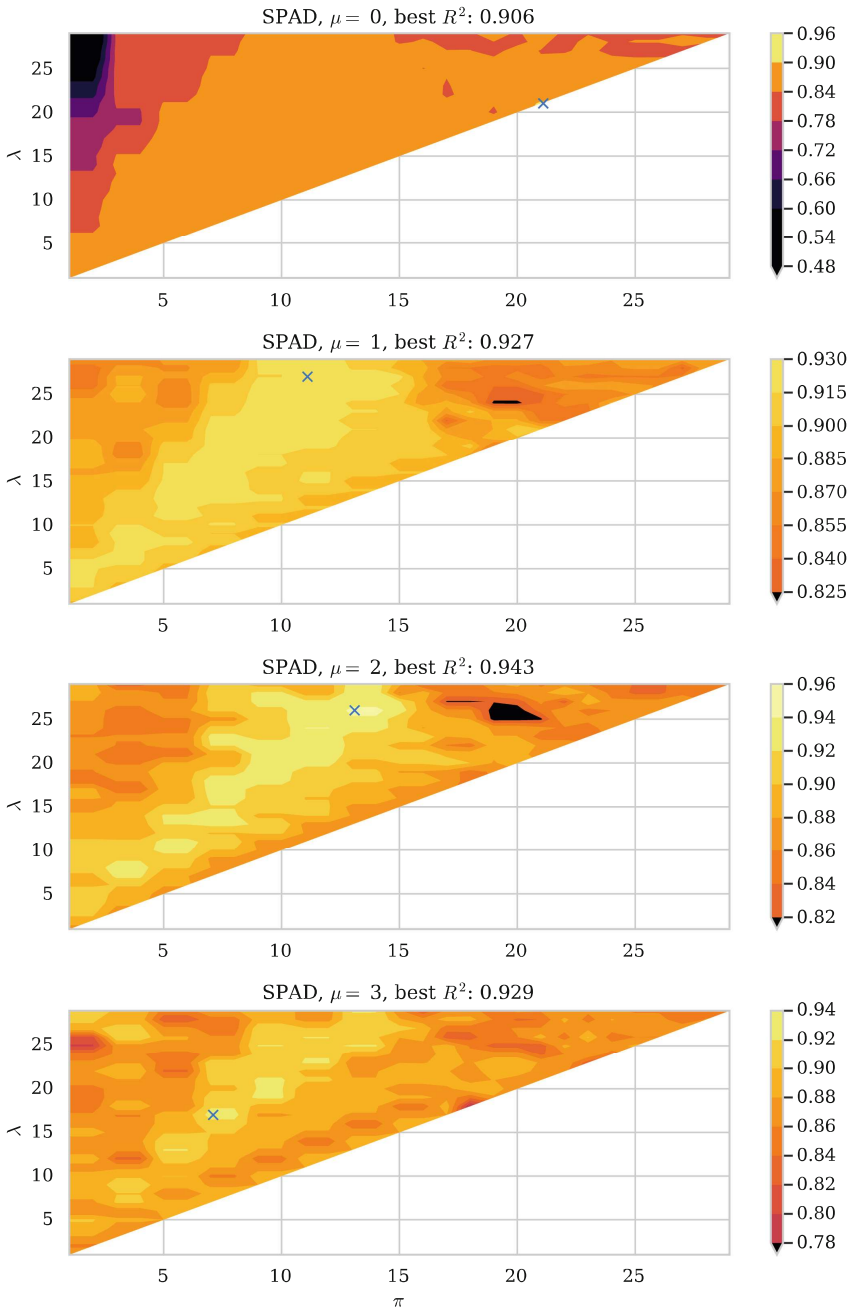


Fig. 2. The R^2 metric obtained for various filtering and feature extraction settings. The light blue × marker indicates the best configuration for each gradient level (μ). (Color figure online)

configurations, which remain consistent for the gradient-extraction levels (μ). However, the exact position of the best parameterization differs across μ's.

The optimized models outperform *Baseline* (the best-known R^2 values from the literature [11])—in Table 1, we report the optimized hyperparameters for the models offering the best regression. The R^2 measure notably increased for all parameters, whereas the regression errors, e.g., MAE, decreased by 53.5%, 23.3%, 30.9%, and 25.0% for SPAD, FvFm, PI, and RWC. The experiments

indicated that appropriate feature engineering, which involves Savitzky-Golay filtering of the median spectral curves followed by feature extraction, allows the elaboration of high-quality models for estimating chlorophyll-related parameters.

Table 1. The best results for all quality metrics, elaborated for the best parameterization (Savitzky-Golay filtering, feature extraction and regression models).

Param.	**Model configuration**	μ	π	λ	**MAPE** $\downarrow$	R^2 $\uparrow$	**MAE** $\downarrow$	**MSE** $\downarrow$
SPAD	Elastic net ($\alpha = 10^{-1}$, $L_{1ratio} = 0.5$)	2	26	12	0.035	0.943	0.756	3.012
	Linear regr. with L_2 ($\alpha = 2.5 \times 10^{-5}$)		—		0.072	0.827	1.625	9.095
FvFm	Linear regr. with L_2 ($\alpha = 10^{-3}$)	0	22	17	0.030	0.764	0.016	0.001
	Linear regr. with L_2 ($\alpha = 5 \times 10^{-5}$)		—		0.036	0.727	0.021	0.001
PI	SVM with linear kernel ($C = 2$, $\mathcal{I} = 10^3$)	1	29	20	0.401	0.837	0.194	0.083
	Linear regr. with L_2 ($\alpha = 10^{-11}$)		—		0.532	0.677	0.280	0.169
RWC	Linear regr. with L_1 ($\alpha = 10^{-1}$)	2	22	11	0.010	0.911	0.706	1.089
	Linear regr. with L_2 ($\alpha = 10^{-3}$)		—		0.013	0.859	0.941	1.731

4 Conclusions

We exploited HSIs for the non-invasive estimation of chlorophyll-related parameters in plants and proposed a machine learning technique for this task. To deal with subtle signal noise, we utilized the Savitzky-Golay smoothing filter that is followed by the gradient-based feature extractor. The experiments revealed that our techniques outperform the state of the art, as quantified using four chlorophyll-related parameters. The coefficient of determination (R^2) achieved by our techniques reached 0.943 (compared to 0.827 reported for the best model in [11], therefore we obtained the improvement of 14%), 0.764 (0.727, improvement of 5%), 0.837 (0.667, improvement of 25%), 0.911 (0.859, improvement of 6%) for SPAD, FvFm, PI, and RWC. Also, we showed that employing the Savitzky-Golay smoothing brings improvements in the generalization of a model trained over the gradient-based features extracted from such filtered signal.

References

1. Adão, T., et al.: Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry. Remote Sens. 9(11) (2017)
2. Chen, J., Lu, M., Chen, X., Chen, J., Chen, L.: A spectral gradient difference based approach for land cover change detection. ISPRS J. Photogramm. Remote. Sens. **85**, 1–12 (2013)
3. Chicco, D., Oneto, L., Tavazzi, E.: Eleven quick tips for data cleaning and feature engineering. PLoS Comput. Biol. **18**(12), 1–21 (2022)
4. Gorretta, N., Nouri, M., Herrero, A., Gowen, A., Roger, J.M.: Early detection of the fungal disease "apple scab" using SWIR hyperspectral imaging. In: 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1–4 (2019)

5. Guo, C., et al.: Predicting Fv/Fm and evaluating cotton drought tolerance using hyperspectral and 1D-CNN. Front. Plant Sci. **13**, 3700 (2022)
6. Huynh, N.H., Böer, G., Schramm, H.: Self-attention and generative adversarial networks for algae monitoring. Eur. J. Remote Sens. **55**(1), 10–22 (2022)
7. Jin, X., Li, Z., Feng, H., Ren, Z., Li, S.: Deep neural network algorithm for estimating maize biomass based on simulated sentinel 2A vegetation indices and leaf area index. Crop J. **8**(1), 87–97 (2020)
8. Nalepa, J., Myller, M., Kawulok, M.: Validating hyperspectral image segmentation. IEEE Geosci. Remote Sens. Lett. **16**(8), 1264–1268 (2019)
9. Ponnusamy, V., Natarajan, S.: Precision agriculture using advanced technology of IoT, unmanned aerial vehicle, augmented reality, and machine learning. In: Gupta, D., de Hugo, C., Albuquerque, V., Khanna, A., Mehta, P.L. (eds.) Smart Sensors for Industrial Internet of Things: Challenges, Solutions and Applications, pp. 207–229. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-52624-5_14
10. Ruszczak, B., Boguszewska-Mańkowska, D.: Deep potato - the hyperspectral imagery of potato cultivation with reference agronomic measurements dataset: towards potato physiological features modeling. Data Brief **42**, 108087 (2022)
11. Ruszczak, B., Wijata, A.M., Nalepa, J.: Unbiasing the estimation of chlorophyll from hyperspectral images: a benchmark dataset, validation procedure and baseline results. Remote Sens. **14**(21) (2022)
12. Singhal, G., Bansod, B., Mathew, L., Goswami, J., Choudhury, B., Raju, P.: Estimation of leaf chlorophyll concentration in turmeric (curcuma longa) using high-resolution unmanned aerial vehicle imagery based on kernel ridge regression. J. Indian Soc. Remote Sens. **47**, 1–12 (2019)
13. Sishodia, R.P., Ray, R.L., Singh, S.K.: Applications of remote sensing in precision agriculture: a review. Remote Sens. **12**(19) (2020)
14. Wang, J., et al.: Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. ISPRS J. Photogramm. Remote. Sens. **154**, 189–201 (2019)
15. Wen, S., Shi, N., Lu, J., Gao, Q., Yang, H., Gao, Z.: Estimating chlorophyll fluorescence parameters of rice (Oryza sativa L.) based on spectrum transformation and a joint feature extraction algorithm. Agronomy **13**(2) (2023)
16. Yan, T., et al.: Combining multi-dimensional convolutional neural network (CNN) with visualization method for detection of aphis gossypii glover infection in cotton leaves using hyperspectral imaging. Front. Plant Sci. 12 (2021)
17. Yue, J., Zhou, C., Guo, W., Feng, H., Xu, K.: Estimation of winter-wheat aboveground biomass using the wavelet analysis of unmanned aerial vehicle-based digital images and hyperspectral crop canopy images. Int. J. Remote Sens. **42**(5), 1602–1622 (2021)
18. Zhang, J., et al.: Detection of canopy chlorophyll content of corn based on continuous wavelet transform analysis. Remote Sens. **12**(17) (2020)
19. Zhang, Y., et al.: Estimating the maize biomass by crop height and narrowband vegetation indices derived from UAV-based hyperspectral images. Ecol. Ind. **129**, 107985 (2021)

Załącznik 6 – [A6]



Single- and Multi-Objective Genetic Optimization of Hyperspectral Image Pre-Processing

Bogdan Ruszczak*
Opole University of Technology
Opole, Poland
b.ruszczak@po.edu.pl

Sabina Kubiciel-Lodzińska
Opole University of Technology
Opole, Poland
s.kubiciel-lodzinska@po.edu.pl

Katarzyna Rudnik
Opole University of Technology
Opole, Poland
k.rudnik@po.edu.pl

Zineb Bougriche
Opole University of Technology
Opole, Poland
z.bougriche@student.po.edu.pl

Agata M. Wijata
Silesian University of Technology
Gliwice, Poland
agata.wijata@polsl.pl

Jakub Nalepa
Silesian University of Technology
Gliwice, Poland
jakub.nalepa@polsl.pl

ABSTRACT

Hyperspectral imaging has broad applications, particularly in Earth observation and remote sensing, due to its ability to provide large-scale, non-invasive environmental and agricultural monitoring. However, its high dimensionality and acquisition conditions pose challenges for data transfer, storage, and downstream analysis. Pre-processing algorithms are thus essential for preparing hyperspectral images for further analysis, but they rely on multiple hyperparameters, and their incorrect selection degrades the performance of machine learning models operating on the pre-processed imagery. To address this, we introduce genetic algorithms—both single- and multi-objective—for optimizing pre-processing in multi-target hyperspectral regression. Experiments on real-world hyperspectral data demonstrate that our approach significantly enhances the models for chlorophyll content estimation from hyperspectral images.

CCS CONCEPTS

• **Computing methodologies** → **Bio-inspired approaches**; *Supervised learning*; • **Applied computing** → *Agriculture*.

KEYWORDS

Genetic algorithms, Hyperspectral imaging, Hyperparameter optimization, Multi-objective optimization, Supervised learning,

ACM Reference Format:
Bogdan Ruszczak, Sabina Kubiciel-Lodzińska, Katarzyna Rudnik, Zineb Bougriche, Agata M. Wijata, and Jakub Nalepa. 2025. Single- and Multi-Objective Genetic Optimization of Hyperspectral Image Pre-Processing. In *Genetic and Evolutionary Computation Conference (GECCO '25 Companion), July 14–18, 2025, Malaga, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3712255.3726774

1 INTRODUCTION

Extracting knowledge and actionable insights from the hyperspectral images (HSIs) [4, 12], capturing up to hundreds of narrow spectral bands within the electromagnetic spectrum, is challenging due to many reasons. Such imagery is highly-dimensional [5], with potentially redundant and noisy information. Of note, the noise level commonly results from various calibration limitations, the influence of varying lighting conditions, and other interferences [10]. To address such issues, a variety of HSI pre-processing algorithms have been introduced so far [2, 11]. In [8], it was shown that the application of successive image transformations can substantially improve the downstream models operating over such prepared HSIs. This method exploited a multi-step pre-processing procedure:

- it **extracts the spectral curves** from the HSI,
- it **applies the Savitzy-Golay filter** to the spectral curves,
- it **applies the gradient transformation and standard scaler** to the resulting data.

As the study reports, such pre-processing can substantially improve regression tasks, but it also has some serious limitations:

- The method is heavily parameterized (both filtering and gradient transformations), and the incorrectly selected hyperparameter values notably deteriorate its performance.
- There is no universal recommendation on how to set those parameters for different downstream tasks. In [8], four different examples were presented, and the hyperparameter values were substantially different for all of them,
- The exhaustive (grid) search for optimizing these hyperparameters would require performing an infeasible number of evaluations, especially in the embedded approach, which necessitates training a downstream machine learning model to determine the quality of hyperparameters.

We address the problem of automated selection of the pivotal hyperparameters affecting a pre-processing chain used to prepare HSIs for downstream analysis. We specifically target the problem of extracting multiple chlorophyll content parameters from HSIs. Our experiments, performed over a benchmark of real-world HSIs, showed that the evolutionary algorithms—both single- and multi-objective—lead to substantial improvements in the chlorophyll content estimation performed by the downstream machine learning models. Our framework allows to elaborate hyperparameters which are either *specialized*, i.e., obtained for a specific target and model (single-objective optimization by a genetic algorithm, GA) or more *universal* (multi-objective optimization using the non-dominated sorting GA, NSGA-II), hence independent of the target and model.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

GECCO '25 Companion, July 14–18, 2025, Malaga, Spain

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1464-1/2025/07
https://doi.org/10.1145/3712255.3726774

2 MATERIALS AND METHODS

2.1 Benchmark Dataset

We utilized the CHESS (**CH**lorophyll **ES**timation Data**S**et) set [6, 7], which also provides initial benchmarking results for the set of four different estimation tasks (related to estimating the chlorophyll content from HSIs: *SPAD*, *FvFM*, *RWC*, and water content: *RWC* in leaves visible in HSIs). The in-situ and image acquisitions were performed during the 2020 campaign in Poland (June—July, three rounds of unmanned aerial vehicle-attached hyperspectral camera flights, 4 weeks apart from each other). The 72 acquired HSIs (150 bands, 460–902 nm, with the 2.2 cm ground sampling distance) represent 12 experimental potato plots for which the corresponding measurements were carried out shortly after each flight.

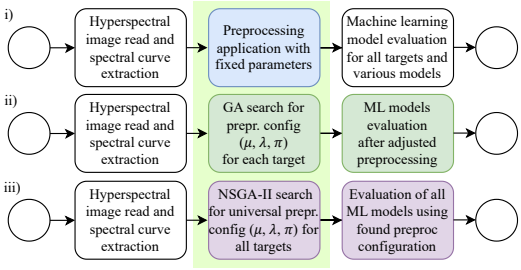


Figure 1: Variants of the pre-processing optimization process.

2.2 Evolving Pre-Processing Hyperparameters Using Single- and Multi-Objective GAs

In this study, we exploit a *i*) classic genetic algorithm (GA) with elitism [13], and the *ii*) non-dominated sorting GA (NSGA-II) [3] to evolve the hyperparameter values of the HSI pre-processing algorithm (Figure 1). There are three pivotal parameters that need to be carefully selected—two Savitzky-Golay filter parameters [9]: window size (denoted here as λ) and the polynomial order (π), and the order of the gradient applied in the next pre-processing step (μ). This gradient can be repeated multiple times and hence $\mu\times$. Therefore, μ reflects the order of the derivative of the original data, and for $\mu = 0$, we preserve the original spectral curve. Thus, each *individual* in an evolved population, using both GA and NSGA-II, encodes the values of those integer-type hyperparameters: $\mu \in (0 \dots 5), \lambda \in (0 \dots 31)$, and $\pi \in (1 \dots 30)$ (the ranges for each hyperparameter are selected as suggested in [8]). Of note, λ cannot have values lower than or equal to π, hence the potential candidate solutions violating this constraint are pruned from the population.

2.2.1 Single-Objective Optimization: GA. In the **GA** evolutionary process, a population of $N = 100$ individuals evolves for up to $G_{\max} = 500$ consecutive generations (the evolution is terminated if the best fitness does not change for 75 generations). In each generation, parental pairs are selected for *crossover* with the uniform crossover probability—two children are generated, each inheriting the hyperparameter value from one of the parental solutions. Afterwards, each offspring is *mutated* with probability $\mathcal{P}_m$ by randomly perturbing its hyperparameters (the parameter is replaced with a random value from the corresponding feasible range). The next generation encompasses N solutions—here, we employ elitism, and always include 5% best individuals from the previous generation. The quality of the encoded hyperparameter values is quantified using the *fitness* function, following the embedded technique. Here, assess the quality of the hyperparameters by investigating the performance of the downstream supervised machine learning model trained using HSIs pre-processed with the encoded hyperparameter values—we calculate the fitness within the training set following two-fold cross-validation. Therefore, the unseen test set is kept aside and never used during evolution and is utilized to quantify the generalization capabilities of the resulting models. To thoroughly validate the quality of the individual p, we focus on all regression targets $T_i \in (SPAD, RWC, FvFm, PI)$ and verify multiple machine learning regression models separately. In this study, we build upon an array of established models, including the following model architectures: linear regression model with L_1 regularization (later denoted as Lasso), linear regression model with L_2 regularization (Ridge), ElasticNet (with simultaneous L_1 and L_2 regularization) and the support vector machine-based regressor with linear kernel (LinearSVR). To maintain comparability of the results, for all models, we used their hyperparameter configurations reported in [7]. As GA is *single-objective*; therefore, to optimize hyperparameters for all chlorophyll parameter targets and models, we need to perform 16 evolutions (four targets, four models).

2.2.2 Multi-Objective Optimization: NSGA-II. A multi-objective evolutionary optimization approach might help identify a hyperparameter configuration that could be universal for different targets and models used to quantify the fitness. To tackle this issue, we exploit NSGA-II, and calculate four fitness values for each individual during the evolution (for all targets: *SPAD*, *RWC*, *FvFm*, *PI*). In NSGA-II, we utilize an adaptive mutation operator, as suggested in [1], and the mutation probability becomes $\mathcal{P}_m = 0.2$. Having a single, universal configuration of hyperameters can not only be elaborated faster when compared to performing multiple parallel optimization processes, but also may it be easier to generalize to new, perhaps unknown target variables that should be estimated from pre-processed HSIs. A multi-objective NSGA-II optimization can bring both functional and non-functional improvements in the parameter estimation system operating on HSIs.

3 EXPERIMENTAL STUDY

The objective is to verify whether using single- and multi-objective evolutionary algorithms may lead to improving the performance of downstream models operating on real-world HSIs pre-processed using the routines with the optimized hyperparameters. The generalization of the models is quantified using the mean absolute error (*MAE*), mean percentage absolute error (*MAPE*) and the mean squared error (*MSE*) obtained over the test set. To ensure consistency with previous works, we strictly follow the validation procedure suggested in [7]. To help the reader conveniently compare the results of this study with the benchmark [7], we introduced a new relative metric: *rMAE*. It takes into account the outcome of our investigation, and for each *i-th* configuration found by the evolutionary algorithm for which we measured the error $MAE_{GA,i}$ with the benchmark result $MAE_{B,i}$ reported in [7]: $rMAE_i = MAE_{GA,i} \,/\, MAE_{B,i}$. Finally, a grid searching the entire

solution space of all hyperparameter combinations would require **2480 evaluations for each model-target combination**.

Table 1: The configurations evolved for all targets and machine learning models using a single-objective GA. The boldfaced entries indicate optimal hyperparameters ($rMAE = 1.0$).

Target	Model	μ	λ	π	*MAPE*	*MAE*	*MSE*	*rMAE*
SPAD	Lasso	3	26	9	0.038	0.702	4.176	**1.000**
	Ridge	4	26	4	0.041	0.718	4.397	**1.000**
	ElasticNet	2	27	12	0.039	0.698	3.401	**1.000**
	LinearSVR	1	31	13	0.054	1.069	6.120	**1.000**
RWC	Lasso	2	8	3	0.010	0.521	1.247	**1.000**
	Ridge	3	21	12	0.010	0.443	1.333	**1.000**
	ElasticNet	3	21	12	0.011	0.578	1.492	0.999
	LinearSVR	0	29	20	0.012	0.730	1.522	**1.000**
FvFm	Lasso	4	18	14	0.031	0.011	0.001	0.926
	Ridge	4	27	20	0.033	0.011	0.001	**1.000**
	ElasticNet	2	14	4	0.033	0.011	0.001	0.999
	LinearSVR	0	24	21	0.034	0.014	0.001	**1.000**
PI	Lasso	3	3	2	0.460	0.148	0.137	0.896
	Ridge	4	31	7	0.570	0.154	0.135	0.879
	ElasticNet	4	16	12	0.382	0.154	0.111	0.970
	LinearSVR	1	14	2	0.373	0.135	0.151	**1.000**

As reported in Table 1, a single-objective GA independently optimized the hyperparameters for each target chlorophyll parameter, leading to substantially different configurations obtained for the target chlorophyll parameters and underlying machine learning models, as both aspects affect the hyperparameters' choice. For almost all models applied to estimate the *SPAD* and *RWC* targets, the results were of the same quality as offered by the hyperparameters returned by the exhaustive grid search ($rMAE \approx 1.000$). Figure 2 further shows the convergence capabilities of a single-objective GA, specialized in optimizing the hyperparameters for a given chlorophyll target parameter and regression model.

These results were obtained in a fraction of the time required to traverse the entire solution space by exhaustive search, as in Table 2. It shows the practical utility of the framework, as—for the pre-processing routines that are parameterized with a larger number of hyperparameters—the exhaustive search would quickly become impossible, especially for wrapper techniques, as they require training a model to assess the quality of hyperparameters. Additionally, Table 2 shows that the selection of a machine learning model notably affects the overall GA search time—therefore, selecting efficient (fast to train) models or building surrogate models that approximate the fitness landscape should be further researched to accelerate the optimization procedure. Overall, the proportion of the total optimization time measured for our single-objective GA and the exhaustive search indicates that GA needed about 5%, or less time than grid searching to reach the solution with similar quality. The number of evaluations was much lower for GA, ranging from 79 to 184 (when compared to 2480 evaluations for each model-target combination in the exhaustive search).

On the one hand, elaborating hyperparameter values for specific models and target variables that are to be estimated from HSIs allow for obtaining high-quality models targeting a well-defined downstream task. However, such optimizations may easily lead to a large number of hyperparameter sets (in this study, 80 solutions). Additionally, some of the configurations can optimize a single quality metric and still fail in the others—this was visible for e.g., *MAPE*,

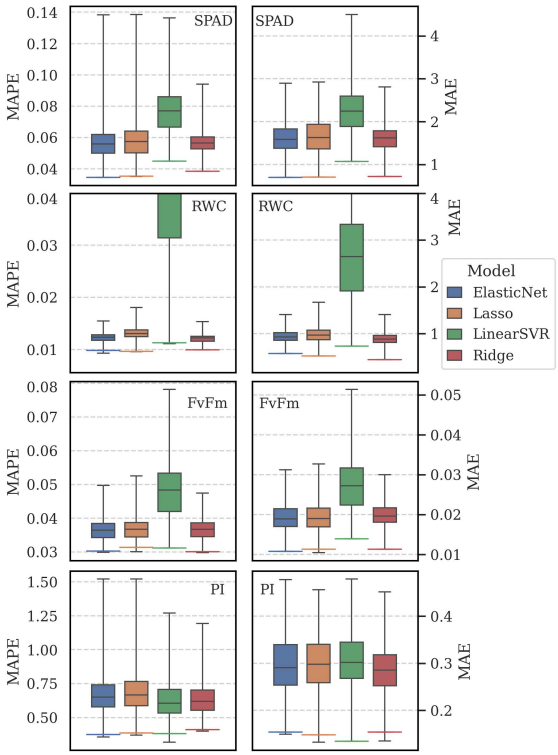


Figure 2: The final results for the GA search (horizontal colored line) and box plot for the exhaustive search population for all configurations. The results were gathered for all four target parameters and models, and compared using *MAPE* (left) and *MAE* (right)—both should be minimized.

Table 2: The optimization time (τ_{sum}, in seconds) of a single-objective GA compared with the exhaustive grid searching (denoted as GS) of the entire solution space.

Model	Target	GA search Iter.	GA search τ_{train}[s]	GA search τ_{search}[s]	GA search τ_{sum}[s]	GS τ_{train}[s]	GS to GA ratio
Lasso	SPAD	119	26.0	2.4	28.4	542.5	5.2%
	RWC	94	20.8	1.8	22.6	548.6	4.1%
	FvFm	88	18.9	1.7	20.6	532.7	3.9%
	PI	123	26.7	2.3	29.0	538.7	5.4%
Ridge	SPAD	87	8.8	1.7	10.5	249.9	4.2%
	RWC	86	8.8	1.7	10.5	253.5	4.1%
	FvFm	87	8.9	1.7	10.6	253.4	4.2%
	PI	79	8.1	1.6	9.7	254.7	3.8%
ElasticNet	SPAD	87	22.6	1.7	24.3	644.7	3.8%
	RWC	91	24.2	1.8	26.0	660.2	3.9%
	FvFm	88	20.7	1.7	22.5	584.5	3.8%
	PI	89	22.0	1.7	23.7	613.2	3.9%
LinearSVR	SPAD	87	11.0	1.7	12.7	312.3	4.1%
	RWC	184	25.2	3.5	28.7	339.8	8.5%
	FvFm	88	13.7	1.7	15.3	385.0	4.0%
	PI	126	18.5	2.5	21.0	363.8	5.8%

especially with the *PI* chlorophyll-related parameter. Interestingly, even when *PI* was the main parameter driving the optimization process, the resulting hyperparameters did not allow for obtaining acceptable-quality estimations from HSIs. It indeed means that the

captured images might not be detailed enough to precisely extract *PI*—this parameter was removed from our further considerations, as it was clearly outlying.

Table 3: The results for the NSGA-II optimization for the fitness functions (fit) with the Pareto-optimal configurations.

Found parameter settings			Fitness function values (*MAPE*)		
μ	λ	π	fit_{SPAD}	fit_{RWC}	fit_{FvFm}
2	8	2	0.0452	0.0097	0.0332
2	**14**	**6**	0.0465	**0.0096**	0.0381
2	16	6	0.0411	0.0128	0.0304
2	17	6	0.0405	0.0125	0.0330
2	20	8	0.0375	0.0122	0.0333
2	22	9	0.0366	0.0131	0.0374
2	22	11	0.0398	0.0099	0.0330
2	23	10	0.0394	0.0118	0.0351
2	25	13	0.0377	0.0099	0.0370
2	**28**	**9**	**0.0353**	0.0127	0.0428
3	29	23	0.0513	0.0127	0.0322
4	18	13	0.0498	0.0119	0.0324
4	20	14	0.0527	0.0120	0.0321
4	**20**	**17**	0.0592	0.0133	**0.0301**
4	28	27	0.0527	0.0115	0.0307

To verify a hypothesis saying that a multi-objective genetic optimization approach (we used NSGA-II) can help identify a hyperparameter configuration that could be universal for different targets and models, we report the obtained parameterizations in Table 3, together with the corresponding fitness values calculated for each parameter (note that all of those metrics should be optimized by NSGA-II). As mentioned above, we focus on the *SPAD*, *RWC*, *FvFm* chlorophyll parameters (without an outlying *PI*). This table gathers the fitness levels (fit_{SPAD}, fit_{RWC}, fit_{FvFm}) for all Pareto-optimal combinations of μ, λ and π. We can observe that 15 Pareto-optimal configurations were elaborated by NSGA-II (versus 80 configurations obtained using a single-objective GA). However, using the NSGA-II-optimized hyperparameters and the results obtained for each target (various objectives), one could select a configuration according to their preference for the modeled data. For instance, if one were to prioritize the model's performance for the *SPAD* estimation and wanted to obtain the lowest *MAPE* for this target chlorophyll parameter, the best possible configuration for this setup would be one that is close to the optimal for other parameters as well. In this particular example, the parameters would become $\mu = 2, \lambda = 28$ and $\pi = 9$, with the final score of $MAPE_{SPAD} = 0.0353$.

In Figure 3, we compare the selected Pareto-optimal configurations. To enhance the readability of this plot, we increased the size of the points according to the error level of the *RWC* estimation—the $fit(RWC)$ value is rendered on the OY–axis. Additionally, we colored the *FvFm* estimation errors, so that the smaller errors are dark green and the larger errors go from yellow to red for the OZ–axis that reflects $fit(FvFm)$. The three-dimensional (3D) surface of the Pareto-optimal solutions is sparse and should allow the selection of the configuration supporting appropriate criteria.

4 CONCLUSIONS

The practical application of HSIs requires a robust pre-processing method that would be effective and, if possible, does not require a time-consuming tuning process. In this study, we introduced the single- and multi-objective genetic algorithms to the task of optimizing the pivotal hyperparameters of the HSI pre-processing

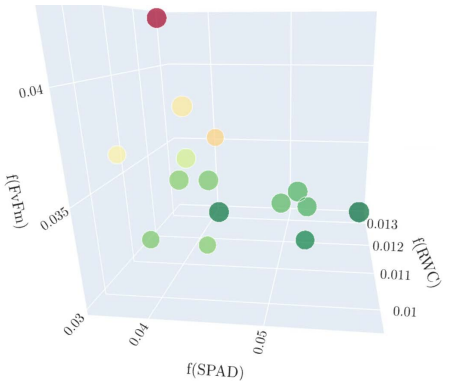


Figure 3: The fitness for three objectives (*SPAD*, *RWC*, *FvFm*) that reflect the *MAPE* for the machine learning models using the NSGA-II-optimized hyperparameter configurations.

routines, and verified their usability in a real-world application concerned with estimating various chlorophyll-related parameters from such imagery. The experimental study, performed over an established hyperspectral benchmark revealed that the proposed algorithms lead to obtaining high-quality parameterizations which, in turn, lead to well-generalizing machine learning models. The evolutionary techniques offered fast operation, converging in a fraction of the time required by exhaustive grid searching.

ACKNOWLEDGMENTS

This work was financially supported by the Opole University of Technology as part of the DELTA project no. 272/24, as well as by The National Centre for Research and Development of Poland under project POIR.04.01.04-00-0009/19.

REFERENCES

[1] A.G. Carvalho and A.R. Araujo. 2009. Improving NSGA-II with an adaptive mutation operator. In *Gecco'09*. ACM. https://doi.org/10.1145/1570256.1570387
[2] D. Cozzolino et al. 2023. An overview of pre-processing methods available for HSI applications. *Microchem J* (2023). https://doi.org/10.1016/j.microc.2023.109129
[3] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE TEVC* 6, 2 (2002), 182–197.
[4] M.F. Guerri et al. 2024. Deep learning techniques for HSI analysis in agriculture: A review. *IOJPRS* 12 (2024). https://doi.org/10.1016/j.ophoto.2024.100062
[5] D. Ruiz et al. 2021. Dim. reduction of HSI of vegetation and crops based on self-organized maps. *InfProcessAgric* (2021). https://doi.org/10.1016/j.inpa.2020.07.002
[6] B. Ruszczak and D. Boguszewska-Mańkowska. 2022. Deep potato – The hyperspectral imagery of potato cultivation with reference agronomic measurements dataset. *Data in Brief* 42 (2022), 108087. https://doi.org/10.1016/j.dib.2022.108087
[7] B. Ruszczak, A.M. Wijata, and J. Nalepa. 2022. Unbiasing the Estimation of Chlorophyll from Hyperspectral Images: A Benchmark Dataset, Validation Procedure and Baseline Results. *Rem. Sens.* 14, 21 (2022). https://doi.org/10.3390/rs14215526
[8] B. Ruszczak, A.M. Wijata, and J. Nalepa. 2023. Estimating Chlorophyll Content from Hyperspectral Data Using Gradient Features. In *ICCS'23*, Vol. 14074. Springer, 196–203. https://doi.org/10.1007/978-3-031-36021-3_18
[9] R.W. Schafer. 2011. What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Processing Mag.* 28, 4 (2011), 111–117. https://doi.org/10.1109/MSP.2011.941097
[10] O. Torun, Seniha E. Yuksel, et al. 2024. Hyperspectral image denoising via self-modulating convolutional neural nets. *Signal Processing* 214 (2024), 109248.
[11] M. Vidal and J.M. Amigo. 2012. Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometr. Intell. Lab. Syst.* 117 (2012), 138–148.
[12] C. Wang, B. Liu, et al. 2021. A review of deep learning used in the hyperspectral image analysis for agriculture. *Artificial Intelligence Review* 54, 7 (01 Oct 2021), 5205–5253. https://doi.org/10.1007/s10462-021-10018-y
[13] A.C. Wook and R.S. Ramakrishna. 2003. Elitism-based compact genetic algorithms. *IEEE Transactions on Evolutionary Computation* 7, 4 (2003), 367–385.

Załącznik 7 – [A7]

SoftwareX 23 (2023) 101476

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software Publication

OXI: An online tool for visualization and annotation of satellite time series data

Bogdan Ruszczak [a,b,1,*], Krzysztof Kotowski [b,1], Jacek Andrzejewski [b], Christoph Haskamp [c], Jakub Nalepa [b,d]

[a] Department of Informatics, Faculty of Electrical Engineering, Automatic Control and Informatics, Opole University of Technology, Prószkowska 76, 45-316 Opole, Poland
[b] KP Labs, Konarskiego 18C, 44-100 Gliwice, Poland
[c] Airbus Defence and Space GmbH, Airbus-Allee 1, 28199 Bremen, Germany
[d] Department of Algorithmics and Software, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

ARTICLE INFO

Article history:
Received 14 May 2023
Received in revised form 28 June 2023
Accepted 11 July 2023

Dataset link: https://oxi.kplabs.pl/static/files/sample_trainset.csv

Keywords:
Annotation tool
Annotations
Time series
Ground truth
Satellite telemetry

ABSTRACT

Satellite telemetry data is a special case of multivariate time series characterized by large volumes (in terms of both the number of series and samples), varying sampling rates (including time gaps), redundant sensors, and many interconnections between the series. Special tools are needed to handle visualization, analysis, and, most importantly, annotation of such data. Manually prepared annotations are crucial when designing and evaluating algorithms for satellite telemetry analysis, including anomaly detection in telemetry data. Although there are many applications for time series analysis, there are no tools that would smoothly handle typical satellite telemetry and at the same time provide a free user-friendly interface accessible from any computer without the need for installation, registration, data sharing, or special training. As a solution to this technology gap, we propose the web-based OXI tool written in JavaScript and publicly available at https://oxi.kplabs.pl/. Based on the TRAINSET application (https://trainset.geocene.com/), we optimized each part of the interface to handle real satellite telemetry data from European Space Agency (ESA) missions. The most important improvements include the redesigned data loading and visualization to effectively handle large datasets (up to 2,000,000 samples), different annotation modes (1D/2D and single/multi-series), autosaving annotation status in the browser storage, and dozens of user experience enhancements (autoscaling of axes, color-coded selectors, additional hotkeys, and many others). The tool was designed and evaluated in cooperation with domain experts from Airbus and ESA. Finally, we provide an extensive description of the software, design choices, and their impact on creating and evaluating machine learning models for anomaly detection in satellite telemetry.

© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Code metadata

Current code version	*v1.0.0*
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-23-00311
Permanent link to reproducible capsule	*Not applicable. The software is not an algorithm but a web-based tool accessible online at https://oxi.kplabs.pl*
Legal code license	*MIT*
Code versioning system used	*git*
Software code languages, tools and services used	*JavaScript*
Compilation requirements, operating environments and dependencies	*Any modern Internet browser, Node.js, Vue.js, jQuery, D3.js, Bootstrap, Papa Parse, Luxon.*
If available, link to developer documentation/manual	https://oxi.kplabs.pl/help
Support email for questions	kkotowski@kplabs.pl, jandrzejewski@kplabs.pl, info@kplabs.pl

* Corresponding author at: Department of Informatics, Faculty of Electrical Engineering, Automatic Control and Informatics, Opole University of Technology, Prószkowska 76, 45-316 Opole, Poland.
E-mail addresses: b.ruszczak@po.edu.pl (Bogdan Ruszczak), kkotowski@kplabs.pl (Krzysztof Kotowski), jandrzejewski@kplabs.pl (Jacek Andrzejewski), christoph.haskamp@airbus.com (Christoph Haskamp), jnalepa@ieee.org (Jakub Nalepa).
[1] Joint first authors

https://doi.org/10.1016/j.softx.2023.101476
2352-7110/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Motivation and significance

Anomaly detection in time series is an active area of research in which multiple classic and machine learning algorithms were proposed [1]. Satellite telemetry data is a special case of multivariate time series characterized by high dimensionality (up to thousands of channels per satellite), missing data (idle states and varying sampling rates during a mission), and – most importantly – limited information about known anomalies. Detecting such abnormal events in telemetry data is of paramount practical importance in spacecraft operations, and it can help respond to anomalies quicker. This, in turn, can ultimately lead to ensuring the safe operation of a satellite, and recovering from abnormal events or gracefully responding to them [2].

According to the latest taxonomy of anomaly types in multivariate time series [1], anomalies can be categorized by three different aspects. The first aspect is the number of consecutive points affected by the anomaly, i.e., a single point, subsequence, or a whole time series. The second aspect is the range of values representing the anomaly, whether they are inside (local) or outside (global) the nominal range of values. The last aspect is the number of series (channels) affected by the anomaly, i.e., univariate or multivariate. It requires extensive expert knowledge and proper tools to manually identify these different types of anomalies in a specific case, i.e., a satellite mission. To automate the process of anomaly detection from telemetry data, several research venues have been explored in the literature so far. The out-of-limit detection engines, in which we verify if the actual telemetry values fall into pre-defined ranges, can easily spot global anomalies, and various data-driven expert systems have been proposed to cover other events [3]. Although there are machine learning algorithms for detecting anomalies from telemetry data (encompassing both supervised and unsupervised techniques [4]), they are commonly heavily parameterized, and the incorrectly tuned hyperparameters can significantly deteriorate their performance [3]. In recent years, we have been witnessing the unprecedented success of deep learning in virtually all scientific and industrial fields, with anomaly detection from sequential data not being an exception here [5]. A class of artificial neural networks, known as recurrent neural networks is particularly suitable for forward prediction based on time series input — in deep learning-powered anomaly detection systems, the predicted signal is commonly confronted with the actual telemetry captured on board spacecraft, and if the difference between those two signals is sufficiently large, the anomaly alarm is triggered [4,6]. Such methods have been validated over various real-life datasets, including the NASA benchmark spacecraft data and the hydrogen clock data of the Beidou Navigation Satellite [7]. It is of note, however, that classic machine learning models may outperform deep learning algorithms for specific types of anomalies [8]. Therefore, the process of selecting an appropriate technique for anomaly detection is not trivial and should be based on objective criteria, such as the effectiveness, efficiency, and robustness of the algorithms [9], quantified over the available benchmark time series datasets [10]. Classic machine learning methods have been effectively deployed on board the OPS-SAT satellite operated by ESA, further proving the feasibility of on-board anomaly detection in such an extreme and hardware-constrained environment [11].

The rapid increase in the number of classic and deep machine learning supervised learners for anomaly detection from satellite telemetry data results in a similarly fast-growing data collection demand, and the main concern during this process is to ensure a reproducible data labeling process, that allows to keep the high ground-truth (GT) quality level and to reduce the time and effort required to generate GT for massively large (potentially multi-channel) time series data. In telemetry data, annotating GT events is directly concerned with the beginning and the end of such anomalous events within time series, effectively leading to the creation of labels for each data point (either nominal or anomalous). Although there exist software tools that can be utilized for annotating such data (Table 1), they suffer from numerous important shortcomings, hampering their real-world adoption. Those disadvantages span across the high complexity of such tools, also related to the complex annotation process which induces long training times for annotators, their inability to deal with large and multi-channel time series data (and to deal with multiple GT labels), and the necessity of deploying them within large software ecosystems.

In this work, we introduce OXI — a highly-accessible and easy-to-use web-based tool for efficient and seamless exploratory analysis and annotation of satellite telemetry (and other similar multi-channel time series data). OXI allows for manual annotation of single- and multi-channel satellite telemetry and effectively handles large data files which are commonly acquired on board operating satellites. It does not require any additional installation or extensive training while handling large-scale satellite telemetry signals captured by real-life space missions currently operating in orbit. OXI is based on the open-source TRAINSET application by Geocene Inc. from which it inherits a couple of features desirable for wider adoption of the tool in practice:

(a) **Ease of use and accessibility** — no additional training is required to operate OXI; no registration is required to use OXI,
(b) **Usability and multi-platform use** — OXI is a web-based application, hence it is multiplatform, it does not require any additional database to store the data, and it provides a decent list of useful keyboard shortcuts; finally, it handles the most popular and widely-established text file format (CSV),
(c) **Security** — the tool operates on a client side, hence does not require sending any user (potentially sensitive) data over the Internet,
(d) **Open license** — OXI is an MIT-licensed software tool,
(e) **Generalizability** — although OXI was developed for handling satellite telemetry data, it can be effectively utilized for analyzing any other type of time-series datasets.

Importantly, we addressed the most important shortcomings of TRAINSET, and significantly improved it in terms of performance and user experience. We implemented an array of features for working with long multi-channel time series (such as satellite telemetry) — the most important improvements include better parsing of files and active downsampling to avoid out-of-memory errors, multi-channel annotation mode, different autoscaling options, and autosaving. The detailed list of added features is given in Section 2.2.

We believe that OXI can become an indispensable tool for analyzing satellite telemetry data, as it directly addresses the immediate needs of the space operations community. In this work, we bundled a unique team of professionals from various fields (spacecraft operations, artificial intelligence, data analytics, and software engineering) to design, implement and deploy in practice a tool that can accelerate the process of gathering GT satellite telemetry data. This will, in turn, help speed up the adoption process of classic and deep machine learning supervised models for anomaly detection from such data, as they require high-quality GT not only for training but also for validation and exploratory analysis of the results elaborated using such algorithms.

Table 1
A list of related tools for time series analysis and annotation [12–17].

Tool name	Reference and link to the source code (if available)	Online	Year of the latest release	Application domain	Multi-channel	Multiple labels	Comments
TRAINSET	https://github.com/Geocene/trainset	YES https://trainset.geocene.com/	2022	Clean cooking	YES	YES	It is a simple web-based tool for time series annotation that we extended to satellite telemetry.
MTV	[12] https://github.com/sintel-dev/MTV	NO	2022	Satellite telemetry	YES	YES	Web-based tool for collaborative satellite telemetry annotation. High complexity. Data is sent over the Internet.
Visplore	https://visplore.com/ (commercial tool)	NO	2022	Universal	YES	YES	Available only for Windows. High complexity. The free version limits data size to 5 attributes and 5000 data records.
TagAnomaly	https://github.com/Microsoft/TagAnomaly	NO	2021	Universal	YES	NO	Very slow and tedious process of labeling long series [13].
ANNO	[13]	NO	2020	NILM*	YES	YES	Only single-point annotations (not time ranges).
WDK	[14] https://github.com/avenix/WDK	NO	2020	Wearables	YES	YES	A part of a larger ecosystem for analysis of signals from wearable devices coupled with videos.
WebTCAD	[15]	NO	2018	Satellite telemetry	YES	NO	A tool mainly for visualization purposes.
feel	[16]	Not active aveiro.m-iti/feel	2018	NILM*	NO	NO	Offers simple semi-automatic anomaly detection.
Curve	https://github.com/baidu/Curve	NO	2018	Universal	NO	NO	Only single-point annotations (not time ranges). Allows for searching for similar anomalies.
time-series-annotator	https://github.com/CrowdCurio/time-series-annotator	NO	2018	EEG*	YES	NO	Not straightforward to use. Needs to be integrated into a separate web application.
CAFED	[17] https://github.com/caoh/CAFED (just data and description, no source code)	Not active https://cafed.inf.ethz.ch/	2015	NILM*	NO	NO	A collaborative framework for annotating time series. Handles fragments of a 1-day span.

*NILM – non-intrusive load monitoring, EEG- electroencephalography

2. Software description

OXI is a tool for the analysis and annotation of multivariate time series data. The tool was designed specifically for satellite telemetry data but can be used with any similar (or less demanding) time series. It offers an intuitive, efficient, and easily accessible interface for creating and evaluating annotated datasets, e.g., for benchmarking algorithms for anomaly detection and classification, or time series segmentation.

OXI is an open-source web-based application freely accessible at https://oxi.kplabs.pl/. No installation, registration, or special training is needed to use it. Any updates to the software are immediately accessible to all users. The only requirement is a PC with any modern web browser installed. We tested the application using Google Chrome (v109), Mozilla Firefox (v109), Microsoft Edge (v109), and Vivaldi (v5.6). OXI can be also easily set up locally by downloading the code from GitHub https://github.com/kplabs-pl/OXI and building it using npm (Node.js package manager) according to the instructions in the README file.

2.1. Software architecture

OXI is built on the foundations of the TRAINSET application by Geocene® (https://trainset.geocene.com/, GitHub: https://github.com/Geocene/trainset). We redesigned and/or optimized each part of the TRAINSET application to handle real satellite telemetry data and to improve the user experience as assessed by experts from Airbus, KP Labs, and ESA.

OXI is written in JavaScript as a client-side application. It means that all data processing is done locally on the user's PC, any annotation progress is saved in the local browser storage, and no data is sent over the Internet to any external servers, this is often crucial in terms of data privacy in satellite missions.

There are only three main views of the application:

- **Starting index view** (a main page) with a description of the software and a button for loading or dropping files to import.
- **Labeler view** is the main view of the application which implements all annotation functionalities. The Labeler view is displayed right after loading the file in the Starting index view.
- **Help view** with detailed instructions for users accessible after clicking Help in the top navigation bar.

The Starting index view (Index.vue) has a simple backend that loads CSV files and initializes the Labeler view with them. The backend of Labeler view is contained in a single JavaScript file (LabelerD3.js) which initializes the plots and implements the main functionality. The Labeler view has a simple event-driven architecture that responds to user actions in the first-in-first-out (FIFO) manner. Hence, instead of the diagram of the software architecture, we present a use case diagram in Fig. 1. It contains four common user actions in the following order: uploading a CSV file (1.), selecting a reference series and zooming in (2.), modifying annotations (3.), and exporting data to a file (4.). Each action triggers a series of events represented by high-level function names from the OXI code. The events can be divided into three different kinds: load, visualize, and annotate. The code of all load events is included in the Index.vue, the latter two types are a part of LabelerD3.js.

OXI uses Vue.js (https://vuejs.org/) and D3.js (https://d3js.org/) as frontend frameworks. Vue.js is used to build a skeleton of the application (i.e., views, buttons, selectors, and navigation bars) and to define communication between the components (i.e., sending data from the Starting index view to the Labeler view, or triggering changes in the main plot after user actions).

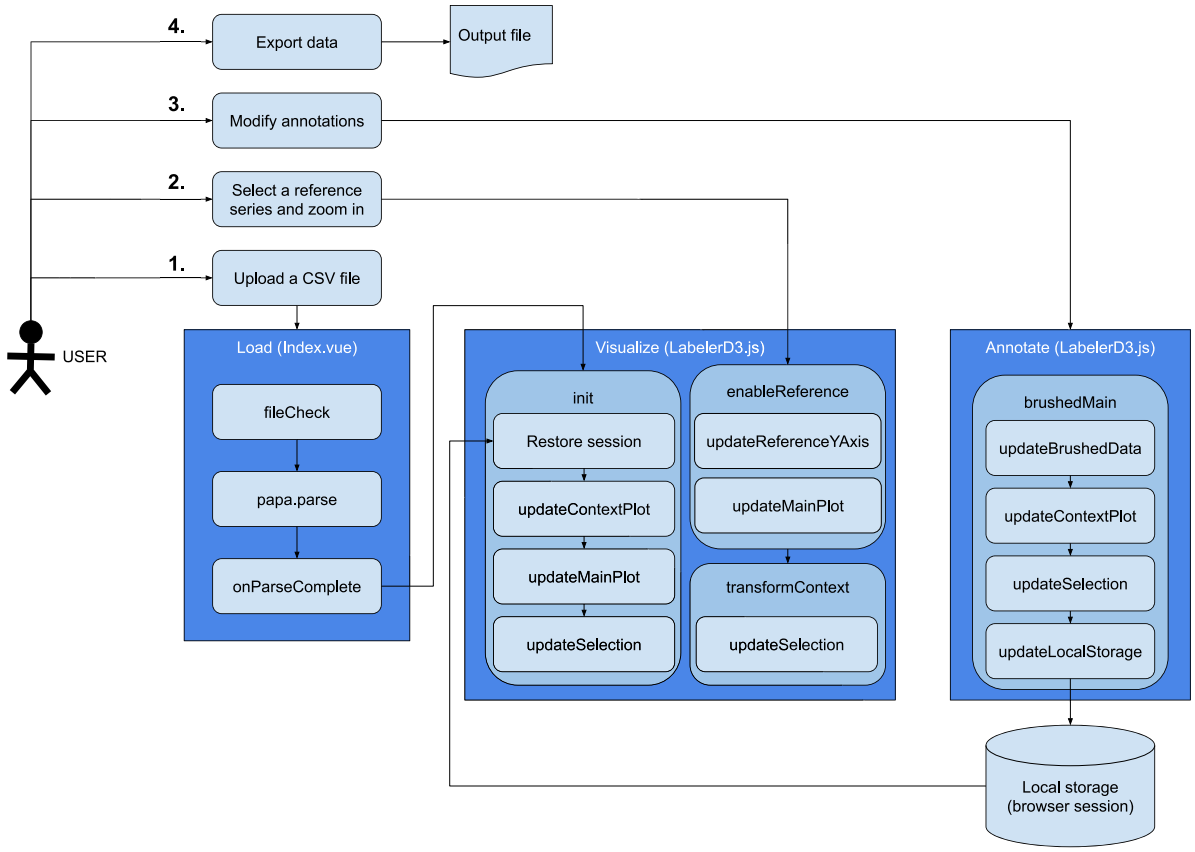


Fig. 1. The use case diagram of the OXI application.

D3.js is used to render plots, select points to annotate, and handle the key shortcuts. Additionally, the Bootstrap toolkit (https://getbootstrap.com/) is used as a main CSS library.

2.2. Software functionalities

The main functionality of the software is to visualize multivariate time series (especially satellite telemetry) and to provide easy-to-use functions for its analysis and annotation. The full list of major functions is given below, the underlined fragments represent new functions in relation to TRAINSET:

- Files import/export
  - Loading files up to 2,000,000 rows
  - Data format supporting varying sampling frequencies between channels
  - Adding a timestamp of export to the file name
  - Drag-n-drop option when importing files
- Visualization:
  - Visualizing two series at once (active and reference)
  - Main plot for annotation and context plot for selecting time ranges
  - Handling very large datasets (up to 2,000,000 rows) without out-of-memory errors
  - Interpolation method adjusted for satellite telemetry data (zero-order hold)
  - Ensuring that all labeled fragments of the series are always visible in the plots despite downsampling
- Annotation:
  - Two annotation modes: single-series and multi-series
  - Two annotation brushes: 1D (in time only) and 2D (in time and values)
  - Multiple annotation labels
  - Autosaving the status of annotations in the local browser memory
- Analysis and user experience:
  - Hotkeys for manipulating plots, switching series, and enabling a reference plot
  - Sorted and color-coded selectors
  - Three autoscaling options: local, global, active/reference split
  - Zooming using the mouse wheel
  - User interface optimized based on the feedback from mission control experts

The functions are described in detail in the following subsections.

2.2.1. Files import/export

OXI imports and exports files in a predefined CSV format containing 4 columns with a header (series, timestamp, value, label) and a maximum of 2,000,000 rows (around 3 days of satellite telemetry data with 80 channels and 0.1 Hz sampling rate). Each row represents one sample in the dataset. We tested that larger data may cause an out-of-memory error on some browsers, so we warn the user and limit the number of rows when reading too large files.

The first *series* column defines the name of the series to which the data sample belongs. The header and values of this column can be any non-empty string because the term “series” can be replaced with “channel”, “parameter”, “sequence”, “stream”, etc. — depending on the domain, but all whitespaces from both

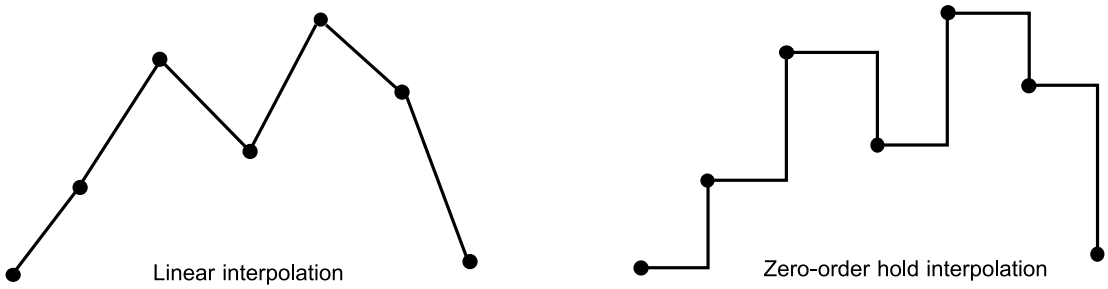


Fig. 2. The difference between linear (left) and zero-order hold interpolation (right).

ends will be trimmed during loading. The second *timestamp* column defines the timestamp of the sample in ISO8601 string format [18] supported by the Luxon library (https://moment.github.io/luxon). The *value* column is parsed as a floating point number and represents the sample value. The last *label* column defines the name of the label assigned to the sample, it should be empty if the sample is not labeled. It is of note that this data format is not optimal in terms of disk memory usage when all the series have the same sampling frequency and timestamps. However, it enables handling a varying sampling frequency between series which is a common situation in satellite telemetry.

Unlike TRAINSET which uses a standard JavaScript *FileReader* object, we use the Papa Parse library (https://www.papaparse.com/) for parsing CSV files. We chose Papa Parse because of its very fast multi-threaded processing, especially with disabled dynamic typing and enabled *fastMode* (if the CSV file does not contain quotation marks). Together with our simplified data validation, it allowed us to import data around 15% faster (7.5 s vs 8.7 s as measured for a file with 1.5 million rows). Files exported from OXI have the same format and length as the input files, only the *label* column may change in the process of annotation. The default output file name is created by adding a suffix with an export timestamp to the input file name.

2.2.2. Visualization

The interface was intentionally designed to be as simple as possible and focused on the annotation functionality, so it is maximally intuitive and does not require any specific training. This is because we found applications like MTV [12] or Visplore (https://visplore.com/) quite complex to set up and use (without proper training) even by experts in the domain.

The most important elements of the Labeler view are the main plot and the context plot (see Fig. 7). Two different series can be visualized at once in the main plot: an active series (a series being annotated in the single-series annotation mode) and a reference series (visualized as a reference for the active series, using smaller markers and gray color), each with its own Y axis. We decided that visualizing more than two series at once may be too confusing for users and may cause too long rendering times. The context plot below the main plot shows the whole time range of an active series and is used to select the time window of interest for the main plot. A step chart with zero-hold interpolation (keeping the last known value) between points (a step-after mode in D3.js) was applied as suggested by the authors of the MTV tool for satellite telemetry [12]. Our experts from Airbus and ESA confirmed that this interpolation method helps them to track the time gaps in the data. Additionally, this interpolation method is well suited for visualizing binary and quantized signals that are common in satellite telemetry (i.e., telecommands and measurements from analog-to-digital converters) because, unlike the linear interpolation in TRAINSET, it does not create any artificial, impermissible values between points, as presented in Fig. 2.

To prevent long rendering times and out-of-memory errors, we limited the number of samples per series in the main plot to 4320 (12 h of data for 0.1 Hz sampling frequency). This specific number was selected based on the qualitative assessment of the maximal number of points that can be distinguished from each other in the plot on the screen of 1080p resolution. The number of samples is limited using the Largest Triangle Three Buckets (LTTB) downsampling algorithm [19] implemented as a D3.js component in the D3FC library (https://d3fc.io/). This algorithm aims to select the best subset of points which keeps the visual representation of the original series. In short, the algorithm maximizes areas of triangles formed by consecutive points in the downsampled series. The same downsampling is applied in TRAINSET to limit the number of points in the context plot to 1000. This algorithm has two important features in the context of satellite telemetry visualization: (1) just like zero-order hold interpolation, it does not create any artificial or impermissible values in quantized signals; (2) it focuses on areas of high variability (i.e., anomalies) and removes redundant points from signal plateaus. We measured the impact of such downsampling on the rendering time (the time between mouse scroll and displaying the downsampled plot) of two series (active and reference) when increasing the number of points in both series. As presented in Fig. 3, there is a linear relationship between the number of points and rendering times when no limit is applied, up to the point where the browser crashes due to the out-of-memory error. Thus, limiting the number of samples helps avoid crashes and notably reduces the slope of the linear relationship for the numbers of samples above the limit.

When limiting the number of points, there may be situations where short annotated fragments are removed from the downsampled plot. This is a very unwanted behavior, especially in terms of the context plot which may incorrectly suggest a lack of anomalies in a previewed series. To tackle this problem, we proposed a procedure for restoring representative points for annotated fragments after downsampling. The procedure is presented in Code snippet 1. It iterates over all annotated fragments from the original data and successively adds representative points, i.e., middle points, to the downsampled data array if necessary.

OXI was implemented using responsive web design, that is, the content of the website dynamically scales to fit a browser window size. However, the height of the main plot is fixed to 500 pixels (in comparison to 390 pixels in TRAINSET), so for the best experience, we recommend using at least 1080p resolution. We do not recommend using the application on mobile devices.

2.2.3. Annotation

A user can annotate samples by brushing over the main plot. The default annotation brush (called 1D from 1-dimensional brushing) allows for selecting all the samples in the time range between two timestamps. The second annotation brush (called 2D from 2-dimensional brushing) allows for selecting samples in a specific rectangular region of interest across timestamps and

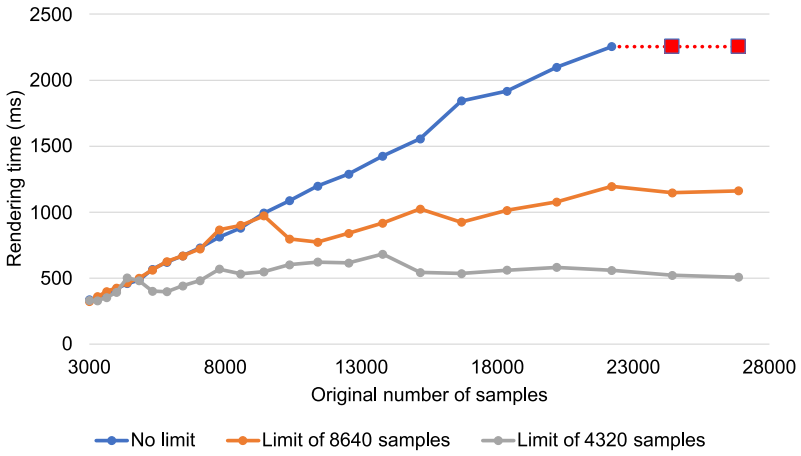


Fig. 3. The impact of limiting the number of samples in the visualization on the rendering time of the main plot. The red squares represent out-of-memory errors of the browser.

series values. It is especially useful when selecting multiple outlying samples (peaks) separated by nominal samples. However, it may lead to accidental time gaps in annotations (e.g., when some samples are outside the field of view because of axes scaling), so the 1D brush was introduced as the default mode. To help users define the interval (1D) or area (2D) to be annotated, the current cursor position in the plot is displayed in the "Sample details" widget (see Fig. 7). The widget shows the exact time below the cursor and, depending on the annotation mode, the exact value of the active Y axis below the cursor (2D) or the value and label of the closest sample (in terms of time) of the active series (1D). Annotations can be removed by selecting the samples with the Shift key pressed. Single points can be annotated (and unannotated) by clicking on them. The samples are annotated with the active label selected using the label selector (see Fig. 7). Importantly, to simplify the visualization, file format, and analysis, each sample can have only one label assigned. It is of note that introducing the possibility of assigning multiple class labels to each sample would have a significant impact on the user experience and complexity of OXI.

There are two main annotation modes: single-series (the default mode in which only the active series is annotated) and multi-series. The multi-series annotation mode is especially useful when working with satellite telemetry because there are usually groups of interconnected channels, e.g., currents from a set of solar panels, temperatures of components in the same subsystem, or propulsion engine on-times with attitude and orbit control measurements. If an anomaly appears in one such channel it usually appears also in all interconnected channels in the same time window. It would be very impractical to annotate each channel separately in such a case. We implemented a multi-series selector using the Vue-multiselect component (https://vue-multiselect.js.org/) as presented in Fig. 4. In the multi-series mode, all series marked in the selector will be annotated together in the same time window (even if they are not visible in the main plot). We decided that brushing in 2D in the multi-series mode would be confusing and impractical, so only the 1D brush is available in that mode.

A current annotation status is autosaved in the local browser storage after each change to prevent losing annotation progress in case of out-of-memory errors or accidentally closing the browser tab. Before saving, a list of labels for consecutive samples in each series is compressed using run-length encoding [20] to reduce the memory footprint. It is a very fast and effective method if there are relatively few continuous annotated fragments in the data, which is the case for anomalies in satellite telemetry. The annotation status is automatically restored when the user loads the file of the same name.

As there are a few computationally-demanding processes run after each annotation (i.e., modifying the data, autosaving, downsampling, and restoring missing annotations in the context plot processes), we analyzed the latency between the user input and the response from the tool. The measured latencies for single-point and whole-series annotations are presented in Fig. 5. There is a linear relationship between the series length and the annotation latency which is in line with the fact that all annotation-related processes have a complexity of $O(n)$, where n is the number of samples (modifying, autosaving, downsampling) in the series or the number of annotated samples (restoring missing annotations). For short series (<100k samples), the annotation latencies are less than 200 ms. For the longest possible series of 2,000,000 samples, the annotation process takes 2.5 s.

2.2.4. Analysis and user experience

After each development iteration, we consulted the interface with system engineers from KP Labs and/or domain experts from Airbus and ESA. It resulted in a series of improvements over TRAINSET to make the manual analysis of satellite telemetry more user-friendly and effective. A list of all new functions is presented in Table 2.

Additional hotkeys, autoscaling options, and coloring of selectors were assessed as the most effective changes that drastically decreased the time of manual analysis. Hotkeys help to smoothly iterate over consecutive series without using selectors and turning the reference series on and off when comparing it to the active series. Local and global autoscaling options modify the range of the Y axis to the range of values in the selected fragments (local) or in the whole series (global). Global scaling is a default option when initializing the application, but it is not the best scaling when strong outliers are present in the data (which is common to satellite telemetry data). Additionally, when the reference series is displayed, it is possible to split the main plot into top and bottom halves for active and reference series respectively (see the example in Fig. 7). It is especially useful when comparing two series with very different value ranges which is a common situation in satellite telemetry. Finally, the color-coding of selectors is very useful when searching for series containing samples annotated with a specific active label. Such series have names colored using the color of an active label as shown in Fig. 6.

The numbers on axes are displayed using scientific notation limited to two decimal places, i.e., 2.12e−2, so they always have similar number of characters for all series. It helps to avoid

```
function fillMissingAnnotationsAfterDownsampling(downsampled_data,
all_labels) {
    if (all_labels.length === 0 || downsampled_data.length === 0) {
      return downsampled_data;
    }

    let points_to_fill = [];
    let current_data_iter = 0;
    let current_data_label = downsampled_data[0].label;
    let current_data_timestamp = downsampled_data[0].time;

    for (let i = 0; i < all_labels.length; i++) {
      let labels_timestamp = all_labels[i].time;
      if (labels_timestamp < current_data_timestamp) {
        continue
      }

      for (let j = current_data_iter; j < downsampled_data.length; j++){
        current_data_timestamp = downsampled_data[j].time;
        if (current_data_timestamp < labels_timestamp) {
          current_data_label = downsampled_data[j].label
          continue
        }

        // add a representative point for a specific label if missing
        if (current_data_label !== all_labels[i].label &&
downsampled_data[j].label !== all_labels[i].label) {
          let series_of_labels = [];
          for (let k = i; k < all_labels.length; k++) {
            if (all_labels[k].time >= current_data_timestamp) {
              break
            }
            series_of_labels.push(all_labels[k])
          }

points_to_fill.push(series_of_labels[Math.floor(series_of_labels.length /
2)])
        }

        current_data_iter = j;
        current_data_label = downsampled_data[j].label
        break
      }
    }

    if (points_to_fill.length > 0) {
      downsampled_data = downsampled_data.concat(points_to_fill).sort((a, b)
=> a.time - b.time);
    }
    return downsampled_data;
  }
```

Code snippet 1. A function for filling downsampled data with annotated regions, so they never disappear from the visualizations.

problems with axes positioning and visualization for very small and very large numbers. The disadvantage is that sometimes two decimal places are not enough to properly analyze series with very small variations, however, the value with 5 decimal places (together with exact timestamp and label) of any sample can be inspected in the upper right corner of the main plot after moving the cursor over that sample (see example in Fig. 7).

3. Illustrative examples

In Fig. 7, we provide a detailed description of the Labeler view and all functions provided by its interface. In this particular example, the active series_a and the reference series_b in the main plot are split vertically using active/reference split autoscaling function (button 3 in Fig. 7). The scaling of axes can be changed separately for active and reference series using buttons on the left and right side of the main plot (elements 1 and 2 in Fig. 7, respectively). The black points in the main plot are samples without any label assigned and violet, green, and orange points represent samples annotated with 3 different labels, respectively. The user can annotate (or unannotate) the samples in the active series by brushing over them. As visible in the label selector (element 4 in Fig. 7), the active label is orange, and additional annotation options are disabled (elements 5 and 6), so the brush is set up in 1D annotation mode for the active series only. The label selector allows for adding (“+”

Table 2
The list of additional functions and improvements over TRAINSET targeted for more effective analysis of satellite telemetry.

Function or improvement	Purpose
Hotkeys for iterating over active/reference series and turning the reference series on/off	More effective comparisons between series
Sorting and color-coding of label and series selectors	Easier search and faster identification of series annotated with a selected label
Additional autoscaling options for axes: local scaling, global scaling, split between active and reference series	Easier comparisons between series and more effective analysis of series with outliers
Consistent formatting of numbers on axes and fixed position of axes and scaling buttons	Better user experience
Zooming using the mouse scroll	More intuitive analysis of series
Updating the main plot during the time window selection in the context plot	Better user experience
A larger area of the main plot	Larger working area and more detailed analysis. Better user experience

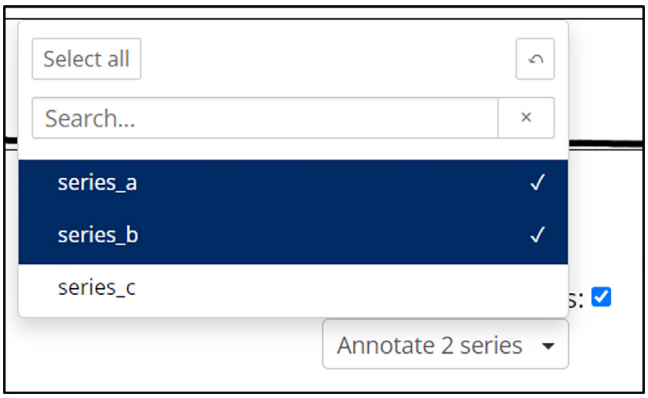


Fig. 4. The selector of multiple series for annotation.

button) and removing specific labels ("x" button). The main plot includes only samples from the time window of interest selected using the context plot. The context plot is placed just below the main plot and it visualizes the whole active series with the gray area representing the current time window of interest. The time window of interest can be manipulated by dragging or resizing the gray area, or by using mouse scroll.

The instructions for the user are placed as static text in the bottom left corner of the Labeler view. In the bottom right corner, there are active and reference series selectors. Each selector is a drop-down list of all series in the dataset which can be selected by clicking on them or by using the hotkeys mentioned in the instructions. There are two options to turn off the reference series in the main plot: using the tick box (element 9 in Fig. 7) or setting the reference series to be the same as the active series. Finally, by using the global functions in the upper right corner of the website, the user can display the Help view, export, or clear the annotations.

4. Impact

OXI fills the technological gap in the process of multivariate time series visualization and annotation. It is a handy tool for anyone working with satellite telemetry (or similar data). It can be used to prepare and analyze datasets for different research projects connected with time series analysis, including multivariate time series anomaly detection which is an increasingly popular task in the scientific literature [21]. In recent years, the space operation community is actively working on the topic, including teams from NASA [4], CNES [22], DLR [23], JAXA [24], EUMETSAT and Airbus among others [2,25–28].

OXI was created and is currently used in the ongoing project "ESA Anomalies Dataset for International AI Anomaly Detection

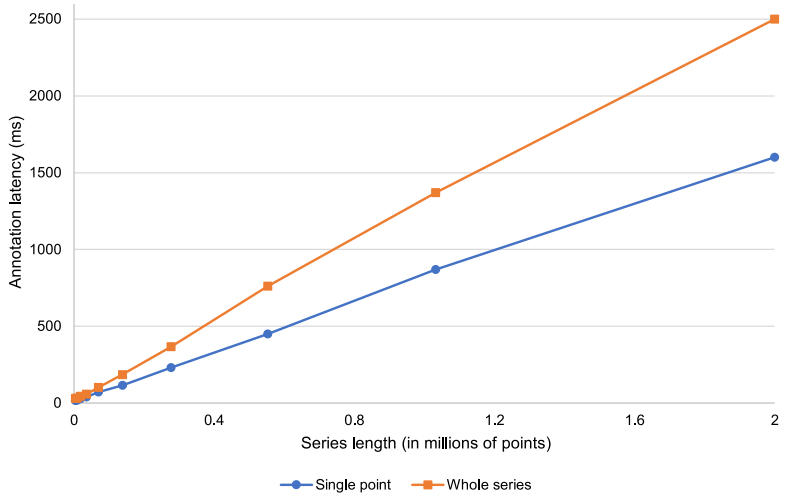


Fig. 5. The impact of the active series length on the latency of single-point and whole-series annotations.

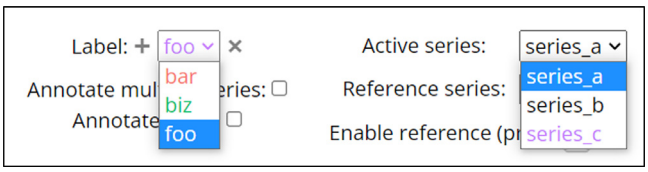


Fig. 6. The example of color-coding of label and series selectors.

Benchmark" funded by ESA and realized by a consortium of Airbus Defence and Space and KP Labs companies. The main purpose of this project is to prepare an extensive public dataset for anomaly detection in real-life satellite telemetry which could generate an impact in this domain similar to the impact of ImageNet [29] in the domain of image classification. OXI is used in the project for manual annotations and verification of the results from anomaly detection algorithms. The crucial feature of OXI in the project is the ease of access to the tool by anyone in the consortium (software developers, machine learning specialists, system engineers, and mission control experts) from any computer, without any special installation or training. It helps to inspect and discuss interesting fragments remotely. OXI will be a recommended visualization tool for the community when the dataset is released publicly. This should effectively spread the information and boost the popularity of the tool among users and various research and industrial communities. In the meantime, we plan to promote the tool among students, clients and via the KP Labs social media.

OXI was first published online (https://oxi.kplabs.pl/) on September 2022. Recently, it was used to prepare datasets for anomaly detection experiments in the OPS-SAT satellite telemetry [11] as a part of ESA-funded project "On-Board Anomaly Detection From The OPS-SAT Telemetry Using Deep Learning" [30]. It is also used internally by system engineers in KP Labs to annotate different satellite telemetry datasets, e.g., from the PW-Sat2 satellite. Outside the space sector, the QZ Solutions company uses OXI to analyze time series data from smart farming machinery.

5. Conclusions

Detecting anomalies in satellite telemetry data plays an increasingly important role in space operations, as spotting abnormal events quickly can allow for undertaking appropriate actions to ensure its safety. Although supervised machine learning approaches have been blooming at a steady pace in the field, collecting heterogeneous and high-quality GT datasets that could be used for training large-capacity models is still a challenging task, especially given that the manual annotation process is time-consuming, hard to reproduce, cumbersome and requires specialized tools for handling single- and multi-channel signal data. In this work, we tackled this issue and introduced OXI — an online tool which is specifically designed for the analysis and annotation of satellite telemetry data, while still offering a high level of generalizability for other time series. We presented a wide array of functionalities of OXI which have been already proven in various ESA-funded projects and missions, and it will be a recommended tool in the emerging ones. We believe that OXI can become a state-of-the-art tool for analyzing satellite telemetry due to its usability, robustness, ease of use, and portability, which effectively can accelerate its widespread adoption.

Due to its simplicity, OXI has numerous limitations as well. It does not offer any algorithms for automated analysis or annotation, and it does not resolve the issue of reproducible labeling. Also, it does not offer any advanced collaborative features. The data size is limited to 2,000,000 rows and the latency of annotation increases proportionally to the series length. Thanks to the simple architecture and clean implementation of OXI, the additional functionalities and extensions can be relatively easily added as part of future work. The most important ones span across the undo function, advanced data analysis and exploration settings (including various interpolation methods, and controlling the limit of samples in the exploratory analysis), handling different file formats and structures, adding uncertainty annotations, allowing for user comments, and providing multiple labels

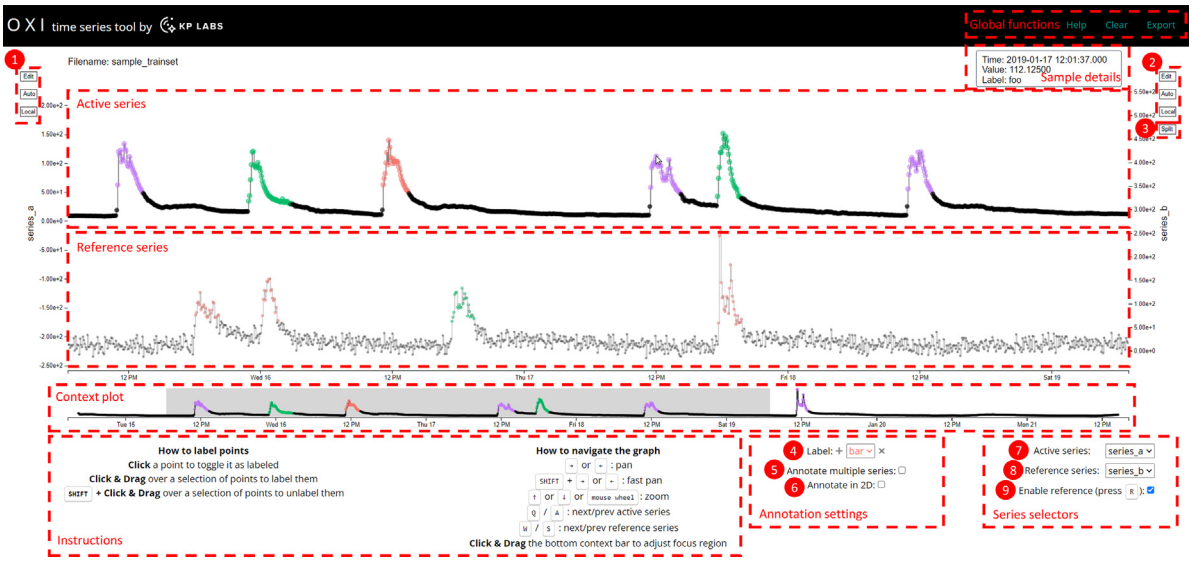


Fig. 7. The description of the main parts of the Labeler view. 1 — Scaling buttons of the active series axis. 2 — Scaling buttons of the reference series axis. 3 — Button for splitting series vertically. 4 — Label selector with options to add ("+") and remove labels ("x"). 5 — Switch between single (default) and multiple series annotation modes. 6 — Switch between 1D (default) and 2D annotation modes. 7 — Active series selector. 8 — Reference series selector. 9 — Switch for turning reference series on/off.

per sample. Finally, it would be useful to allow for exporting publication-ready screenshots, in order to make the utilization of the software easier for the research community as well. However, any changes should be preceded by a complexity vs. usability analysis to maintain the greatest strengths of OXI [31].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

An example data file is made publicly available as a part of OXI (https://oxi.kplabs.pl/static/files/sample_trainset.csv).

Acknowledgments

The work was financially supported by European Space Agency as a part of the "ESA Anomalies Dataset for International AI Anomaly Detection Benchmark" project (contract number 4000137682/22/D/SR) and "Few-shot anomaly detection in satellite telemetry" project (contract number 4000141301). JN was supported by the Silesian University of Technology funds through the grant for maintaining and developing research potential.

We would like to thank Geocene® team (Rush Kapoor, Ajay Pillarisetti, Jeremy Coyle, Skot Croshere, Marc Paré, Hamza Benkhay, and Danny Wilson) for developing the foundations of our application. We would like to acknowledge the work of all testers of our application and Alicja Musiał (KP Labs) for proposing a set of user experience improvements. Finally, a word of thanks goes to the ESA teams who supported the contracts for their invaluable feedback.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.softx.2023.101476.

References

[1] Blázquez-García A, Conde A, Mori U, Lozano JA. A review on outlier/anomaly detection in time series data. ACM Comput Surv 2021;54(3):56:1–33. http://dx.doi.org/10.1145/3444690.

[2] Nalepa J, Myller M, Andrzejewski J, Benecki P, Piechaczek S, Kostrzewa D. Evaluating algorithms for anomaly detection in satellite telemetry data. Acta Astronaut 2022. http://dx.doi.org/10.1016/j.actaastro.2022.06.026.

[3] Benecki P, Piechaczek S, Kostrzewa D, Nalepa J. Detecting anomalies in spacecraft telemetry using evolutionary thresholding and LSTMs. In: Proceedings of the genetic and evolutionary computation conference companion. New York, NY, USA: ACM; 2021, p. 143–4.

[4] Hundman K, Constantinou V, Laporte C, Colwell I, Soderstrom T. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018, p. 387–95. http://dx.doi.org/10.1145/3219819.3219845.

[5] Wang Y, Gong J, Zhang J, Han X. A deep learning anomaly detection framework for satellite telemetry with fake anomalies. Int J Aerosp Eng 2022;2022:e1676933. http://dx.doi.org/10.1155/2022/1676933.

[6] Xu J, Wu H, Wang J, Long M. Anomaly transformer: Time series anomaly detection with association discrepancy. 2021, http://dx.doi.org/10.48550/arXiv.2110.02642.

[7] Yang L, Ma Y, Zeng F, Peng X, Liu D. Improved deep learning based telemetry data anomaly detection to enhance spacecraft operation reliability. Microelectron Reliab 2021;126:114311. http://dx.doi.org/10.1016/j.microrel.2021.114311.

[8] Rewicki F, Denzler J, Niebling J. Is it worth it? comparing six deep and classical methods for unsupervised anomaly detection in time series. Appl Sci 2023;13(3):3. http://dx.doi.org/10.3390/app13031778.

[9] Schmidl S, Wenig P, Papenbrock T. Anomaly detection in time series: A comprehensive evaluation. Proc VLDB Endow 2022;15(9):1779–97. http://dx.doi.org/10.14778/3538598.3538602, Lipiec.

[10] Jacob V, Song F, Stiegler A, Rad B, Diao Y, Tatbul N. Exathlon: A benchmark for explainable anomaly detection over time series. Proc VLDB Endow 2021;14(11):2613–26. http://dx.doi.org/10.14778/3476249.3476307, Październik.

[11] Nalepa J, et al. Toward on-board detection of anomalous events from opssat telemetry using deep learning. In: Presented at the 8th international workshop on on-board payload data compression. Athens: Zenodo; 2022, p. 7. http://dx.doi.org/10.5281/zenodo.724499.

[12] Liu D, Alnegheimish S, Zytek A, Veeramachaneni K. MTV: Visual analytics for detecting, investigating, and annotating anomalies in multivariate time series. Proc ACM Hum-Comput Interact 2022;6(CSCW1):103:1–30. http://dx.doi.org/10.1145/3512950.

[13] Huchtkoetter J, Reinhardt A, Hossain S. ANNO: A time series annotation tool to evaluate event detection algorithms. In: Gunkelmann N, Baum M, editors. Simulation science. Communications in computer and information science, Cham: Springer International Publishing; 2020, p. 70–87. http://dx.doi.org/10.1007/978-3-030-45718-1_5.

[14] Haladjian J. The Wearables Development Toolkit (WDK) [Computer software]. 2019, [Online]. Available: https://github.com/avenix/WDK. [Accessed: 09 September 2022].

[15] Sanchez F, Pankratz C, Lindholm DM, Christofferson R, Osborne D, Baltzer T. WebTCAD: A tool for ad-hoc visualization and analysis of telemetry data for multiple missions. In: 2018 SpaceOps conference. Marseille, France: American Institute of Aeronautics and Astronautics; 2018, http://dx.doi.org/10.2514/6.2018-2616.

[16] Pereira L, Ribeiro M, Nunes N. Engineering and deploying a hardware and software platform to collect and label non-intrusive load monitoring datasets. In: 2017 sustainable internet and ICT for sustainability. 2017, p. 1–9. http://dx.doi.org/10.23919/SustainIT.2017.8379791.

[17] Cao H-A, Wijaya TK, Aberer K, Nunes N. A collaborative framework for annotating energy datasets. In: 2015 IEEE international conference on big data. 2015, p. 2716–25. http://dx.doi.org/10.1109/BigData.2015.7364072.

[18] International organization for standardization, data elements and interchange formats – Information interchange – Representation of dates and times (ISO 8601:2004). 2004, [Online]. Available: https://www.iso.org/standard/40874.html. [Accessed: 13 March 2023].

[19] Steinarsson S. Downsampling time series for visual representation [Thesis], 2013, [Online]. Available: https://skemman.is/handle/1946/15343. [Accessed: 14 March 2023].

[20] Tsukiyama T, Kondo Y, Kakuse K, Saba S, Ozaki S, Itoh K. Method and system for data compression and restoration, US4586027A, Apr. 29. 1986, [Online]. Available: https://patents.google.com/patent/US4586027A/en. [Accessed: 15 March 2023].

[21] Li G, Jung JJ. Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. Inf Fusion 2023;91:93–102. http://dx.doi.org/10.1016/j.inffus.2022.10.008.

[22] Fuertes S, Pilastre B, D'Escrivan S. Performance assessment of Nostradamus & other machine learning-based telemetry monitoring systems on a spacecraft anomalies database. In: 2018 SpaceOps conference. SpaceOps conferences, American Institute of Aeronautics and Astronautics; 2018, http://dx.doi.org/10.2514/6.2018-2559.

[23] O'Meara C, Schlag L, Wickler M. Applications of deep learning neural networks to satellite telemetry monitoring. In: 2018 SpaceOps conference. SpaceOps conferences, American Institute of Aeronautics and Astronautics; 2018, http://dx.doi.org/10.2514/6.2018-2558.

[24] Yairi T, Takeishi N, Oda T, Nakajima Y, Nishimura N, Takata N. A data-driven health monitoring method for satellite housekeeping data based on probabilistic clustering and dimensionality reduction. IEEE Trans Aerosp Electron Syst 2017;53(3):1384–401. http://dx.doi.org/10.1109/TAES.2017.2671247.

[25] Su Y, Zhao Y, Niu C, Liu R, Sun W, Pei D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '19, New York, NY, USA: Association for Computing Machinery; 2019, p. 2828–37. http://dx.doi.org/10.1145/3292500.3330672.

[26] Tariq others S. Detecting anomalies in space using multivariate convolutional LSTM with mixtures of probabilistic PCA. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '19, New York, NY, USA: Association for Computing Machinery; 2019, p. 2123–33. http://dx.doi.org/10.1145/3292500.3330776.

[27] Zhang C, et al. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the thirty-third AAAI conference on artificial intelligence and thirty-first innovative applications of artificial intelligence conference and ninth aaai symposium on educational advances in artificial intelligence. AAAI'19/IAAI'19/EAAI'19, Honolulu, Hawaii, USA: AAAI Press; 2019, p. 1409–16. http://dx.doi.org/10.1609/aaai.v33i01.33011409.

[28] Pilastre B, Boussouf L, D'Escrivan S, Tourneret J-Y. Anomaly detection in mixed telemetry data using a sparse representation and dictionary learning. Signal Process 2020;168:107320. http://dx.doi.org/10.1016/j.sigpro.2019.107320.

[29] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. 2009, p. 248–55. http://dx.doi.org/10.1109/CVPR.2009.5206848.

[30] Nebula public library, on-board anomaly detection from the ops-sat telemetry using deep learning. 2023, https://nebula.esa.int/content/board-anomaly-detection-ops-sat-telemetry-using-deep-learning. [Accessed 16 March 2023].

[31] Mai R, Hoffmann S, Schwarz U, Niemand T, Seidel J. The shifting range of optimal web site complexity. J Interact Mark 2014;28(2):101–16. http://dx.doi.org/10.1016/j.intmar.2013.10.001.

Załącznik 8 – [A8]

Jiří Mikyška · Clélia de Mulatier ·
Maciej Paszynski ·
Valeria V. Krzhizhanovskaya ·
Jack J. Dongarra · Peter M. A. Sloot (Eds.)

LNCS 14073

Computational Science – ICCS 2023

23rd International Conference
Prague, Czech Republic, July 3–5, 2023
Proceedings, Part I



Springer

Lecture Notes in Computer Science **14073**

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*
Wen Gao, *Peking University, Beijing, China*
Bernhard Steffen, *TU Dortmund University, Dortmund, Germany*
Moti Yung, *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Jiří Mikyška · Clélia de Mulatier ·
Maciej Paszynski · Valeria V. Krzhizhanovskaya ·
Jack J. Dongarra · Peter M. A. Sloot
Editors

Computational Science – ICCS 2023

23rd International Conference
Prague, Czech Republic, July 3–5, 2023
Proceedings, Part I

Springer

Editors
Jiří Mikyška
Czech Technical University in Prague
Prague, Czech Republic

Maciej Paszynski
AGH University of Science and Technology
Krakow, Poland

Jack J. Dongarra
University of Tennessee at Knoxville
Knoxville, TN, USA

Clélia de Mulatier
University of Amsterdam
Amsterdam, The Netherlands

Valeria V. Krzhizhanovskaya
University of Amsterdam
Amsterdam, The Netherlands

Peter M. A. Sloot
University of Amsterdam
Amsterdam, The Netherlands

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-35994-1 ISBN 978-3-031-35995-8 (eBook)
https://doi.org/10.1007/978-3-031-35995-8

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Machine Learning Detects Anomalies in OPS-SAT Telemetry

Bogdan Ruszczak[1,3(✉)], Krzysztof Kotowski[3], Jacek Andrzejewski[3], Alicja Musiał[3], David Evans[4], Vladimir Zelenevskiy[4,5], Sam Bammens[4], Rodrigo Laurinovics[4,6], and Jakub Nalepa[2,3(✉)]

[1] Faculty of Electrical Engineering, Automatic Control and Informatics, Department of Informatics, Opole University of Technology, Prószkowska 76, 45-758 Opole, Poland
b.ruszczak@po.edu.pl
[2] Faculty of Automatic Control, Electronics and Computer Science, Department of Algorithmics and Software, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
jnalepa@ieee.org
[3] KP Labs, Konarskiego 18C, 44-100 Gliwice, Poland
{bruszczak,kkotowski,jandrzejewski,amusial,jnalepa}@kplabs.pl
[4] European Space Agency/ESOC, Robert-Bosch-Str. 5, 64293 Darmstadt, Germany
{David.Evans,Sam.Bammens}@esa.int,
{Vladimir.Zelenevskiy,Rodrigo.Laurinovics}@ext.esa.int
[5] Telespazio Germany GmbH, Europaplatz 5, 64293 Darmstadt, Germany
[6] IrbGS ltd, "Viraki", Ance Parish, Irbene LV-3601, Latvia

Abstract. Detecting anomalies in satellite telemetry data is pivotal in ensuring its safe operations. Although there exist various data-driven techniques for the task of determining abnormal parts of the signal, they are virtually never validated over real telemetries. Analyzing such data is challenging due to its intrinsic characteristics, as telemetry may be noisy and affected by incorrect acquisition, resulting in missing parts of the signal. In this paper, we tackle this issue and propose a machine learning approach for detecting anomalies in single-channel satellite telemetry. To validate its capabilities in a practical scenario, we build a dataset capturing the nominal and anomalous telemetry data captured on board OPS-SAT—a nanosatellite launched and operated by the European Space Agency. Our extensive experimental study showed that the proposed algorithm offers high-quality anomaly detection in real-life satellite telemetry, reaching 98.4% accuracy over the unseen test set.

Keywords: machine learning · anomaly detection · feature engineering · satellite telemetry

This work was partially supported by the: "On-board Anomaly detection from the OPS-SAT telemetry using deep learning" project funded by the European Space Agency under contract No. 4000137339/22/NL/GLC/ov. JN was supported by the Silesian University of Technology Rector's grant (02/080/RGJ22/0026).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
J. Mikyška et al. (Eds.): ICCS 2023, LNCS 14073, pp. 295–306, 2023.
https://doi.org/10.1007/978-3-031-35995-8_21

1 Introduction

Many critical systems are being monitored using various sensors for detecting their malfunction [11,20]. This is extremely important for satellites to ensure their safe operations and to appropriately respond to any abnormal events that are reflected in the telemetry data. In general, there are three types of anomalous events which are commonly observed in satellite telemetry. In point anomalies, telemetry signal values fall outside the nominal operational range, and hence can be easily detected using out-of-limit checks. On the other hand, collective anomalies correspond to the overall sequences of consecutive telemetry values that are abnormal, whereas in contextual anomalies, the single telemetry values are anomalous within their local neighborhood [17]. There exist data-driven algorithms for detecting such types of abnormal events [18], spanning across classic techniques exploiting expert systems [25], unsupervised approaches [8,14] and deep learning models [10] (often benefiting from recurrent neural networks which are especially well-fitted to process time-series data [1,9]). Such algorithms, however, are virtually never validated over real-life telemetry data. Also, they often require long time-series data to build a model reflecting the nominal operation of the satellite—capturing it on board is tedious and time-consuming, thus data-level digital twins have been blooming to simulate the correct telemetry [2]. Benchmark datasets commonly exploited to validate detection algorithms contain time-series data, where each time series is split into its training and test parts, presenting similar characteristics. Such data is not affected by the practical challenges commonly observed in on-board telemetry, such as data noisiness or missing data, due to e.g., inappropriate signal acquisition. Therefore, the estimated anomaly detection capabilities of data-driven techniques may easily become over-optimistic, and the experimental scenarios are often flawed by methodological issues in the field [23].

In this paper, we tackle the issue of thorough validation of machine learning techniques for anomaly detection in satellite telemetry data. We not only propose an end-to-end pipeline for detecting such abnormal events in signal data. We also investigate its capabilities over a dataset capturing real-life telemetry data captured on board OPS-SAT—a nanosatellite launched and operated by the European Space Agency (ESA). OPS-SAT is a *flying laboratory*, providing a platform for on-board experiments (e.g., those which may be too risky to be executed on other operational satellites [7]), with one of them being anomaly detection from telemetry data using the pipeline discussed in this work. Other experiments performed on-board OPS-SAT include, among others:

- on-board unsupervised machine learning for spacecraft autonomy [13,22],
- the compression of housekeeping telemetry [5],
- assessing the stability of the attitude during inertial pointing mode using on-board images of the sky [21],
- deploying and maintaining the MO/MAL ground infrastructure with this aircraft [15],
- testing the on-board thermal vacuum [12],
- and addressing fail-safety and redundancy design challenges [24].

Here, we build upon the available OPS-SAT telemetry data to build a representative dataset that is used to train and validate our anomaly detection techniques in a practical setting, reflecting real on-board telemetry challenges, e.g., missing data and aperiodicity of the signal [6]. Our experimental validation performed over a carefully bundled OPS-SAT telemetry dataset revealed that our approaches exploiting hand-crafted feature extractors (which are independent of the length of the signal and can effectively operate with missing data) and classic supervised learners offer high-quality anomaly detection, reaching 98.4% classification accuracy over the unseen test set.

This paper is structured as follows. In Sect. 2, we discuss the main objective tackled in this paper and present the OPS-SAT telemetry channels of interest, together with our end-to-end machine learning pipeline for detecting abnormal events from time-series data. Our experimental study is reported and discussed in Sect. 3. Section 4 concludes the paper and highlights the most exciting future research pathways which may emerge from our work.

2 Materials and Methods

In this section, we discuss the investigated OPS-SAT telemetry (Sect. 2.1), together with our approach for detecting anomalies in such data (Sect. 2.2).

2.1 Dataset

In this work, we investigate the telemetry signals indicated by the ESA OPS-SAT team as the most interesting from the operational point of view, and we downloaded such data from the WebMUST client telemetry directory available for OPS-SAT [4]. Those telemetry channels include the magnetometer readouts, alongside the photo diode (PD) values—the following signals are collected:

- **Magnetometer telemetry channels:** I_B_FB_MM_0 (CADC0872) (see its fragment rendered in Fig. 1), I_B_FB_MM_1 (CADC0873), I_B_FB_MM_2 (CADC0874),
- **PD channels:** I_PD1_THETA (CADC0884), I_PD2_THETA (CADC0886), I_PD3_THETA (CADC0888), I_PD4_THETA(CADC0890), I_PD5_THETA (CADC0892), I_PD6_THETA (CADC0894).

It is of note that there are several practical challenges can be identified while exploring telemetry data captured by experimental satellites. They encompass—but are not limited to—the following issues:

- **High fragmentation of telemetry signals.** The data was registered mainly during the important stages of the OPS-SAT mission.
- **Missing data**, reflected as a high number of "gaps" in the telemetry readouts.
- Presence of the **recurring parts with noisy fragments** or with an unusual number of peaks (not necessarily abnormal) in the signal.

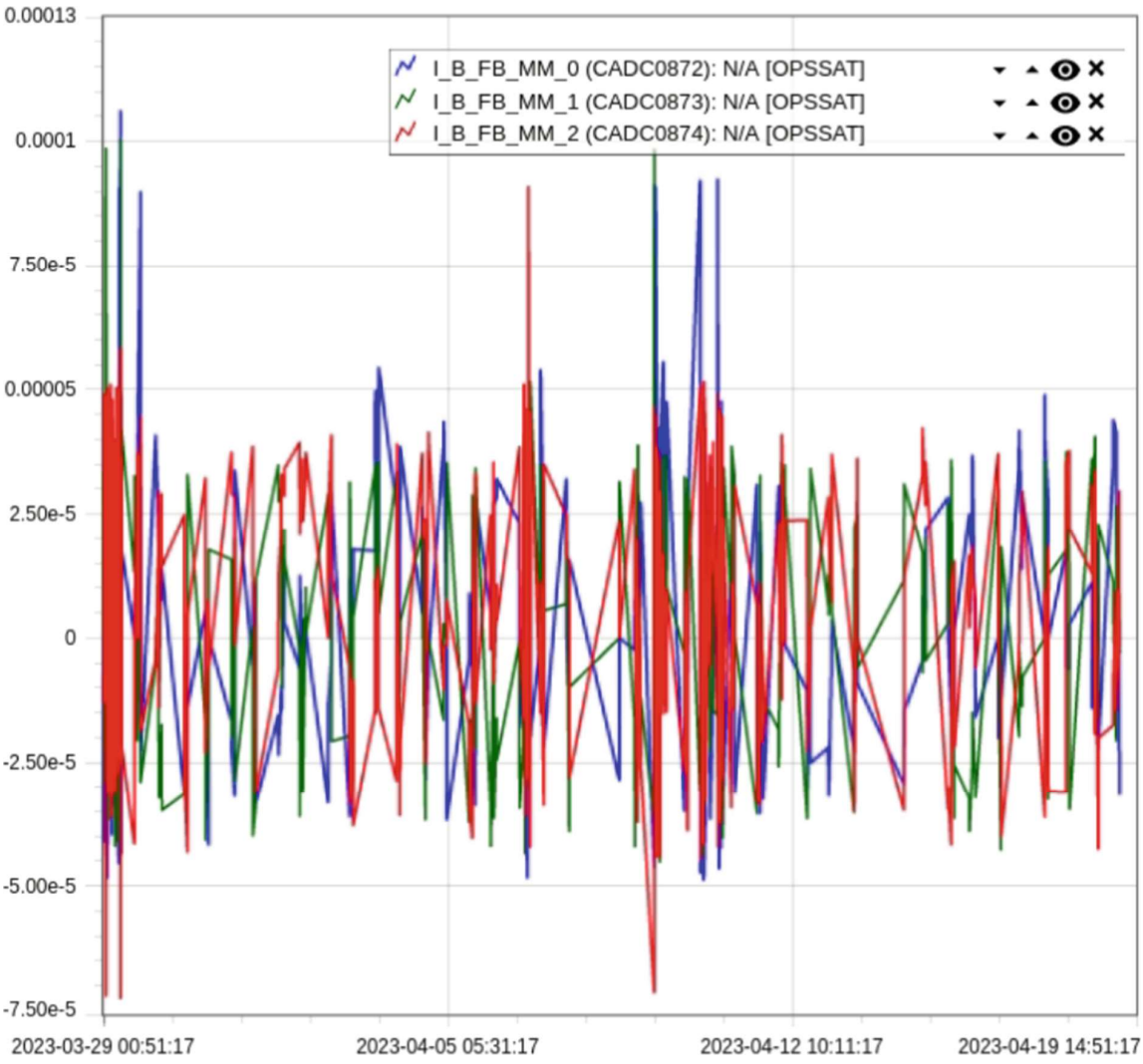


Fig. 1. A preview of three different OPS-SAT telemetries for a long (several weeks) timeframe. We can observe that the signal can easily become aperiodic and noisy, hence exploiting the approaches which focus on predicting the nominal signal and confronting it with the actual readouts [9] can easily become infeasible.

- **Non-uniform acquisition frequency rates** (the signal may be acquired hourly, daily, or even monthly), and there might be some radical changes in the signal characteristics along the analyzed months (due to e.g., various changes in satellite operations or hardware characteristics).
- **Lack of periodic and consistent nominal data** which would be captured over a long period of time.

The above-discussed telemetry channels were split into sub-parts, roughly corresponding to their periods, and they were labeled as nominal or anomalous by a human rater, and then the assigned labels were verified by the OPS-SAT

Operations Team using the OXI labeling system[1]. Overall, we collected **1273 nominal and 328 anomalous telemetry fragments for training**, and **416 nominal and 117 anomalous fragments for testing**, with the latter fragments, never seen during training (this training-test dataset split is stratified and it was generated randomly over all available telemetry fragments, to include approximately 25% of all telemetry fragments in the test set).

2.2 Detecting OPS-SAT Anomalies Using Machine Learning

In our machine learning anomaly detection pipeline, we **(1) extract an array of discriminative features from the telemetry fragments**[2], and such feature vectors are later **(2) fed to a supervised learner which performs prediction** (here, we tackle a binary classification task of determining abnormal vs. nominal telemetry fragments). For each telemetry segment, we extract the following 18 features (referred to as (*a*–*p*)), with some of them already proven to be discriminative for handling the classification tasks in telemetry data [16]:

- The features reflecting the **basic statistics** within the telemetry fragment (those include its *(a)* duration, *(b)* length, *(c)* mean value, *(d)* variance, and *(e)* standard deviation of a segment).
- The **number of peaks** within the fragment. More specifically, we sum the number of peaks with prominence of at least 10% of the segment value, and check this statistic for several segment representations, i.e., for *(f)* raw segment, alongside for its two variants after applying a smoothing transformation, using the "lighter" *(g)*, and "stronger" smoothing *(h)*, and for its *(i)* first and *(j)* second derivative.
- The **segment's variance** using its *(k)* first and *(l)* second derivative.
- The squared sum of *(m)* the **number of missing readouts**.
- The segment's *(n)* **weighted length** (relative to its sampling that also differs across different telemetry segments in the dataset), and *(o)* the **variance relative to segment's length** and *(p)* **to the segment's duration**.

Our extractors are independent of the length of the segment, and they can be effectively exploited to extract features from the segments with missing readouts—this may not be possible while operating over the raw telemetry data.

To tackle the issue of the relatively small amount and limited representativeness of OPS-SAT training samples, we introduce several data augmentation techniques specifically targeting the telemetry segments. Although they can be effectively used to synthesize both nominal and anomalous segments based on the actual samples, we augment the nominal examples only, in order to ensure that the abnormal signal segments were validated by the OPS-SAT Operations

[1] The OXI labeling system has been developed by KP Labs and it is available at https://oxi.kplabs.pl. OXI allows for not only investigating time-series data, together with the ground-truth information but also for generating ground truth.

[2] It is of note, however, that our approach can be applied to any time-series data, not only satellite telemetry.

Team (and they are *not* the result of synthetic data augmentation). We introduce the following data transformations to augment such telemetry samples:

- **Mirroring a segment over the OX axis** (ω_1)—a segment is flipped along the horizontal axis.
- **Mirroring a segment over the OY axis** (ω_2)—in a single telemetry channel, two consecutive nominal segments are flipped along their median value, starting from the half length of the first segment to the half of the second one.
- **Shifting a segment** (ω_3)—in a single telemetry channel, two consecutive nominal segments are shifted for a certain number of steps (steps of 15% and 25% of the segments' length were applied). The resulting segment starts at the shifted (by a given number of steps) starting position, and it finishes at the shifted ending position of the first segment. Hence, the last point of the first (shifted) segment will overlap the second segment—to avoid information leaks across the training and test sets, both neighboring segments that undergo this augmentation step must be originally included in the training set.

Finally, once the features are extracted and the training set has been potentially augmented, such nominal and anomalous examples are fed to the classifier to perform training. We may easily exploit any supervised learner in the proposed approach—this flexibility of our processing chain is shown in Sect. 3.

3 Experimental Validation

The main objective of our experimental study is to investigate the classification performance of the proposed classification engine for detecting anomalies in real-life OPS-SAT telemetry data. On top of that, we are aimed at verifying the flexibility of the processing chain, alongside the impact of improving the classification part of the pipeline working on a stable set of extracted features. We focus on classic classification models, including widely-adopted random forests, multi-layer perceptrons, adaptive boosting algorithms, k-nearest neighbors and support vector machines (SVMs) with a linear kernel function. To quantify the performance of the models, we calculate their accuracy, precision, recall and F_1 score, as well as the Matthews correlation coefficient (r_ϕ) over the unseen test set, with r_ϕ being the measure commonly used in imbalanced classification tasks [3]. All metrics should be maximized, with one indicating their perfect score. We split the experimental study into two experiments: in Experiment 1 (Sect. 3.1), we train the classification models over the original data only, whereas in Experiment 2 (Sect. 3.2) we benefit from data augmentation of nominal training samples.

3.1 Experiment 1: Exploiting Original Training Dataset

In the preliminary experiment, we optimized the most important hyperparameters of the classification models using a five-fold cross-validation procedure over the training set [19]. Thus, we ultimately had the maximum depth of random

forests of 20 and a maximum allowed number of features of 50% and 50 estimators; the size of the hidden layer in the multi-layer descriptor was 400; for adaptive boosting, we used 500 estimators; for SVMs, the regularization (C) parameter was set to 8; and $k = 5$ for k-nearest neighbors. The results obtained for such optimized models are gathered in Table 1 (they are sorted according to r_ϕ). Although precision exceeds 0.92 for the k-nearest neighbor classifier, we can observe rather low r_ϕ values across all investigated algorithms, indicating that the number of false positives (i.e., abnormal telemetry segments incorrectly classified as nominal) is large. This kind of error is, however, unacceptable in practice, as it could lead to missing anomalous events, e.g., in the on-board fault detection, isolation, and recovery system. On the other hand, a too large number of false negatives can deteriorate the usability of the system, if such incorrectly classified nominal segments are to be reviewed by the operations team before taking action in response to potential on-board anomalies. In practice, the latter issue is commonly a more severe problem, especially if the number of false alarms gets (very) large. In this work, we hypothesize that deploying data augmentation routines to synthesize nominal training examples may help deal with those issues by generating a larger and more representative training set that will allow us to elaborate well-generalizing models.

Table 1. The classification results obtained over the test set using machine learning models with default parameterization (sorted by r_ϕ). The best results are **boldfaced**.

Model	Accuracy	Precision	Recall	F_1 score	r_ϕ
Random forest	**0.9294**	**0.8826**	0.8050	**0.8233**	**0.7830**
Multilayer perceptron	0.8958	0.7814	**0.8077**	0.7688	0.7215
Adaptive boosting	0.9101	0.8138	0.7774	0.7646	0.7151
k-nearest neighbors	0.9051	0.9204	0.7495	0.7643	0.7133
SVM with linear kernel	0.8957	0.8249	0.7132	0.7369	0.6765

3.2 Experiment 2: Augmenting Training Datasets

Table 2. The summary of the segment data set for OPS-SAT telemetry anomalies.

Class	Training set	ω_1	ω_2	ω_3	Validation set
Nominal	1273	2677	703	1406	416
Anomalous	328	—	—	—	117

The augmentation methods can substantially increase the size of the training set (Table 2). Several examples of augmented training samples are visualized in Fig. 2, where each column represents one telemetry segment (marked with a solid line) and its augmented version (rendered as a dashed line). In this experiment,

we initially trained a random forest classifier (as it achieved the largest F_1 score in the previous experiment across all supervised models) using all versions of the augmented training sets. To better understand the impact of specific data augmentation routines on the overall performance of the classifier, we employed them separately (and then collectively), hence obtaining four augmented training sets. In Table 3, we gather the performance gain for each augmented dataset separately, and for all of them collectively (see the last row in this table). We can appreciate that the biggest improvement in the random forest's accuracy is obtained while exploiting ω_1 (mirroring training segments over the OX axis), as well as while utilizing all augmentation techniques ($\omega_1 \wedge \omega_2 \wedge \omega_3$). Here, the results are reported for the very same test set to ensure fair comparison.

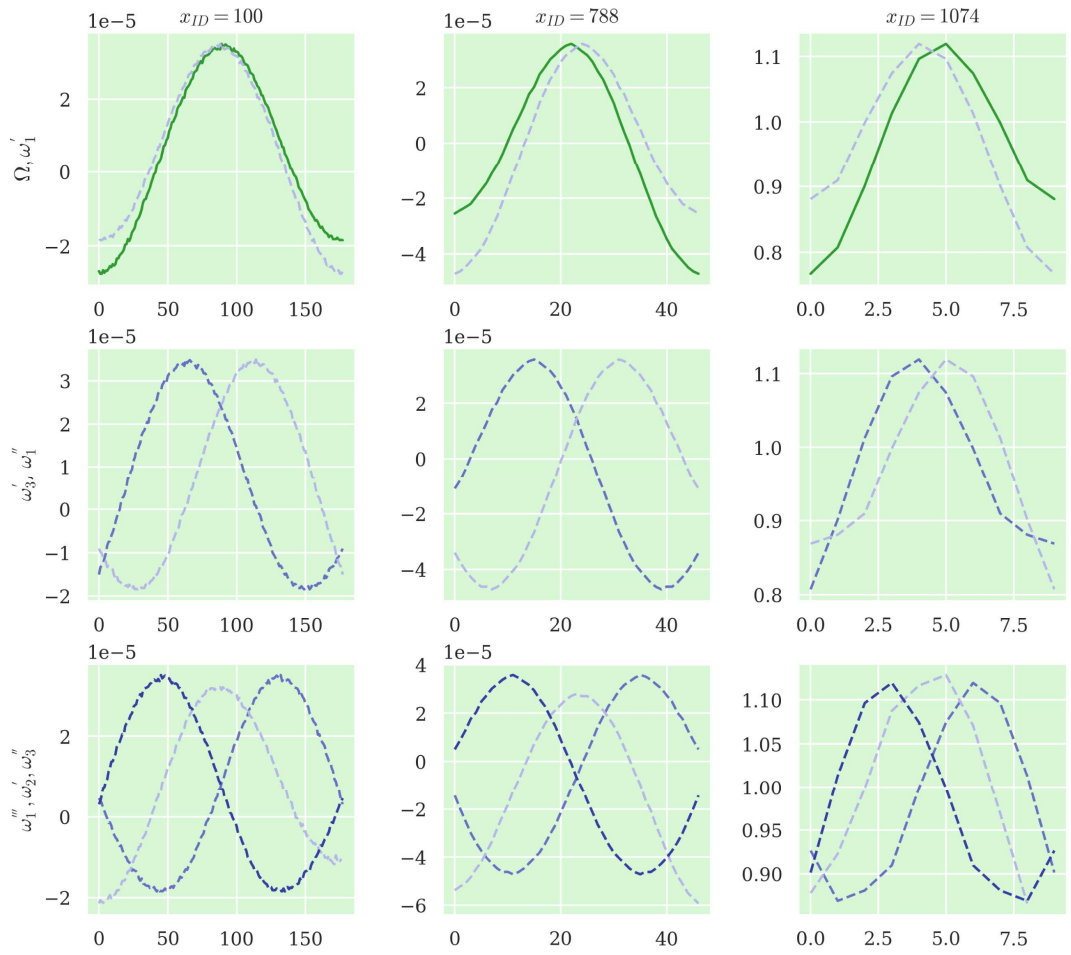


Fig. 2. The examples of the nominal telemetry segments (Ω, green solid lines), with their augmented versions with different parameterizations ($\omega_1, \omega_2, \omega_3$, blue and violet dashed lines). Each column provides one selected telemetry segment and its augmented versions. Some augmentations have been applied more than once, thus we have synthesized segments denoted as: $\omega_1^{'}, \omega_1^{''}, \omega_1^{'''}$. The presented segments are extracted from the following OPS-SAT telemetry channels: $x_{ID} = 100$ from *CADC0872*, $x_{ID} = 788$ from *CADC0874*, and $x_{ID} = 1074$ from *CADC0888*. (Color figure online)

Table 3. The benefit of exploiting data augmentation, quantified as the gain in a metric when compared to the classifier trained over the original training set. The positive and negative values indicate an increase or decrease in the model's performance.

Augmentation	Accuracy	Precision	Recall	F_1 score	r_ϕ
ω_1	0.0491	0.1055	−0.0764	0.0109	0.0535
ω_2	0.0064	−0.0363	−0.0615	−0.0592	−0.0535
ω_3	0.0347	0.0870	−0.0521	−0.0054	0.0243
$\omega_1 \wedge \omega_2 \wedge \omega_3$	0.0548	0.1174	−0.0978	−0.0079	0.0393

Knowing that the latter approach results in the largest accuracy and precision gains (hence, we minimize the number of false positive alarms), we selected this augmented version of the training set for further experiments with all other supervised models. Here, we optimized their hyperparameters following the five-fold cross-validation procedure over the augmented training set, resulting in 400 neurons in the hidden layer of the multi-layer perceptron, 500 estimators in adaptive boosting, $C = 8$ for an SVM, and $k = 5$ for k-nearest neighbors. In Table 4, we summarize the classification measures obtained using all optimized classifiers, with the random forest model (max. depth of 25) exceeding the accuracy of 0.98, and the corresponding r_ϕ amounting to 0.826, significantly outperforming the random forest model trained over the original training set (Table 1).

To provide a comparison for those metrics, a deep learning algorithm [9] obtained precision: 0.855 and recall: 0.855 (it was evaluated for the SMAP Spacecraft data), and precision: 0.926 and recall: 0.694 (for the Curiosity telemetry dataset). Although we are aware that comparing the algorithms over different sets may easily become biased, we can indeed observe that our classification models were able to exceed the reported metrics for very challenging OPS-SAT telemetry data which indicates their significant generalization capabilities and robustness against noisy and difficult time-series data.

Table 4. The results (obtained over the test set) elaborated using all investigated classification models with optimized hyperparameters and trained over the augmented training set ($\omega_1 \wedge \omega_2 \wedge \omega_3$), sorted by r_ϕ. The best results are **boldfaced**.

Model	Accuracy	Precision	Recall	F_1	r_ϕ
Random forest	**0.9843**	**1.0000**	0.7317	**0.8239**	**0.826**1
Multi-layer perceptron	0.9792	0.8716	**0.7437**	0.7795	0.7815
Adaptive boosting	0.9812	0.9422	0.6858	0.7817	0.7893
SVM with linear kernel	0.9729	0.9097	0.6154	0.6683	0.6802
k-nearest neighbors	0.9491	0.8662	0.6637	0.7186	0.7008

Finally, for the best model (random forest), we present the confusion matrix in Fig. 3, which—besides presenting all measures (the number of true positives, true negatives, false positives, and false negatives: N_{TP}, N_{TN}, N_{FP}, and N_{FN},

respectively) brings the examples of correctly and incorrectly classified test segments. The classifier correctly identified 82% abnormal segments (96/117), and 99% (413/416) nominal segments, showing its potential practical utility for onboard anomaly detection. Additionally, we can observe a significant heterogeneity of the test segments, highlighting the difficulty of the classification task.

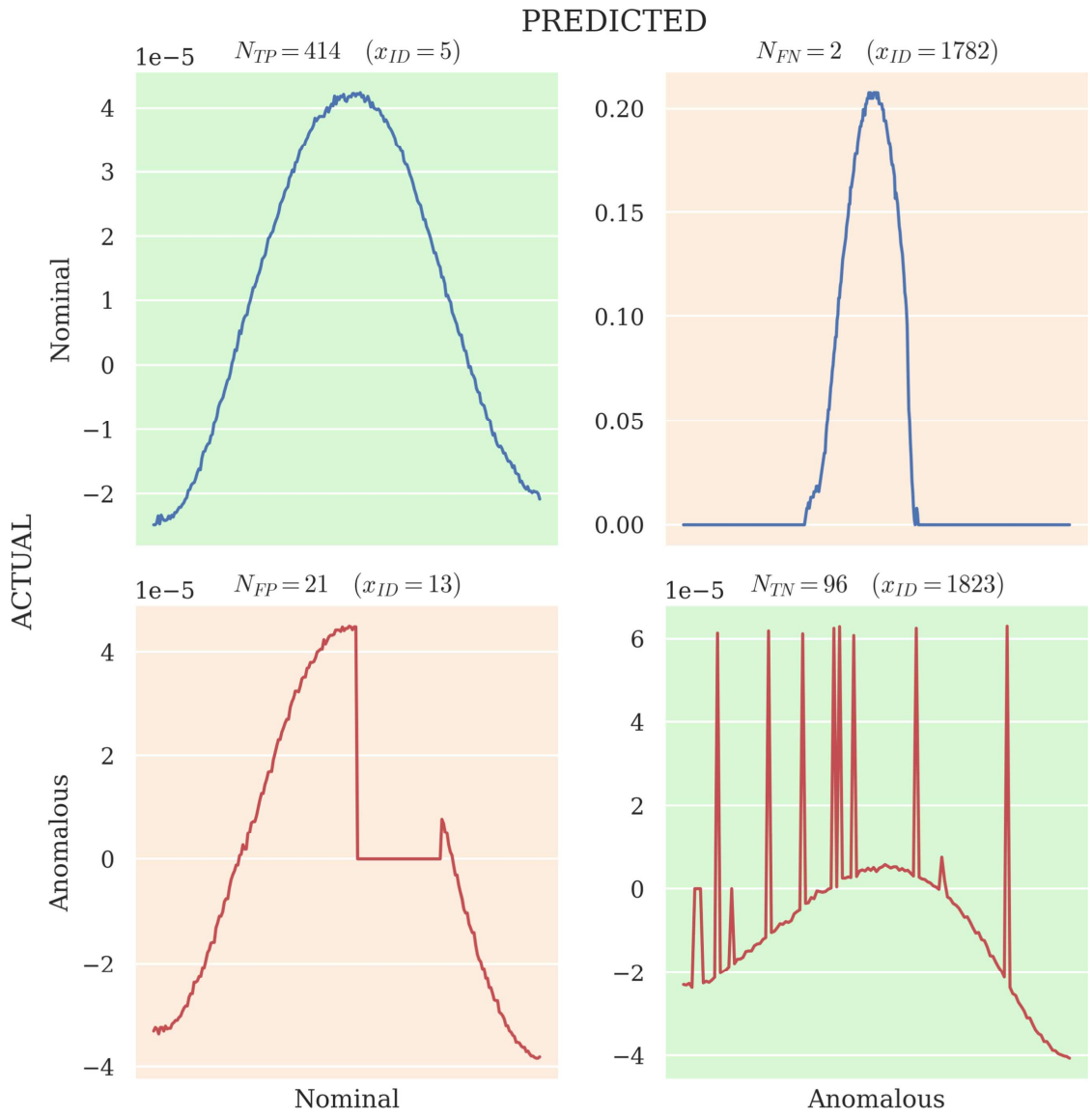


Fig. 3. The confusion matrix for the best classifier (random forest trained over the augmented training set with optimized hyperparameters), and the examples of correctly and incorrectly classified test segments (with the green and orange background, respectively). These example segments were extracted from the following OPS-SAT telemetries: $x_{ID} = 1782$ from *CADC0894*, $x_{ID} \in \{5, 13, 1823\}$ from *CADC0872*. (Color figure online)

4 Conclusions and Future Work

Detecting anomalies in spacecraft telemetry data is of paramount practical importance to appropriately respond to various unexpected events that may happen on-board an operational satellite. Albeit there exist data-driven approaches toward this task, they are virtually never validated over telemetry channels captured by a real spacecraft. In this paper, we tackled this problem and proposed an end-to-end machine learning pipeline for detecting such abnormalities. Our approach benefits from hand-crafted feature extractors which are independent of the length of the telemetry segments which are later classified by a supervised learner as nominal or anomalous. To understand its generalization capabilities, we validated it over a curated (and validated by the ESA Operations Team) set of nominal and abnormal telemetry channels acquired on-board OPS-SAT—a nanosatellite operated by ESA. The experiments indicated that our technique offers high-quality detection of anomalies from OPS-SAT telemetry. Also, we showed that exploiting the suggested data augmentation routines allows for significantly improving the generalization capabilities of the classification models.

Our current research efforts are focused on deploying the proposed anomaly detection system on-board OPS-SAT. It is of note that the model has been already trained on the ground, and we are in the process of uplinking it to the satellite. Also, we are working on utilizing our method for other missions, to further prove its generalizability. Finally, it would be interesting to exploit unsupervised clustering over large-scale telemetry datasets while utilizing our feature extractors. This may help accelerate the process of generating ground truth for supervised models (as pre-classified telemetry segments would have to be verified by humans), and possibly uncover intrinsic characteristics of such time-series data that might have not been observed by the Operations Teams.

References

1. Benecki, P., Piechaczek, S., Kostrzewa, D., Nalepa, J.: Detecting anomalies in spacecraft telemetry using evolutionary thresholding and LSTMs. In: Proceedings of the GECCO, GECCO 2021, ACM, New York, NY, USA, pp. 143–144 (2021)
2. Castellani, A., Schmitt, S., Squartini, S.: Real-world anomaly detection by using digital twin systems and weakly-supervised learning. CoRR abs/2011.06296 (2020). https://arxiv.org/abs/2011.06296
3. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 6 (2020)
4. ESA: WebMUST - web client for OPS-SAT directory (2021). https://opssat1.esoc.esa.int/webclient-must
5. Evans, D., et al.: Implementing the New CCSDS Housekeeping Data Compression Standard 124.0-B-1 (Based on POCKET+) on OPS-SAT-1. Small Satellite Conference (2022). https://digitalcommons.usu.edu/smallsat/2022/all2022/133
6. Evans, D.J., Ortega, M., Zeif, R., Sergert, T.: OPS-SAT: FDIR design on a mission that expects bugs - and lots of them. In: SpaceOps 2016 Conference. American Institute of Aeronautics and Astronautics (2016). https://doi.org/10.2514/6.2016-2481

7. Fratini, S., Policella, N., Silva, R., Guerreiro, J.: On-board autonomy operations for OPS-SAT experiment. Appl. Intell. **52**(6), 6970–6987 (2022)
8. Gao, Y., et al.: An unsupervised anomaly detection approach for spacecraft based on normal behavior clustering. In: Proceedings of the ICICTA, pp. 478–481 (2012)
9. Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Söderström, T.: Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: Proceedings of the SIGKDD, pp. 387–395. ACM (2018)
10. Kacker, S., Meredith, A., Cahoy, K., Labrèche, G.: Machine learning image processing algorithms onboard OPS-SAT. In: Small Satellite Conference (2022). https://digitalcommons.usu.edu/smallsat/2022/all2022/65
11. Kopp, M., Pevný, T., Holeňa, M.: Anomaly explanation with random forests. Expert Syst. Appl. **149**, 113187 (2020)
12. Kubicka, M., Zeif, R., Henkel, M., Hörmer, A.J.: Thermal vacuum tests for the ESA's OPS-SAT mission. Elektrotech. Informationstechnik **139**(1), 16–24 (2022)
13. Labreche, G., et al.: OPS-SAT spacecraft autonomy with TensorFlow lite, unsupervised learning, and online machine learning. In: Proceedings of the IEEE Aerospace Conference. IEEE Computer Society (2022)
14. Li, K., et al.: A novel method for spacecraft electrical fault detection based on FCM clustering and WPSVM classification with PCA feature extraction. Proc. Inst. Mech. Eng. **231**(1), 98–108 (2017)
15. Marszk, D., Evans, D., Mladenov, T., Labrèche, G., Zelenevskiy, V., Shiradhonkar, V.: MO services and CFDP in action on OPS-SAT. In: Small Satellite Conference (2022). https://digitalcommons.usu.edu/smallsat/2022/all2022/67
16. Nalepa, J., et al.: Toward On-Board Detection Of Anomalous Events From Ops-Sat Telemetry Using Deep Learning (2022). https://www.researchgate.net/publication/367091759
17. Nalepa, J., Myller, M., Andrzejewski, J., Benecki, P., Piechaczek, S., Kostrzewa, D.: Evaluating algorithms for anomaly detection in satellite telemetry data. Acta Astronaut. **198**, 689–701 (2022)
18. Nassif, A.B., Talib, M.A., Nasir, Q., Dakalbab, F.M.: Machine learning for anomaly detection: a systematic review. IEEE Access **9**, 78658–78700 (2021)
19. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
20. Primartha, R., Tama, B.A.: Anomaly detection using random forest: a performance revisited. In: Proceedings ICoDSE, pp. 1–6 (2017)
21. Segret, B., et al.: On-board images to characterize a CubeSat's ADCS. In: Small Satellite Conference (2022). https://digitalcommons.usu.edu/smallsat/2022/all2022/174
22. Toussaint, F., Thomassin, J., Laurens, S.: ASTERIA in-orbit testing on OPSSAT: an on-board autonomous orbit control solution including collision risks avoidance. In: Small Satellite Conference (2022). https://digitalcommons.usu.edu/smallsat/2022/all2022/103
23. Wu, R., Keogh, E.J.: Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress (extended abstract). In: 2022 IEEE 38th International Conference on Data Engineering (ICDE), pp. 1479–1480 (2022)
24. Zeif, R., Henkel, M., Hörmer, A., Kubicka, M., Wenger, M., Koudelka, O.: The redundancy and fail-safe concept of the OPS-SAT payload processing platform. In: 69th International Astronautical Congress, IAC 2018 (2018)
25. Zeng, Z., Jin, G., Xu, C., Chen, S., Zhang, L.: Spacecraft telemetry anomaly detection based on parametric causality and double-criteria drift streaming peaks over threshold. Appl. Sci. **12**(4), 1–24 (2022)

Załącznik 9 – [A9]

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The OPS-SAT benchmark for detecting anomalies in satellite telemetry

Bogdan Ruszczak[1,2] ✉, Krzysztof Kotowski[2], David Evans[3] & Jakub Nalepa[4]

Detecting anomalous events in satellite telemetry is a critical task in space operations. It is time-consuming, error-prone and human dependent, thus automated data-driven algorithms have been emerging at a steady pace. However, there are no available datasets of real satellite telemetry with annotations to verify anomaly detection models. We address this gap and introduce the AI-ready benchmark dataset (OPSSAT-AD) containing the telemetries acquired on board OPS-SAT—a CubeSat mission, operated by the European Space Agency. The dataset is accompanied with the baseline results obtained using 30 supervised and unsupervised classic and deep machine learning algorithms. They were evaluated using the training-test dataset split introduced in this work, and we suggest a set of quality metrics which should be calculated to confront the new algorithms for anomaly detection while exploiting OPSSAT-AD. We believe that this work may become an important step toward building a fair, reproducible and objective validation procedure that can be used to quantify the capabilities of the emerging techniques in an unbiased and fully transparent way.

Background & Summary

The anomaly detection (AD) domain encompasses a diverse array of methodologies for the identification of anomalous patterns in data of various modalities. These approaches can be applied to a multitude of data types, including images, text, and time series data, among others. However, the development and evaluation of real-world anomaly detection applications are dependent on the availability of real-world data. Currently, there is a considerable number of datasets available for a wide range of scenarios[1], but the satellite telemetry data for AD is an extremely underrepresented category in this catalogue. This kind of data is difficult and costly to obtain, often confidential, and requires expert knowledge to annotate properly. The only two widely accessible and used collections of this type include the NASA Soil Moisture Active Passive (SMAP) and Mars Science Laboratory (MSL) datasets[2]. They offer short fragments of signals and related commands from 55 and 27 telemetry parameters, respectively, with a total of 105 annotated anomalies. However, the recent consensus in the community is that they should not be used for time series AD benchmarking due to their unrealistic anomaly density, many trivial anomalies, mislabelled ground truth, distributional shifts, and a lack of meaningful correlation between commands and channels[3–5]. Other well-known satellite telemetry datasets, such as Mars Express[6] or NASA WebTCAD[7], do not contain annotations of anomalous events. There is an ongoing activity to publish a large-scale AD dataset by European Space Agency (ESA) solving all the mentioned issues[8,9], but it will primarily address the needs of large-scale, complex and relatively stable missions.

The dataset introduced in this article, dubbed OPSSAT-AD, is fundamentally different from those available in the literature, as it tackles a very specific ESA OPS-SAT mission—a CubeSat flying laboratory, for which we might expect a noticeable number of abnormal events[10]. The raw telemetry from OPS-SAT is characterized by many data gaps, artifacts, sampling frequency changes, and signal amplitude variations. The dataset was collectively curated by space operations engineers and machine learning experts to make it useful for building and validating data-driven anomaly detection techniques. It includes a selection and the corresponding ground-truth annotation of 2123 short single-channel satellite telemetry fragments (univariate time series) captured within

[1]Computer Science Department, Opole University of Technology, Prószkowska Str. 76, 45-758, Opole, Poland. [2]KP Labs, Bojkowska Str. 37J, 44-100, Gliwice, Poland. [3]European Space Agency/ESOC, Robert-Bosch-Str. 5, 64293, Darmstadt, Germany. [4]Faculty of Automatic Control, Electronics and Computer Science, Department of Algorithmics and Software, Silesian University of Technology, Akademicka Str. 16, 44-100, Gliwice, Poland. ✉e-mail: b.ruszczak@po.edu.pl

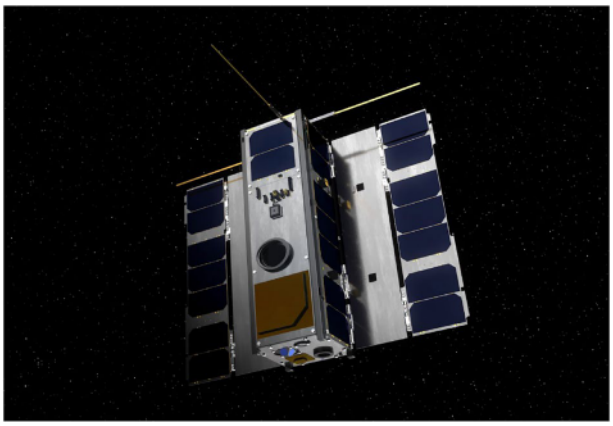


Fig. 1 The ESA OPS-SAT frontal view. Image credits: European Space Agency.

9 telemetry channels. Due to the underlying nature of the OPS-SAT mission, anomalous fragments account for 20% of the dataset. Such fragments contain raw data with many aforementioned real-life challenges, and they differ in their length and sampling frequency. For each telemetry fragment, the dataset also contains a set of 18 handcrafted features used in the actual machine learning AD algorithm validated on board OPS-SAT[11]. These features are exploited in this article to benchmark 30 other supervised and unsupervised machine learning algorithms for anomaly detection. All of them were trained on 1494 and tested on 529 telemetry segments, and assessed using 7 metrics suggested for quantifying the operational capabilities of anomaly detection algorithms—this training-test dataset split is included in our benchmark as well.

Overall, the benchmark (including the dataset, training-test dataset split, suggested quality metrics, and our baseline results) introduced in this paper shall help the community to create and compare their approaches to detecting anomalies in real-life satellite telemetry in a fair and unbiased way. Therefore, we also address the reproducibility crisis currently observed in the (not only) machine learning community[12]. While the OPS-SAT spacecraft completed its atmospheric reentry at the end of May 2024, its successor—OPS-SAT VOLT—is going to be launched in late 2025 and will make a great opportunity to validate the algorithms developed based on our benchmark in the wild after deploying them on-board an operational satellite. Although future satellites may be equipped with different sensor sets, with possible variations in sampling rates or data ranges, and other telemetry may be important to monitor, we believe that a collection of real data such as the one examined in this paper is of great value to the community interested in on-board deployments.

Methods

Data acquisition and annotation. The telemetry data delivered[13] in this paper was acquired from the ESA OPS-SAT satellite (Fig. 1). It is a small 3-unit (3U, where 1U=10 cm^3) CubeSat launched in December 2019 with the primary objective of being a technological demonstrator for in-orbit data processing. It finished its mission with the atmospheric reentry on 22 May 2024, but it generated lots of useful data during more than 4 years of its operations, including satellite imagery[14] and telemetry[11].

OPS-SAT offered a unique opportunity for researchers to run their experiments and algorithms in orbit. While these experiments were carried out, all telemetry data was simultaneously collected and recorded in the ESA archive. The archive was monitored for potential anomalies to ensure the mission's stable and uninterrupted operation. Our dataset consists of telemetry fragments recommended by the OPS-SAT operation engineers as the most "interesting" (according to their subjective assessment) for anomaly detection. The actual data collection process was carried out using the data exchange platform WebMUST[15] used in the European Space Operations Centre (ESOC). This platform is restricted to the authorized ESA partners only, but the data included in our dataset package does not have to be requested through it, and thus is made publicly available.

The online OXI tool for visualization and annotation of satellite telemetry (https://oxi.kplabs.pl/)[16] was used to enable a collaborative labeling process of the dataset. Using this application, domain experts were able to manually extract and annotate telemetry segments representing periods of nominal and anomalous operation. The initial selection of anomalies was provided by 3 ESA spacecraft operations engineers and further curated by 2 machine learning experts (with more than 10 years of experience each). The curated annotations were finally reviewed by the three spacecraft operations engineers. The detailed satellite telemetry annotation process, together with the visual artefacts generated throughout it, are discussed in[11,17,18].

The telemetry segments have been designated as anomalous for the cases that the operators should be notified about. The anomalous segments included in the collection are characterized by instances of unusual signal shapes, unexpected sensor behavior (e.g., single or multiple peaks), signal fragments with zero values, or gaps in the signal. In such cases, it is imperative to initiate the shutdown or restart of pertinent aircraft systems or sensors, or alternatively, to transition their operation to safe mode.

Feature extraction. Due to the characteristics of satellite telemetry, the segments of raw data selected by the domain experts have varying lengths and sampling frequency. As such, they could not be handled by most machine learning algorithms without performing an additional preprocessing or feature extraction. Thus, 18

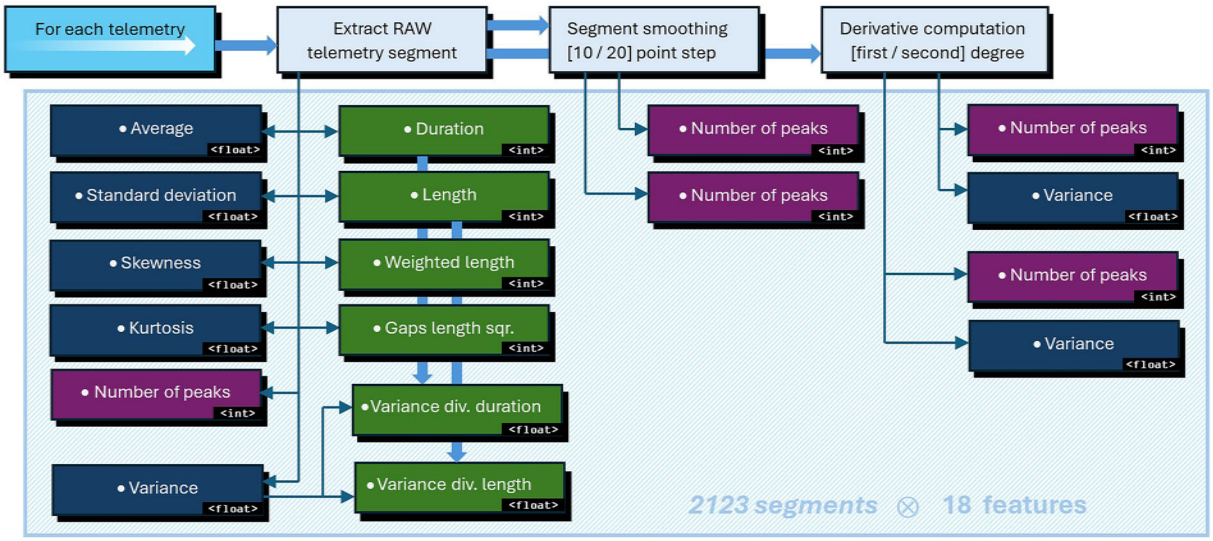


Fig. 2 The features extracted for each segment, with the corresponding data type. The meaning of the colors: dark blue for popular statistics, violet for peak counters for various converted segments, and green for the length or duration-related features.

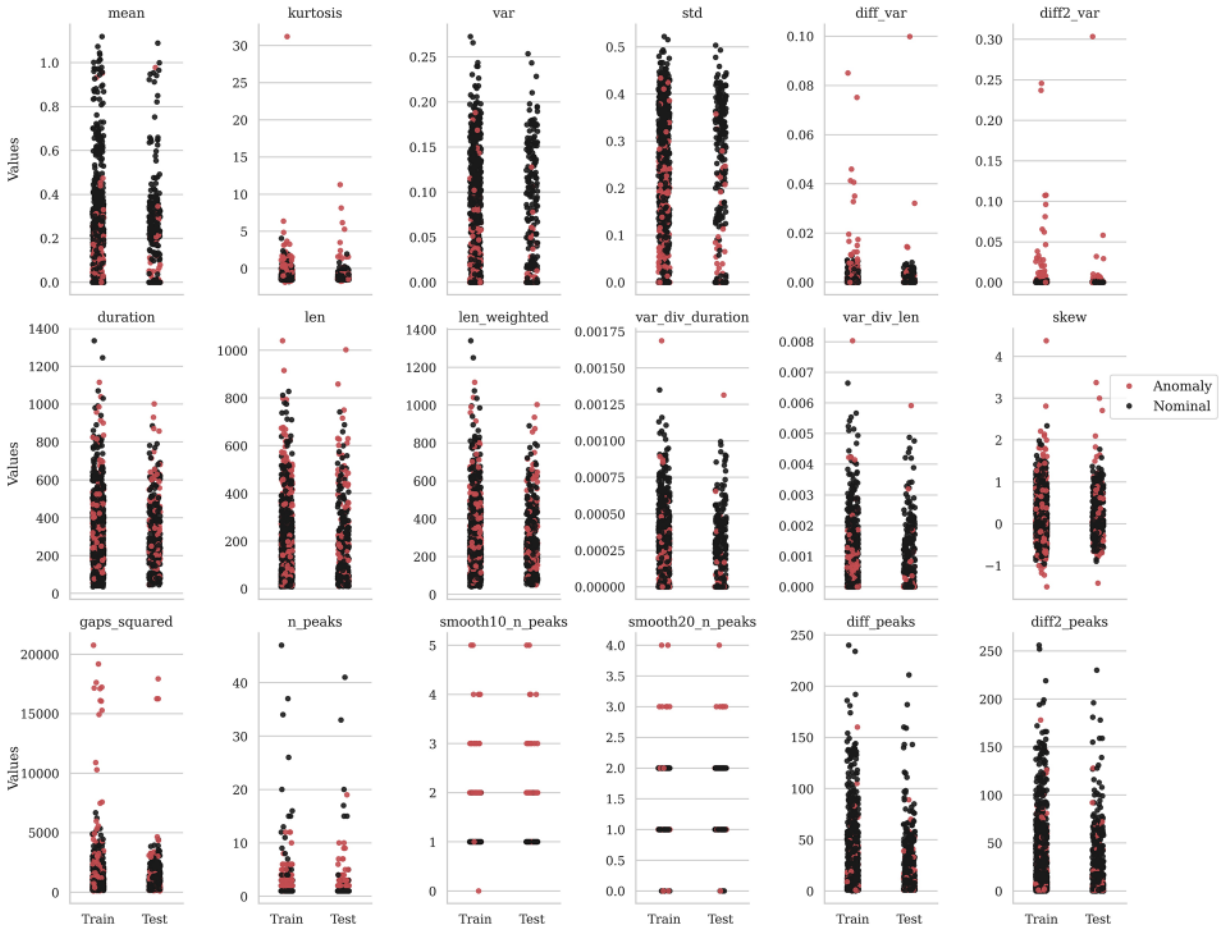


Fig. 3 Dataset features for the training and validation subsets with the indication of anomalies (marked in red).

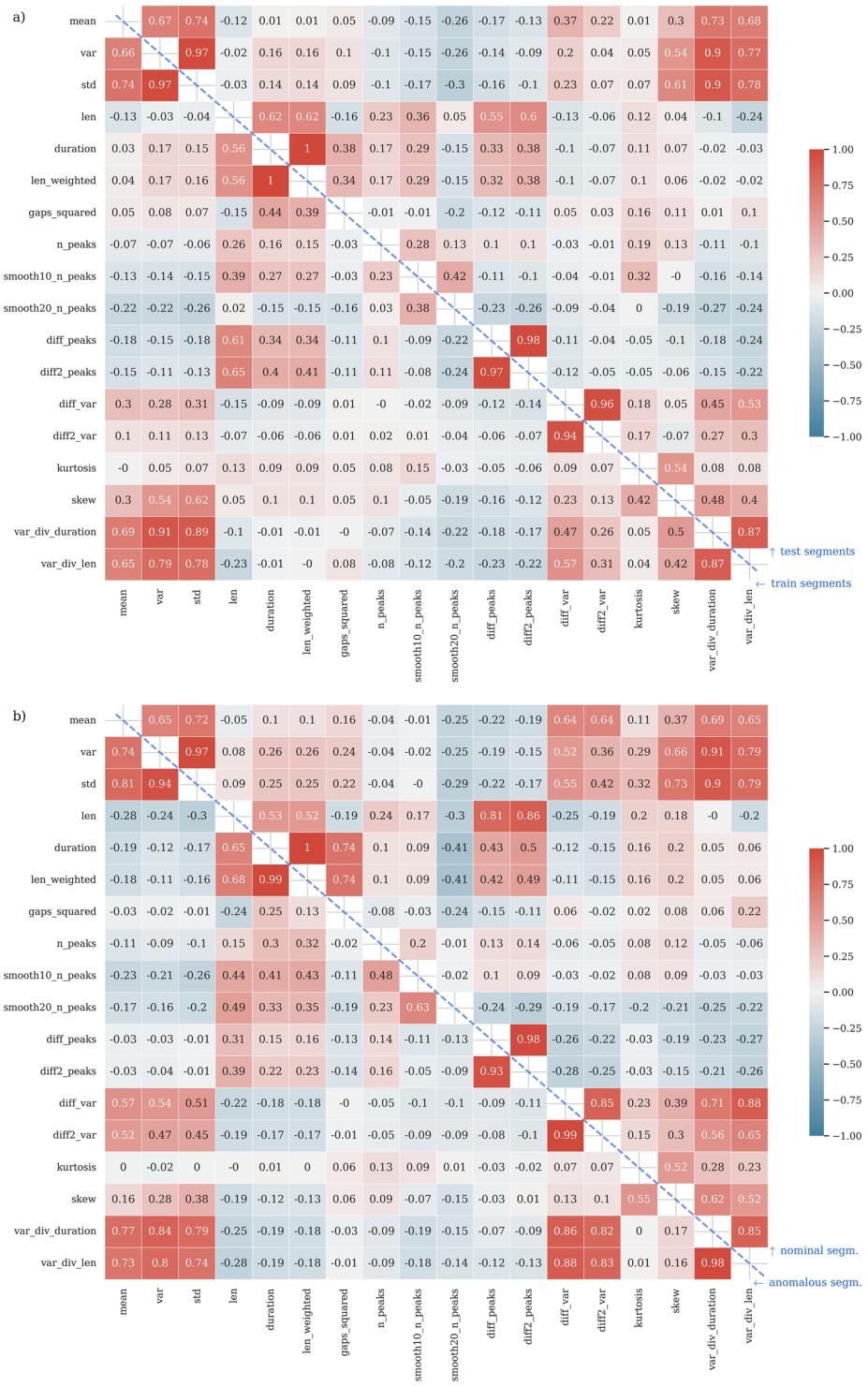


Fig. 4 The coefficient of correlation computed between each feature, for the training (below the blue dashed diagonal line) and test set (above the same diagonal line). We employed: (**a**) Pearson's correlation coefficient and (**b**) Spearman's Rank correlation coefficient.

handcrafted features were designed for the task of anomaly detection[11]—they were calculated separately for each segment, and they are included in our benchmark. An algorithm operating on such features was already validated in our previous work focusing on the application of data-driven anomaly detection on board OPS-SAT[11].

The features extracted for each telemetry segment are presented in Fig. 2. They are divided into three groups:

- **12 features extracted from raw segments**, including basic statistics, such as the arithmetic average of the signal values, their standard deviation, skewness, kurtosis and variance ($\langle mean \rangle$, $\langle std \rangle$, $\langle skew \rangle$, $\langle kurtosis \rangle$,

Class	Training set (T)	Test set (Ψ)	Total
Nominal	1273	416	1689
Anomalous	321	113	434
Sum	1494	529	2123

Table 1. The number of telemetry segments included in the training (T) and test (Ψ) sets.

Abbreviation	Year	Algorithm
Supervised algorithms		
Linear+L2[30]	2006	Linear classifier with L_2 regularization
LR[31]	2008	Logistic regression
AdaBoost[32]	2009	Adaptive Boosting
LSVC[33]	2013	Support Vector Classifier with the squared hinge linear loss
XGBOD[34]	2018	Extreme Gradient Boosting Outlier Detection
FCNN[35]	2019	Fully Connected Neural Network with dropout and batch normalization
RF+ICCS[11]	2023	Random Forest based model with segment augmentation
Unsupervised algorithms		
PCA[36,37]	1996	Principal Component Analysis
LMDD[38]	1996	Linear Method for Deviation Detection
COF[39]	2002	Connectivity-based Outlier Factor
KNN[40]	2002	K-Nearest Neighbors
CBLOF[41]	2003	Cluster-Based Local Outlier Factor
ABOD[42]	2008	Angle-based Outlier Detector
IForest[43]	2008	Isolation Forest
SOD[44]	2009	Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data
SOS[45]	2012	Stochastic Outlier Selection
VAE[46]	2013	Variational Autoencoder
OCSVM[47]	2016	One-Class Support Vector Machine with a polynomial kernel
LODA[48]	2016	Lightweight On-line Detector of Anomalies
GMM[37]	2017	Gaussian Mixture Model
AnoGAN[49]	2017	Generative Adversarial Networks for AD
DeepSVDD[50]	2018	Deep one-class classification
ALAD[51]	2018	Generative Adversarial Networks for AD
INNE[52]	2018	Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles
SO-GAAL[53]	2020	Single-objective Generative Adversarial Active Learning
MO-GAAL[53]	2020	Multi-objective Generative Adversarial Active Learning
COPOD[54]	2020	Copula-based outlier detection
ECOD[55]	2022	Empirical Cumulative Distribution Functions
LUNAR[56]	2022	Unified Local Outlier Detection with Graph Neural Networks
DIF[57]	2023	Deep Isolation Forest

Table 2. Anomaly detection methods investigated in this study.

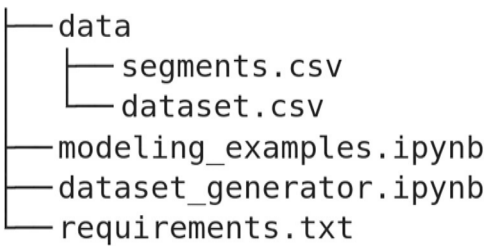


Fig. 5 The layout of the OPS-SAT benchmark for anomaly detection.

and ⟨*var*⟩), but also the number of peaks (of the minimum of 10% prominence, with a peak prominence measuring how much a peak "stands out" in relation to the signal, while considering its height and location: ⟨*n_peaks*⟩), duration (in seconds: ⟨*duration*⟩) and the length (in the number of telemetry points: ⟨*len*⟩), the

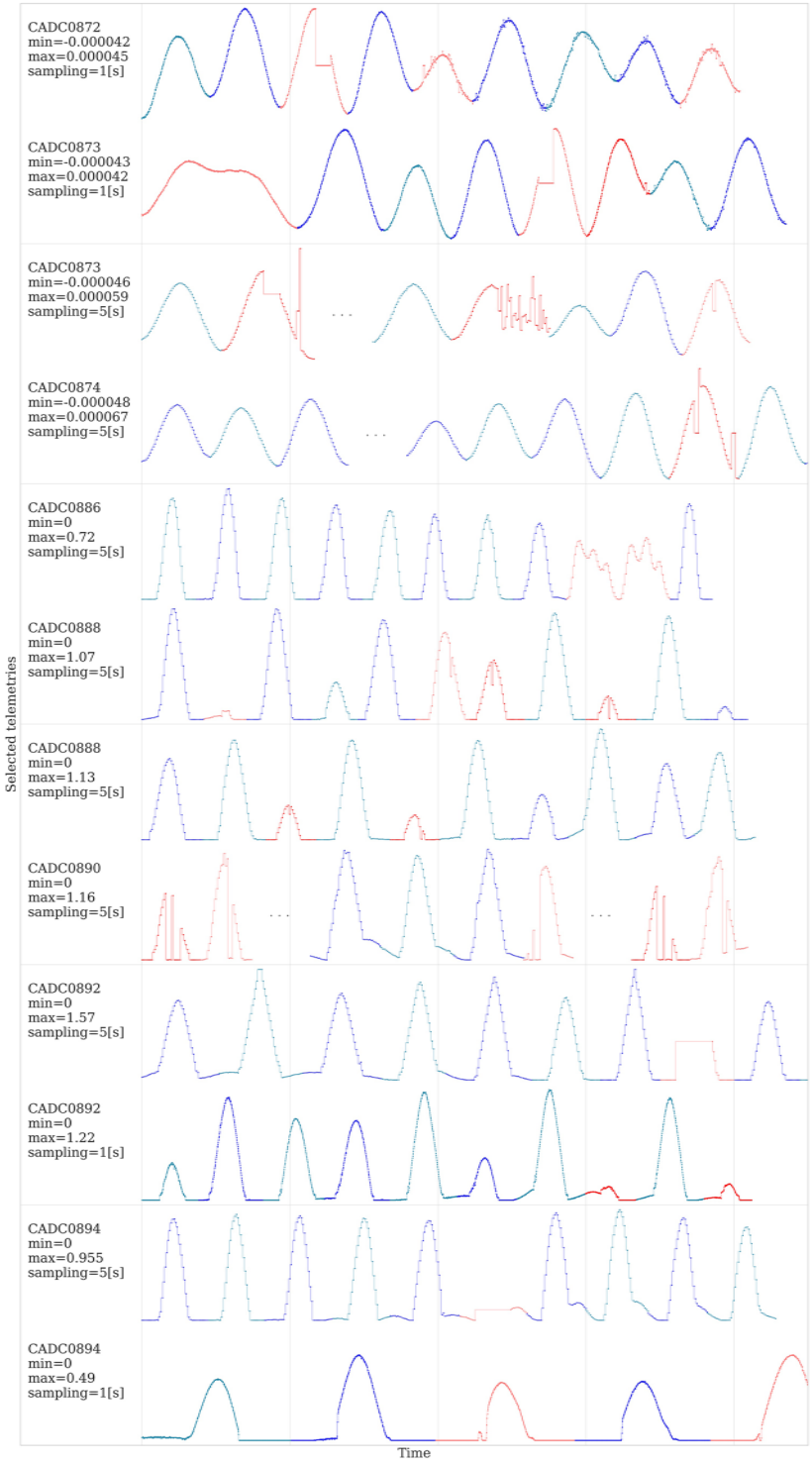


Fig. 6 Selected segments from out OPS-SAT dataset. Several types of signal distortions are depicted, including peaks, deformations, noise (`CADC0873`), irregular periodicity (`CADC0886`), short (`CADC0892`, `CADC0894`) and long data gaps (`CADC0874`). Anomalous segments are plotted in red. For brevity, we omit the axis values, but provide data ranges and the sampling information for each channel.

weighted length (weighted by sampling: ⟨*len_weighted*⟩), the gaps' length (the squared number of missing data points: ⟨*gaps_squared*⟩), and the weighted variance (weighted by the duration and by the length: ⟨*var_div_duration*⟩, ⟨*var_div_len*⟩).

- **2 features extracted from the smoothed segments (using the uniform interpolation**[19]**)**, including the number of peaks (extracted using the 10 and 20 points smoothing steps: ⟨*smooth*10*_n_peaks*⟩, ⟨*smooth*20*_n_peak*⟩).

Model	AUC_{PR} (↑)	AUC_{ROC} (↑)	*Accuracy* (↑)	F_1 (↑)	*Precision* (↑)	*Recall* (↑)	*MCC* (↑)
Supervised algorithms							
FCNN	**0.979**	0.989	**0.977**	**0.946**	0.963	**0.929**	**0.932**
XGBOD	0.975	**0.992**	0.966	0.918	0.944	0.894	0.897
RF+ICCS	0.963	0.985	0.955	0.883	0.978	0.805	0.862
LSVC	0.934	0.968	0.926	0.808	0.911	0.726	0.771
LR	0.931	0.969	0.924	0.800	0.920	0.708	0.764
AdaBoost	0.923	0.962	0.934	0.836	0.890	0.788	0.797
Linear+L2	0.901	0.958	0.905	0.722	0.970	0.575	0.703
Unsupervised algorithms							
MO-GAAL	0.779	0.865	0.907	0.726	0.985	0.575	0.710
AnoGAN	0.668	0.756	0.868	0.588	0.877	0.442	0.563
SO-GAAL	0.660	0.749	0.885	0.655	0.906	0.513	0.627
OCSVM	0.659	0.787	0.845	0.647	0.630	0.664	0.548
KNN	0.658	0.852	0.824	0.575	0.594	0.558	0.465
ABOD	0.644	0.843	0.832	0.582	0.620	0.549	0.479
INNE	0.643	0.806	0.847	0.646	0.638	0.655	0.549
ALAD	0.629	0.744	0.870	0.596	0.879	0.451	0.570
LMDD	0.623	0.767	0.854	0.628	0.691	0.575	0.542
SOD	0.621	0.797	0.737	0.505	0.423	0.628	0.348
COF	0.603	0.774	0.794	0.576	0.514	0.655	0.448
LODA	0.597	0.748	0.822	0.588	0.583	0.593	0.475
LUNAR	0.540	0.792	0.813	0.407	0.630	0.301	0.342
CBLOF	0.493	0.642	0.756	0.427	0.429	0.425	0.272
DIF	0.465	0.797	0.790	0.035	**1.000**	0.018	0.118
VAE	0.450	0.680	0.796	0.349	0.547	0.257	0.272
GMM	0.426	0.713	0.737	0.393	0.388	0.398	0.225
DeepSVDD	0.375	0.610	0.775	0.279	0.442	0.204	0.184
PCA	0.373	0.612	0.728	0.357	0.360	0.354	0.185
IForest	0.347	0.635	0.701	0.295	0.297	0.292	0.105
ECOD	0.340	0.637	0.720	0.345	0.345	0.345	0.167
COPOD	0.328	0.627	0.703	0.270	0.284	0.257	0.084
SOS	0.308	0.524	0.705	0.264	0.283	0.248	0.081

Table 3. The experimental results for the universal detector, sorted by AUC_{PR}. The globally best results for each metric are **boldfaced**, and the best among the unsupervised algorithms are underlined.

- **4 features extracted from the first and the second derivatives of the segment**, including the number of peaks and variance (⟨*diff_peaks*⟩, ⟨*diff2_peaks*⟩, ⟨*diff_var*⟩, ⟨*diff2_var*⟩).

Employing the duration, the length and the gaps' length features should allow the algorithms to easily capture some "obvious" abnormalities in the telemetry data. This intuitively could lead to promote some less computationally demanding AD methods suitable for on-board applications. The proposed set of features serves as an example and may be easily expanded (or replaced) by the community by (*i*) designing new feature extractors (potentially followed by feature selectors), (*ii*) using other well-established feature sets[20] or (*iii*) benefiting from the automated feature learning[21].

The distribution of the extracted features elaborated for each telemetry segment included in our dataset is depicted in Fig. 3. We compare the distribution for the training ($\boldsymbol{T}$) and test (Ψ) sets, to visualize the effect of the training-test dataset split on the feature distributions. Figure 4 provides a detailed view of the relations between the extracted features for the $\boldsymbol{T}$ and Ψ sets. We rendered this plot to confirm that both subsets represent similar data distributions.

Benchmarking procedure. We provide a procedure that should be followed to confront the AD algorithms over our dataset. The entire dataset of 2123 telemetry segments is split into the training ($\boldsymbol{T}$) and test (Ψ) sets (Table 1), forming an AI-ready dataset. To extract these subsets, we performed stratified random sampling to maintain the original percentage of anomalies in both $\boldsymbol{T}$ and Ψ.

The benchmarking procedure can be summarized by the following steps:

1. Load the dataset from the `dataset.csv` file.
2. Split the dataset into $\boldsymbol{T}$ and Ψ according to the ⟨*train*⟩ attribute included in this file.
3. **[Optionally]** Preprocess the datasets using e.g., data normalization, additional feature extraction, feature selection and other steps directly related to the AD algorithm which undergoes the benchmarking process.

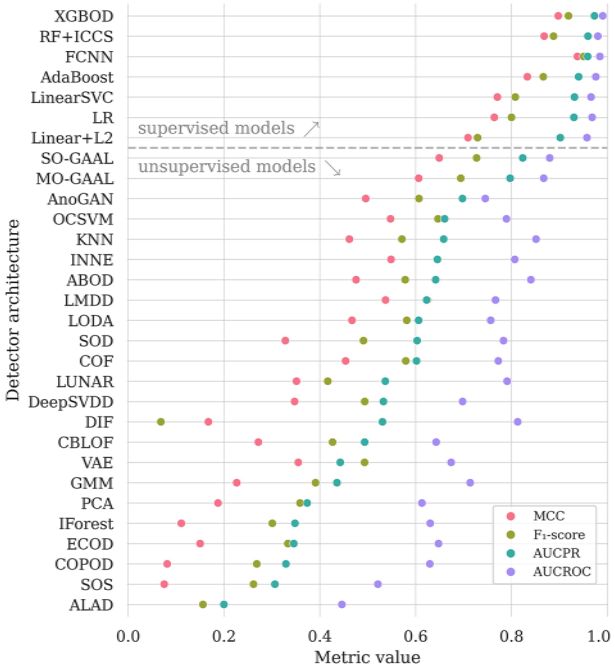


Fig. 7 The results (over Ψ) obtained using the investigated machine learning models (first grouped according to their training strategy, either supervised or unsupervised, then sorted by F_1).

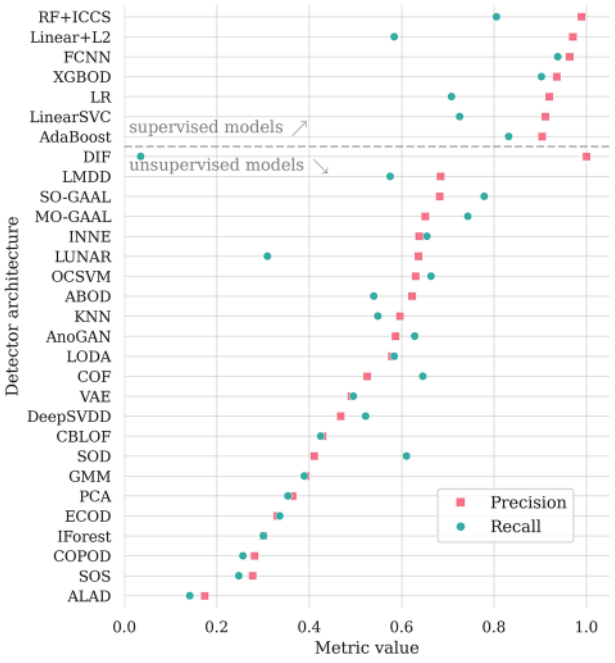


Fig. 8 Precision and recall metrics (over Ψ) obtained using the investigated algorithms. Models are sorted according to the number of misclassified telemetry segments, with the best-performing one rendered on top of the graph.

4. **[Optionally]** Train a machine learning model over $\boldsymbol{T}$.
5. Quantify the algorithm's performance over Ψ using the metrics discussed in the next section.

Quality metrics. The following metrics should always be calculated over the test set Ψ while confronting the AD algorithms (both supervised and unsupervised) over the dataset OPSSAT-AD introduced in this work:

Model	AUC_{PR} (↑)	AUC_{ROC} (↑)	*Accuracy* (↑)	F_1 (↑)	*Precision* (↑)	*Recall* (↑)	*MCC* (↑)
Supervised algorithms							
LR	**0.946**	**0.974**	0.944	0.873	0.924	0.832	0.840
Linear+L2	0.940	0.967	0.949	0.880	0.986	0.798	0.856
LinearSVC	0.939	0.966	0.957	0.898	0.956	0.849	0.873
XGBOD	0.930	0.967	0.945	0.854	0.888	0.829	0.822
FCNN	0.924	0.953	**0.970**	**0.929**	**0.979**	**0.886**	**0.912**
RF	0.918	0.967	0.939	0.816	0.975	0.729	0.801
AdaBoost	0.862	0.920	0.917	0.780	0.796	0.770	0.730
Unsupervised algorithms							
KNN	0.788	0.908	0.868	0.691	0.687	0.718	0.613
SOD	0.761	0.895	0.777	0.627	0.517	0.888	0.545
OCSVM	0.760	0.848	0.844	0.659	0.628	0.715	0.564
COF	0.748	0.901	0.832	0.651	0.606	0.768	0.573
ABOD	0.746	0.891	0.855	0.654	0.636	0.682	0.563
INNE	0.730	0.860	0.847	0.645	0.638	0.669	0.551
CBLOF	0.726	0.815	0.823	0.608	0.606	0.650	0.511
SO-GAAL	0.721	0.797	0.818	0.569	0.667	0.544	0.489
AnoGAN	0.686	0.808	0.814	0.576	0.612	0.591	0.483
DIF	0.676	0.868	0.798	0.305	0.643	0.248	0.289
DeepSVDD	0.647	0.777	0.749	0.513	0.469	0.636	0.381
IForest	0.642	0.806	0.803	0.537	0.595	0.521	0.420
GMM	0.639	0.805	0.782	0.493	0.496	0.518	0.356
LODA	0.639	0.768	0.796	0.517	0.567	0.506	0.407
LUNAR	0.633	0.822	0.819	0.461	0.642	0.373	0.388
VAE	0.605	0.754	0.754	0.411	0.435	0.459	0.290
MO-GAAL	0.594	0.690	0.791	0.453	0.546	0.402	0.344
ECOD	0.577	0.775	0.796	0.478	0.594	0.446	0.388
COPOD	0.575	0.770	0.786	0.439	0.569	0.390	0.344
PCA	0.573	0.734	0.765	0.462	0.473	0.474	0.315
LMDD	0.479	0.650	0.812	0.315	1.000	0.202	0.384
ALAD	0.465	0.704	0.735	0.386	0.399	0.409	0.234
SOS	0.368	0.583	0.719	0.307	0.338	0.289	0.128

Table 4. The experimental results for the specialized models, trained separately for each telemetry channel. The results are averaged for every model, and sorted by AUC_{PR}. The globally best results for each metric are **boldfaced**, and the best among the unsupervised algorithms are underlined.

- **Accuracy:** $(TP + TN)/(TP + TN + FP + FN)$,
- **Precision:** $TP/(TP + FP)$,
- **Recall:** $TP/(TP + FN)$,
- **F_1 score**: $(2 \cdot \text{precision} \cdot \text{recall})/(\text{precision} + \text{recall})$,
- **Matthews' Correlation Coefficient (MCC)**[22]: $(TP \cdot TN - FP \cdot FN)/\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}$,
- **Area under the receiver operating characteristic curve (AUC_{ROC})**,
- **Area under the precision-recall curve (AUC_{PR})**,

Where TP, TN, FP, and FN are the number of true positives (anomalous telemetry segments correctly identified as anomalies), true negatives (nominal telemetry segments correctly identified as nominal), false positives (nominal telemetry segments incorrectly identified as anomalies), and false negatives (anomalous telemetry segments incorrectly identified as nominal). All metrics should be maximized (↑), with one indicating the best score (MCC ranges from -1 to 1, other metrics from 0 to 1).

The baseline: anomaly detection algorithms. Although there are ground-truth AD datasets that may be used to train supervised models for this task, they are extremely limited and, by definition, they cannot capture a representative set of anomalies (otherwise such "anomalies" would not be "anomalies" any longer). In practice, while building data-driven AD algorithms for satellite telemetry, practitioners may not be able to access real-life ground-truth data, hence unsupervised methods have been gaining research attention. Here, we establish a set of baseline results obtained using 30 AD methods, including both supervised and unsupervised algorithms (Table 2).

For all those algorithms that we implemented in the PyOD framework[23] (https://pyod.readthedocs.io/en/latest/), the default parameters (suggested by the authors of these techniques) are used, with the anomaly

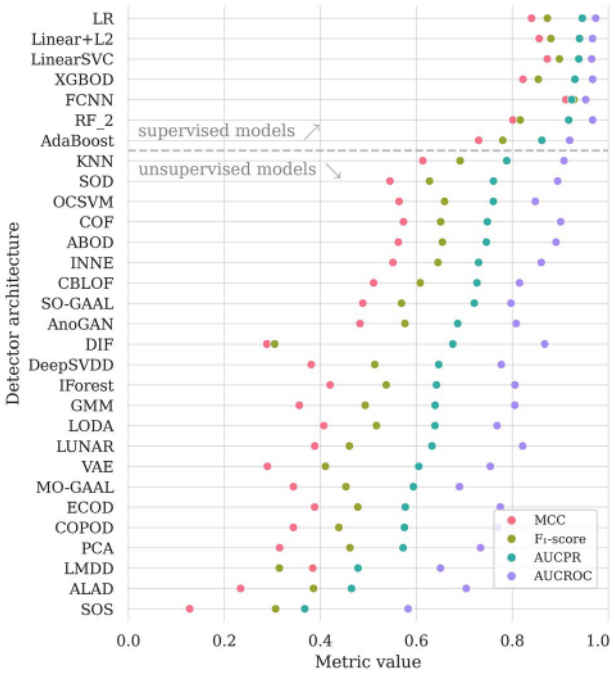


Fig. 9 The results (over Ψ) obtained using the investigated machine learning models (first grouped according to their training strategy, either supervised or unsupervised, then sorted by F_1).

contamination factor set to 0.2, according to the anomaly distribution observed in $\boldsymbol{T}$. To ensure reproducibility, we provide a Jupyter Notebook showing how to execute an example AD algorithm, in a both supervised and unsupervised training regime (*modeling_examples.ipynb*). We note that for the supervised algorithms we employed (denoted in Table 2 as AdaBoost, FCNN, Linear+L2, LR, LSVC, XGBOD, RF+ICCS) the contamination parameter is not exploited.

Data Records

We published the OPSSAT-AD benchmark dataset at Zenodo[13]. It is built of 9 source telemetry channels that were selected by the space operations engineers. They include 3 magnetometer telemetry channels: I_B_FB_MM_0 (`CADC0872`), I_B_FB_MM_1 (`CADC0873`), I_B_FB_MM_2 (`CADC0874`), and 6 photo diode (PD) channels: I_PD1_THETA (`CADC0884`), I_PD2_THETA (`CADC0886`), I_PD3_THETA (`CADC0888`), I_PD4_THETA (`CADC0890`), I_PD5_THETA (`CADC0892`), I_PD6_THETA (`CADC0894`). Here, the names correspond to the source names from the `WebMUST` repository and the OPS-SAT telemetry channel names (in brackets). The layout of the dataset is summarized in Fig. 5—it includes both the raw files, as well as the extracted features in a tabular form.

Raw telemetry data. In Fig. 6, we visualize selected characteristics of the acquired telemetry signals, effectively showing real-world challenges concerned with telemetry data acquired in the wild (e.g., missing readouts, different sampling frequencies). Such segments for all the aforementioned telemetry channels and their selected parts are included in the `data/segments.csv` file. It contains the attributes that identify the registration time: ⟨*timestamp*⟩ (ISO date format), ⟨*channel*⟩ (the channel name), ⟨*value*⟩ (the acquired signal value), and ⟨*label*⟩ (the ground-truth annotation). Additionally, we provide the consecutive segment numbers (⟨*segment*⟩), their sampling rate (⟨*sampling*⟩), and the indication if they are included in $\boldsymbol{T}$ (⟨*train*⟩).

Extracted features. In the tabular version of the dataset (`data/dataset.csv`), we include the extracted features. In the Supplementary Materials, we present the distributions of all of the provided features, rendered for both $\boldsymbol{T}$ and Ψ sets of our dataset (Fig. 3).

Technical Validation

Universal anomaly detectors. For this paper, we perform two alternative series of verifications that resemble two alternative approaches toward building anomaly detectors. We first evaluate a battery of machine learning models in a universal detector scenario, in which an anomaly detection algorithm should allow for monitoring a number of different telemetry channels using the proposed set of intermediate features. Secondly, we implemented and evaluated a set of specialized (channel-wise) detectors, for which we first prepared a separate model for each telemetry channel and aggregated their responses later.

In Table 3, we aggregate the results obtained using all the investigated AD algorithms. Here, we highlighted the globally best results (in bold for each quality metric), and we underlined the best results elaborated by the unsupervised algorithms, as we consider them a different category of the AD solutions. We are aware that some

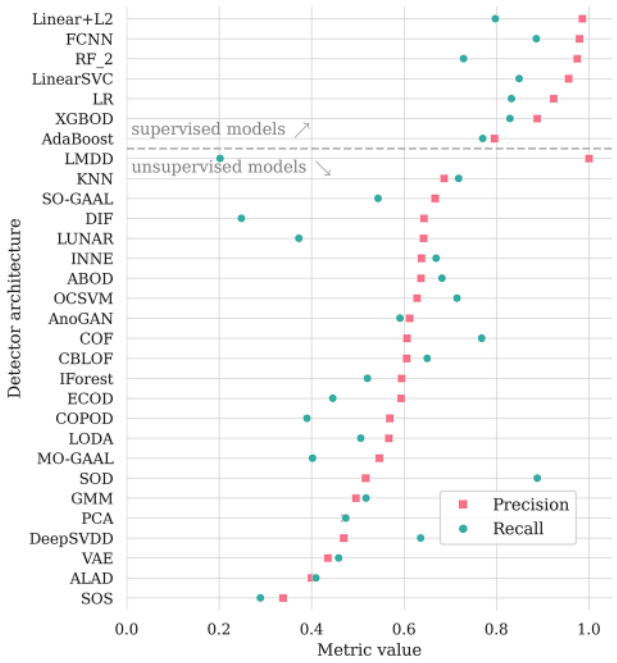


Fig. 10 Precision and recall metrics (over Ψ) obtained using the investigated algorithms. Models are sorted according to the number of misclassified telemetry segments, with the best-performing one rendered on top of the graph.

algorithms from `PyOD` should be rather trained using nominal data only (i.e., OCSVM or autoencoders) to achieve better results. As an example, the OCSVM model achieves AUC_{PR} of 0.659 and AUC_{ROC} of 0.787 in our setting, but when using only the nominal data (without abnormal segments) for training, the corresponding values are 0.762 and 0.815. However, we wanted our baseline to be consistent and to reflect a typical usage of the `PyOD` framework by a non-expert user. Also, the fine-tuning of those algorithms is out of the scope of this study. In Fig. 7, we render the selected metrics for each model. We can indeed observe the better performance of supervised methods, as those could actively benefit from the labeled anomaly examples while building a machine learning model. In Fig. 8, we also display the precision and recall quality metrics. For the fully-connected neural network, we can observe only four false positives and eight false negatives of all Ψ samples, reaching the precision of 0.963, and the recall of 0.929.

The investigation of the unsupervised algorithms reveals that some of them reach a point where they return a small set of mistakenly assessed telemetry samples. Especially the detectors built upon MO-GAAL, SO-GAAL and AnoGAN offered high precision. In terms of the number of misclassified examples, MO-GAAL obtained a better result than the supervised methods, and made one less false detection when compared to FCNN. A number of other unsupervised algorithms, however, tend to either return a large number of false negatives, or to raise many false alarms with a low number of false negatives. In the first group of such methods, we can observe DIF, ALAD, DeepSVDD, and AnoGAN, whereas e.g., COF belongs to the second group here. The usability of such algorithms would be rather limited to situations when avoiding one type of the classification error could be more practically important (e.g., to minimize the overhead induced on the space operations teams that would have to review many incorrectly raised false alarms).

Specialized anomaly detectors. To better reflect a real-life approach of building separate anomaly detectors for each telemetry channel, we follow this scenario in the second experiment. In this experiment, we utilize the same list of machine learning models, but we employed telemetry data from a narrowed list of OPS-SAT channels: `CADC0872`, `CADC0873`, `CADC0874`, `CADC0888`, `CADC0892`, and `CADC0894`. The other channels were filtered out, as they had no anomalous segments in the training or test subsets. Also, we use the same input data (*segments.csv*), but—as we perform the verification for each telemetry channel independently—the dataset which is exploited for each training and test round is a subset of the dataset used in the previous experiment. For the training sets, the number of nominal segments ranges from 95 (`CADC0894`) to 366 (`CADC0873`), and for anomalous segments from 16 (`CADC0894`) to 99 (`CADC0872`). We utilized the same data split to get the test subset, as in the previous experiments. The proportion of nominal and anomalous examples is similar, and specifically, the number of nominal segments in the test set ranges from 28 (`CADC0894`) to 122 (`CADC0873`), and for anomalous segments from 5 (`CADC0894`) to 32 (`CADC0872`).

Additionally, in this experiment, we do not utilize two features, i.e., `len` and `duration`, to make the analysis independent of the telemetry segment size. Also, we encourage the community to build new feature extractors. For this reason, the code we employed to convert the telemetry segments into a final dataset is included in the package delivered with this paper. The transformation of `segments.csv` could be reproduced, or modified, using the Jupyter notebook: `dataset_generator.ipynb`.

The results for the specialized models are on a similar performance level as those obtained for the universal models, although the ranking of the solutions is different. For this evaluation, we get the largest AUC_{PR} and AUC_{ROC} for the linear regression-based models (0.946 and 0.974, respectively). However, the best detection accuracy, F_1 score, and MCC of 0.970, 0.929, and 0.912, respectively, we noted for the fully-connected neural network, as in the previous experiment. For the unsupervised models, we found the KNN-based models with the greatest scores. We provide a detailed set of results in with the full results' table (Table 4) and the figures depicting the metrics (Figs. 9 and 10).

Usage Notes

The dataset[13] contains the data in two different forms: a set of the original telemetry segments and a corresponding set of handcrafted features, both with anomaly labels. Both collections are also encoded in the popular, easy-to-handle CSV format and are ready to to use with various machine learning models (thus, they can be considered AI-ready). All the algorithms were implemented in `Python` using `PyOD`[23] 1.1.2, `TensorFlow`[24] 2.15, and `PyTorch`[25] 2.1.2. Additionally, we used `NumPy`[26] 1.26.2 and `SciPy`[27] 1.13.1 and `Pandas`[28] 2.1.4 for data preparation, `Seaborn`[29] 0.13.0 for the visualizations, as well as `OXI`[16] for the initial data analysis and labeling processes. Finally, our benchmark is accompanied with a Jupyter Notebook, containing an example experiment (`modeling_examples.ipynb`), in order to ensure the experimental reproducibility.

Code availability

The code for working with the OPS-SAT benchmark, including the functionalities used to prepare the numerical results, figures, and tables for this article, is available through the following GitHub repository: https://github.com/kplabs-pl/OPS-SAT-AD under the MIT license.

Received: 10 July 2024; Accepted: 22 April 2025;
Published: 29 April 2025

References

1. Pang, G., Shen, C., Cao, L. & Hengel, A. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)* **54**, 1–38, https://doi.org/10.1145/3439950 (2021).
2. Hundman, K., Constantinou, V., Laporte, C., Colwell, I. & Soderstrom, T. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. *Proceedings Of The 24th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 387–395, https://doi.org/10.1145/3219819.3219845 (2018).
3. Wu, R. & Keogh, E. Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. *IEEE Transactions On Knowledge And Data Engineering* **35**, 2421–2429, https://doi.org/10.1109/TKDE.2021.3112126 (2023).
4. Wagner, D. *et al*. TimeSeAD: Benchmarking Deep Multivariate Time-Series Anomaly Detection. *Transactions On Machine Learning Research*. https://openreview.net/forum?id=iMmsCI0JsS (2023).
5. Amin Maleki Sadr, M., Zhu, Y. & Hu, P. An Anomaly Detection Method for Satellites Using Monte Carlo Dropout. *IEEE Transactions On Aerospace And Electronic Systems* **59**, 2044–2052, https://doi.org/10.1109/TAES.2022.3206257 (2023).
6. Petković, M. *et al*. Machine-learning ready data on the thermal power consumption of the Mars Express Spacecraft. *Scientific Data* **9**, 229, https://doi.org/10.1038/s41597-022-01336-z (2022).
7. Sanchez, F. *et al*. WebTCAD: A Tool for Ad-hoc Visualization and Analysis of Telemetry Data for Multiple Missions. *2018 SpaceOps Conference*. https://doi.org/10.2514/6.2018-2616 (2018).
8. Kotowski, K., Haskamp, C., Ruszczak, B., Andrzejewski, J. & Nalepa, J. Annotating large satellite telemetry dataset for ESA international AI anomaly detection benchmark. *Proceedings Of The 2023 Conference On Big Data From Space (BiDS'23) - From Foresight To Impact - 6-9 November 2023, Austrian Center, Vienna, 2023*. pp. 341–344, https://doi.org/10.2760/46796 (2023).
9. Kotowski, K. *et al*. European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry. arXiv http://arxiv.org/abs/2406.17826 (2024).
10. Evans, D. OPS-SAT: FDIR Design on a Mission that Expects Bugs - and Lots of Them. *SpaceOps 2016 Conference*. https://doi.org/10.2514/6.2016-2481 (2016).
11. Ruszczak, B. *et al*. Machine Learning Detects Anomalies in OPS-SAT Telemetry. *Computational Science - ICCS 2023*. pp. 295–306, (2023).
12. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804, https://doi.org/10.1016/j.patter.2023.100804 (2023).
13. Ruszczak, B., Kotowski, K., Nalepa, J. & Evans, D. OPSSAT-AD - anomaly detection dataset for satellite telemetry. *Zenodo*. https://doi.org/10.5281/zenodo.12588358 (2024).
14. Shendy, R. & Nalepa, J. Few-shot satellite image classification for bringing deep learning on board OPS-SAT. *Expert Systems With Applications* **251**, 123984, https://doi.org/10.1016/j.eswa.2024.123984 (2024).
15. ESA WebMUST - web client for OPS-SAT directory. (European Space Agency, 2021), https://opssat1.esoc.esa.int/webclient-must.
16. Ruszczak, B., Kotowski, K., Andrzejewski, J., Haskamp, C. & Nalepa, J. OXI: An online tool for visualization and annotation of satellite time series data. *SoftwareX* **23**, 101476, https://doi.org/10.1016/j.softx.2023.101476 (2023).
17. Nalepa, J. *et al*. Toward On-Board Detection Of Anomalous Events From OPS-SAT Telemetry Using Deep Learning. *8th International Workshop On On-Board Payload Data Compression*. https://doi.org/10.5281/zenodo.7244991 (2022).
18. Nalepa, J. *et al*. Look ma, no ground truth! On building supervised anomaly detection from OPS-SAT telemetry. *Proceedings Of The International Astronautical Congress, 2023, International Astronautical Federation*. pp. 1–8, https://dl.iafastro.directory/event/IAC-2023/paper/78668/ (2023).
19. Zhang, J. *et al*. Feature interpolation convolution for point cloud analysis. *Computers & Graphics* **99**, 182–191, https://doi.org/10.1016/j.cag.2021.06.015 (2021).
20. Lubba, C. *et al*. catch22: CAnonical Time-series CHaracteristics. *Data Mining And Knowledge Discovery* **33**, 1821–1852, https://doi.org/10.1007/s10618-019-00647-x (2019).
21. Tafazoli, S. *et al*. C22MP: the marriage of catch22 and the matrix profile creates a fast, efficient and interpretable anomaly detector. *Knowledge And Information Systems*. https://doi.org/10.1007/s10115-024-02107-5 (2024).
22. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424, https://doi.org/10.1093/bioinformatics/16.5.412 (2000).
23. Zhao, Y., Nasrullah, Z. & Li, Z. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal Of Machine Learning Research* **20**, 1–7 (2019).
24. Abadi, M. *et al*. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ (2015).

25. Paszke, A. *et al*. Automatic differentiation in PyTorch (2017).
26. Harris, C. *et al*. Array programming with NumPy. *Nature* **585**, 357–362, https://doi.org/10.1038/s41586-020-2649-2 (2020).
27. Virtanen, P. *et al*. Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272, https://doi.org/10.1038/s41592-019-0686-2 (2020).
28. The Pandas development team pandas-dev/pandas: Pandas. Zenodo https://doi.org/10.5281/zenodo.3509134 (2020).
29. Waskom, M. Seaborn: statistical data visualization. *Journal Of Open Source Software* **6**, 3021, https://doi.org/10.21105/joss.03021 (2021).
30. Grüning, M. & Kropf, S. A Ridge Classification Method for High-dimensional Observations. *From Data And Information Analysis To Knowledge Engineering*. pp. 684–691, https://doi.org/10.1007/3-540-31314-1_84 (2006).
31. Fan, R., Chang, K., Hsieh, C., Wang, X. & Lin, C. LIBLINEAR: A Library for Large Linear Classification. *Journal Of Machine Learning Research* **9**, 1871–1874 (2008).
32. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class AdaBoost. *Statistics And Its Interface* **2**, 349–360, https://doi.org/10.4310/SII.2009.v2.n3.a8 (2009).
33. Lee, C. & Lin, C. A study on l2-loss (squared hinge-loss) multiclass SVM. *Neural Comput.* **25**, 1302–1323, https://doi.org/10.1162/NECO_a_00434 (2013).
34. Zhao, Y. & Hryniewicki, M. XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. *2018 International Joint Conference On Neural Networks (IJCNN)*. pp. 1–8, https://doi.org/10.1109/IJCNN.2018.8489605 (2018).
35. Kwon, D. *et al*. A survey of deep learning-based network anomaly detection. *Cluster Computing* **22**, 949–961, https://doi.org/10.1007/s10586-017-1117-8 (2019).
36. Mastrangelo, C., Runger, G. & Montgomery, D. Statistical process monitoring with principal components. *Quality And Reliability Engineering International* **12**, 203–210 (1996).
37. Aggarwal, C. Linear Models for Outlier Detection. *Outlier Analysis*. pp. 65–110, https://doi.org/10.1007/978-3-319-47578-3_3 (2017).
38. Arning, A., Agrawal, R. & Raghavan, P. A Linear Method for Deviation Detection in Large Databases. *Knowledge Discovery And Data Mining* **1141**, 972–981 (1996).
39. Tang, J., Chen, Z., Fu, A. & Cheung, D. Enhancing Effectiveness of Outlier Detections for Low Density Patterns. *Advances In Knowledge Discovery And Data Mining*. pp. 535–548, https://doi.org/10.1007/3-540-47887-6_53 (2002).
40. Angiulli, F. & Pizzuti, C. Fast Outlier Detection in High Dimensional Spaces. *Principles Of Data Mining And Knowledge Discovery*. pp. 15–27, https://doi.org/10.1007/3-540-45681-3_2 (2002).
41. He, Z., Xu, X. & Deng, S. Discovering cluster-based local outliers. *Pattern Recognition Letters* **24**, 1641–1650, https://doi.org/10.1016/S0167-8655(03)00003-5 (2003).
42. Kriegel, H., Schubert, M. & Zimek, A. Angle-based outlier detection in high-dimensional data. *Proceedings Of The 14th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*. pp. 444–452, https://doi.org/10.1145/1401890.1401946 (2008).
43. Liu, F., Ting, K. & Zhou, Z. Isolation Forest. *2008 Eighth IEEE International Conference On Data Mining*. pp. 413–422, https://doi.org/10.1109/ICDM.2008.17 (2008).
44. Kriegel, H., Kröger, P., Schubert, E. & Zimek, A. Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data. *Advances In Knowledge Discovery And Data Mining*. pp. 831–838, https://doi.org/10.1007/978-3-642-01307-2_86 (2009).
45. Janssens, J., Huszár, F., Postma, E. & Herik, H. Stochastic outlier selection. (Tilburg University, Center for Cognition,2012)
46. Kingma, D. & Welling, M. Auto-Encoding Variational Bayes. *2014 International Conference On Learning Representations (ICLR)*. **abs/1312.6114**, https://doi.org/10.48550/arXiv.1312.6114 (2013).
47. Erfani, S., Rajasegarar, S., Karunasekera, S. & Leckie, C. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*. **58** pp. 121–134, https://www.sciencedirect.com/science/article/pii/S0031320316300267 (2016).
48. Pevný, T. Loda: Lightweight on-line detector of anomalies. *Machine Learning* **102**, 275–304, https://doi.org/10.1007/s10994-015-5521-0 (2016).
49. Schlegl, T., Seeböck, P., Waldstein, S., Schmidt-Erfurth, U. & Langs, G. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. *Information Processing In Medical Imaging*. pp. 146–157 (2017)
50. Ruff, L. *et al*. Deep One-Class Classification. *Proceedings Of The 35th International Conference On Machine Learning* **80**, 4393–4402 (2018).
51. Zenati, H., Romain, M., Foo, C., Lecouat, B. & Chandrasekhar, V. Adversarially Learned Anomaly Detection. *2018 IEEE International Conference On Data Mining (ICDM)*. pp. 727–736, https://doi.org/10.1109/ICDM.2018.00088 (2018).
52. Bandaragoda, T. *et al*. Isolation-based anomaly detection using nearest-neighbor ensembles. *Computational Intelligence* **34**, 968–998, https://doi.org/10.1111/coin.12156 (2018).
53. Liu, Y. *et al*. Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Transactions On Knowledge And Data Engineering* **32**, 1517–1528, https://doi.org/10.1109/TKDE.2019.2905606 (2020).
54. Li, Z., Zhao, Y., Botta, N., Ionescu, C. & Hu, X. COPOD: Copula-Based Outlier Detection. *2020 IEEE International Conference On Data Mining (ICDM)*. pp. 1118–1123, (2020)
55. Li, Z. *et al*. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Trans. On Knowl. And Data Eng.* **35**, 12181–12193, https://doi.org/10.1109/TKDE.2022.3159580 (2022).
56. Goodge, A., Hooi, B., Ng, S. & Ng, W. LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks. *Proceedings Of The AAAI Conference On Artificial Intelligence* **36**, 6737–6745, https://doi.org/10.1609/aaai.v36i6.20629 (2022).
57. Xu, H., Pang, G., Wang, Y. & Wang, Y. Deep Isolation Forest for Anomaly Detection. *IEEE Transactions On Knowledge And Data Engineering* **35**, 12591–12604, https://doi.org/10.1109/TKDE.2023.3270293 (2023).

Acknowledgements

This work was partially supported through the following projects: "On-board Anomaly detection from the OPS-SAT telemetry using deep learning" (40001373339/22/NL/GLC/ov) and "Few-shot anomaly detection in satellite telemetry" (4000141301) funded by the European Space Agency. BR was financially supported by the Opole University of Technology as part of the DELTA project no. 272/24. JN was supported by the Silesian University of Technology grant for maintaining and developing research potential.

Author contributions

B.R., K.K. and J.N. conceived and designed the study. B.R. and D.E. acquired the data and performed the experiments. B.R. implemented the computational pipeline and performed the analyses. B.R. drafted the manuscript. K.K. and J.N. edited and improved the manuscript. D.E. revised the manuscript. All authors have read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to B.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025, corrected publication 2025

Załącznik 10 – [A10]



Coupling Temporal Convolutional Networks with Anomaly Detection and CMA-ES for the Electricity Price Forecasting

Bogdan Ruszczak
Opole University of Technology
Opole, Poland
b.ruszczak@po.edu.pl

ABSTRACT

The process of forecasting using temporal convolutional neural networks on multivariate market data is an established research topic due to its practical applications. Although there exists a machine learning model for this task, their performance is still limited when they are deployed for highly volatile markets. In this study, we tackle this challenge through exploiting an additional module within a machine learning model that provides the auxiliary information captured from the input data and reflecting anomalous events observed in such markets. To further enhance the capabilities of the model, its hyperparameters are optimized using the covariance matrix adaptation evolution strategy. The experiments performed on the real-world electricity price quotes for the day-ahead market showed the superiority of our approach which achieved the prediction error of sMAPE of 8.19 and rMAE of 0.319, outperforming the vanilla model, which offered the sMAPE and rMAE of 17.80 and 0.693, and the baseline TCN of 16.34 and 0.603 accordingly.

CCS CONCEPTS

• **Computing methodologies** → **Search methodologies**.

KEYWORDS

hyper-parameter optimization, deep learning, temporal convolutional network, CMA-ES, electricity price forecasting

ACM Reference Format:
Bogdan Ruszczak. 2024. Coupling Temporal Convolutional Networks with Anomaly Detection and CMA-ES for the Electricity Price Forecasting. In *Genetic and Evolutionary Computation Conference (GECCO '24 Companion), July 14–18, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3638530.3664098

1 INTRODUCTION

The temporal convolutional networks (TCN) [2] have been reported as a solid forecasting method that utilizes the built-in filter fitting mechanism for the time series pattern detection. Such models can be employed to effectively predict real market quotes for the day-ahead market, for which 24 hourly predictions are required upfront [5]. However, a serious limitation of them is related to the periods when the market quotes rapidly change in unexpected ways. For such an abnormal market behavior, the TCN model performance decreases significantly. In this study, we tackle this research gap and introduce an additional procedure to address this challenge. The data, before it is fed to the model, is examined to see if it reflects standard, nominal market behavior or whether it is anomalous. This step provides the additional information to the model to help it incorporate such a detected anomaly into the prediction process. We build upon established anomaly detection (AD) methods, and focus on the improved supervised outlier detector with unsupervised representation learning [7], dubbed AD_X, and the iForest detector [3] (AD_I), that exploits randomly selected features to support unsupervised detection. An AD model is used to study each 24-hour data chunk, and its response is later passed to the prediction TCN model (Fig. 1). Ultimately, the TCN model is optimized using the covariance matrix adaptation evolution strategy (CMA-ES) [1].

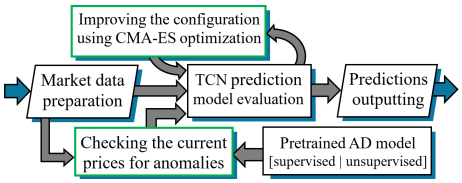


Figure 1: Flowchart of the proposed method.

2 MATERIAL AND METHODS

We benefit from the dataset introduced in [5], containing the modeled day-ahead market quotes [€/MWh] for the Polish power exchange entity. Here, we additionally need the indication of anomalies for the prices. Therefore, each 24-hour long quote time series has been carefully studied. The electricity prices are auto-correlated and exhibit daily, weekly, and yearly periodicity. When labeling the prices, we compared the data values from several past days and the prices from the same weekdays. In Fig. 2, we preview two examples, which were assigned to different categories (nominal and anomalous prices). While it would be ideal to exactly follow the previous dataset split, for the exploited dataset (435 days from January 2022 to March 2023), we observed that the interesting data samples are positioned in the middle of this period. Therefore, we cut the set after August, 1, and provide a similar number of anomalies for both subsets (Fig. 3).

The TCN model enables us to benefit from the covariates, therefore we train it with a wide range of time series, including gas (a) and CO_2 allowances (b) quotes, current solar (c), onshore wind (d) and total electricity generation (e) for the same energy system. We also use the scheduled day-ahead data on exchanges with neighboring countries and weather data (air temperature (j), wind speed (k),

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
GECCO '24 Companion, July 14–18, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0495-6/24/07.
https://doi.org/10.1145/3638530.3664098

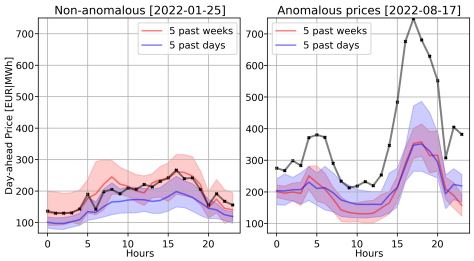


Figure 2: Two examples of electricity market behaviors. The price time-series (in black) of unusual behaviors are visible.

and solar radiation (l)) for central Poland. The features fed into the AD model were inspired by [6]. Therefore, for the raw prices and for its first and second derivatives, the mean, std, variance, number of prominent peaks, and quantiles are computed.

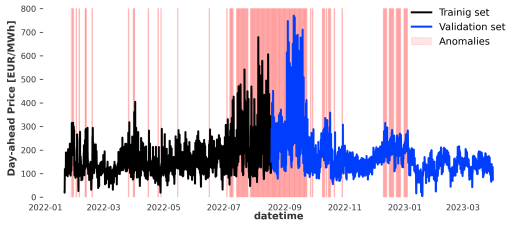


Figure 3: The dataset split and anomalous prices.

3 EXPERIMENTS

The objectives of the experiments are two-fold: to (1) assess whether the optimized model (using CMA-ES, we examined 10^3 combinations for the input data block size, the kernel size, the number of layers and convolutional filters, the dilation base, the normalization presence, and the dropout level) could provide more accurate predictions than its default configuration, and (2) to determine if the TCN model could be enhanced with the AD model. To evaluate the predictions, we employed the MAE and sMAPE metrics, accompanied by the relative metric (rMAE) which incorporates the naıve predictions (as in [4]; all metrics should be minimized). To investigate the AD model, we aggregated the receiver operating characteristic curve (ROC) and the precision-recall curve (PRC). Our techniques were implemented in Python (*Darts* framework for the forecasting, *PyOD* for the AD, and *Optuna* for optimization).

Table 1: The experimental results obtained in this study.

Experiment	*sMAPE*	*MAE*[€]	*rMAE*	*ROC*	*PRC*
Naıve	28.785	83.407	1.000	—	—
Vanilla RNN	17.799	57.848	0.693	—	—
$TCN_{baseline}$	16.344	50.403	0.603	—	—
TCN_{CMA}	10.239	31.480	0.376	—	—
$TCN_{CMA} + AD_I$	9.551	27.960	0.334	0.922	0.835
$TCN_{CMA} + AD_X$	**8.190**	**26.261**	**0.319**	**0.934**	**0.884**

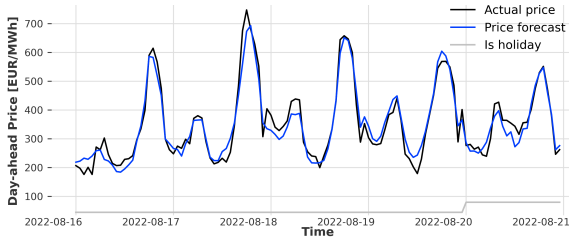


Figure 4: Predictions for the prices with detected anomalies.

Table 1 gathers the experimental results, with reference to the naıve and baseline models. The AD models significantly improved the results. The preview of the predictions for the best-performing model is shown in Fig. 4. Here, AD_X has higher-quality AD metrics and lower forecasting errors. The MAEs are better than those of baseline TCN and naıve models. The metrics lower than reported in [5] can be attributed to the different dataset split and the increased presence of high prices and anomalies in the test set.

4 CONCLUSIONS

In this study, we introduced an approach that benefits from AD to effectively respond to market changes and behaviors in the context of predicting prices. The evolutionary optimization further improved the generalization capabilities of the underlying machine learning model. Interest points of departure for further research emerge from this study. For the currently evaluated configuration, the AD module works only on univariate time series, meaning only anomalies in the price data are investigated. The multivariate approach could help to more precisely filter outlying events. Furthermore, the exploited AD method, which uses the intermediate statistical features, could be replaced with a time series-based AD method. Here, few shot learning approaches to this process should also be considered, as those steps could help to skip the weakest part in the process, that is the necessity of providing the additional expert-driven information.

ACKNOWLEDGMENTS

This work was partially supported by The National Centre for Research and Development of Poland under projects POIR.01.01.01-00-0709/21 and POIR.01.01.01-00-0855/21.

REFERENCES

[1] Y Akimoto and N Hansen. 2022. CMA-ES and Advanced Adaptation Mechanisms *(GECCO '22)*. ACM. https://doi.org/10.1145/3520304.3533648
[2] Yitian C et al. 2020. Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing* (2020). https://doi.org/10.1016/j.neucom.2020.03.011
[3] Fei Tony Liu et al. 2008. Isolation Forest. In *8th IEEE International Conf. on Data Mining*. https://doi.org/10.1109/ICDM.2008.17
[4] K Olivares et al. 2023. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *Int. J. of Forec.* 39 (2023). https://doi.org/10.1016/j.ijforecast.2022.03.001
[5] B. Ruszczak. 2023. Evolutionary configuration of the Temporal Convolutional Network for the electricity price prediction *(GECCO '23)*. ACM. https://doi.org/10.1145/3583133.3596429
[6] B. Ruszczak et al. 2023. Machine Learning Detects Anomalies in OPS-SAT Telemetry *(ICCS '23)*. Springer Nature. https://doi.org/10.1007/978-3-031-35995-8_21
[7] Y. Zhao et al. 2018. XGBOD: Improving Supervised Outlier Det. with Unsupervised Representation Learning *(IJCNN '18)*. https://doi.org/10.1109/IJCNN.2018.8489605

Załącznik 11 – Oświadczenia współautorów

1. oświadczenie współautora – mgr inż. Jacek Andrzejewski,
2. oświadczenie współautora – Mr. Sam Bammens,
3. oświadczenie współautora – Zineb Bougriche, MSc.
4. oświadczenie współautora – mgr inż. Karol Dziubański,
5. oświadczenie współautora – Mr. David Evans,
6. oświadczenie współautora – Christoph Haskamp, MSc,
7. oświadczenie współautora – dr inż. Krzysztof Kotowski,
8. oświadczenie współautora – dr Sabina Kubiciel–Lodzińska,
9. oświadczenie współautora – Mr. Rodrigo Laurinovičs,
10. oświadczenie współautora – mgr inż. Alicja Musiał,
11. oświadczenie współautora – dr hab. inż. Jakub Nalepa, prof. ucz.,
12. oświadczenie współautora – Associate Professor, Dr. Pedro Javier Navarro Lorente,
13. oświadczenie współautora – dr inż. Katarzyna Rudnik,
14. oświadczenie współautora – dr inż. Krzysztof Smykała,
15. oświadczenie współautora – dr hab. inż. Michał Tomaszewski, prof. ucz.,
16. oświadczenie współautora – dr inż. Agata M. Wijata,
17. oświadczenie współautora – Vladimir Zelenevskiy, MSc.

Opole, dn. 20.01.2025

Mgr inż. Jacek Andrzejewski
KP Labs sp z o.o.
Gliwice

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Mgr inż. Jacek Andrzejewski

Dotyczy publikacji:

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics L. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21
Indywidualny wkład B. Ruszczaka: implementacja nadzorowanych detektorów anomalii, zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych, opracowanie metody generowania pośrednich syntetycznych cech danych i doboru cech modelu detektora, prezentacja pracy na konferencji międzynarodowej, współautorstwo artykułu.

Ruszczak B. (30%), Kotowski K. (20%), Andrzejewski J. (20%), Haskamp C. (10%), Nalepa J. (20%), OXI: an online tool for visualization and annotation of satellite time series data. SoftwareX, Vol. 23:101476, 202, ISSN: 2352-7110, DOI: 10.1016/j.softx.2023.101476
Indywidualny wkład B. Ruszczaka: opracowanie testów metody, współautorstwo założeń metody, implementacji, wizualizacji i finalnej wersji artykułu.

Jacek Andrzejewski

Darmstadt, 15.05.2025

Mr Sam Bammens
European Space Agency/ESOC,
Darmstadt, Germany

Co-author statement

I hereby declare, that I have no objections to the percentages and contribution descriptions of our joint publications included in the statement of research, teaching and organizational achievements of **dr Bogdan Ruszczak**, prepared for the purpose of the Bogdan Ruszczak's habilitation procedure (postdoctoral degree).

Mr Sam Bammens

Applies to publication:

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics L. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21

Individual contribution of B. Ruszczak: supervised anomaly detectors implementation, algorithm evaluation procedure development, conducting and evaluation of experiments, results visualization, dataset development, method of intermediate telemetry segments features generation, development, and matching the detector features, presenting the paper at the ICCS conference, and manuscript co-writing.

Sam Bam

Opole, 08.05.2025

MSc Zineb Bougriche
Katedra Informatyki,
Wydział Elektrotechniki, Automatyki i Informatyki,
Politechnika Opolska

Co-author statement

I hereby declare, that I have no objections to the percentages and contribution descriptions of our joint publications included in the statement of research, teaching and organizational achievements of **dr Bogdan Ruszczak**, prepared for the purpose of the Bogdan Ruszczak's habilitation procedure (postdoctoral degree).

MSc Zineb Bougriche



Applies to publication:

Ruszczak B. (60%), Kubiciel-Lodzińska S. (2.5%), Rudnik K. (5%), Bougriche Z. (2.5%), Wijata A. M. (10%), Nalepa J. (20%), Single- and multi- objective genetic optimization of hyperspectral image preprocessing, Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Malaga, Spain, ACM, 2025, DOI: 10.1145/3712255.3726774

Individual contribution of B. Ruszczak: development of the genetic algorithms—both single- and multi-objective—for optimizing pre-processing in multi-target estimation, algorithm evaluation procedure development, conducting and evaluation of experiments, results visualization, dataset development and curation, supervised chlorophyll content regressors implementation, presenting the paper at the GECCO conference, and manuscript co-writing.

Opole, dn. 23.08.2024

Mgr inż. Karol Dziubański

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Mgr inż. Karol Dziubański

Karol Dziubański

Dotyczy publikacji:

Ruszczak B. (70%), Smykała K. (20%), Dziubański K. (10%), The detection of alternaria solani infection on tomatoes using ensemble learning. Journal of Ambient Intelligence and Smart Environments, 12:407–41?, 2020, DOI: 10.3233/AIS-200573

Indywidualny wkład B. Ruszczaka: opracowanie koncepcji badania, manuskryptu, implementacja algorytmów obliczeniowych, zgłoszenie artykułu i odpowiedzi na recenzje, współudział w procesie gromadzenia i przetwarzania danych, analizie wyników, poprawianiu manuskryptu, i pozyskaniu finansowania na publikację.

Darmstadt, 05.05.2025

Mr David Evans
Advanced Operations and OPS-SAT Space Lab Manager,
European Space Agency/ESOC,
Darmstadt, Germany

Co-author statement

I hereby declare, that I have no objections to the percentages and contribution descriptions of our joint publications included in the statement of research, teaching and organizational achievements of **dr Bogdan Ruszczak**, prepared for the purpose of the Bogdan Ruszczak's habilitation procedure (postdoctoral degree).

Mr David Evans

David Evans
Digitally signed by David Evans
Date: 2025.05.06 08:32:45 +02'00'

Applies to publication:

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics L. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21
Individual contribution of B. Ruszczak: supervised anomaly detectors implementation, algorithm evaluation procedure development, conducting and evaluation of experiments, results visualization, dataset development, method of intermediate telemetry segments features generation, development, and matching the detector features, presenting the paper at the ICCS conference, and manuscript co-writing.

Ruszczak B. (50%), Nalepa J. (20%), Kotowski K. (20%), Evans D. (10%), The OPS-SAT benchmark for detecting anomalies in satellite telemetry. Scientific Data, vol. 12, no. 1, pp.1-14, Article number:710, Springer, 2025, DOI: 10.1038/s41597-025-05035-3
Individual contribution of B. Ruszczak: supervised and unsupervised algorithms implementation, algorithm evaluation procedure development, conducting and evaluation of experiments, results visualization, dataset development and preparation for publication, early draft of the manuscript writing and final manuscript writing, manuscript review and editing, and correspondence to the journal and the reviewers.

Bremen, 05.05.2025

Mr Christoph Haskamp
AI Expert, Airbus Defence and Space GmbH,
Bremen, Germany

Co-author statement

I hereby declare, that I have no objections to the percentages and contribution descriptions of our joint publications included in the statement of research, teaching and organizational achievements of **dr Bogdan Ruszczak**, prepared for the purpose of the Bogdan Ruszczak's habilitation procedure (postdoctoral degree).

Mr Christoph Haskamp

Digital signiert von: Christoph Haskamp
Name: CN = Christoph Haskamp email = christoph.haskamp@airbus.com C = DE O = Airbus Defence and Space GmbH OU = TESOG
Datum: 2025.05.05 13:06:07 +02'00'

Applies to publication:

Ruszczak B. (30%), Kotowski K. (20%), Andrzejewski J. (20%), Haskamp C. (10%), Nalepa J. (20%), OXI: an online tool for visualization and annotation of satellite time series data. SoftwareX, Vol. 23:101476, 202, ISSN: 2352-7110, DOI: 10.1016/j.softx.2023.101476

Individual contribution of B. Ruszczak: method tests development, co-authorship for method framing and development, as well as for paper visualisations and final version of the manuscript.

Opole, dn. 08.05.2025

Dr inż. Krzysztof Kotowski
KP Labs sp z o.o.
Gliwice

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Dr inż. Krzysztof Kotowski

Dotyczy publikacji:

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics R. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21
Indywidualny wkład B. Ruszczaka: implementacja nadzorowanych detektorów anomalii, zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych, opracowanie metody generowania pośrednich syntetycznych cech danych i doboru cech modelu detektora, prezentacja pracy na konferencji międzynarodowej, współautorstwo artykułu.

Ruszczak B. (30%), Kotowski K. (20%), Andrzejewski J. (20%), Haskamp C. (10%), Nalepa J. (20%), OXI: an online tool for visualization and annotation of satellite time series data. SoftwareX, Vol. 23:101476, 202, ISSN: 2352-7110, DOI: 10.1016/j.softx.2023.101476
Indywidualny wkład B. Ruszczaka: opracowanie testów metody, współautorstwo założeń metody, implementacji, wizualizacji i finalnej wersji artykułu.

Ruszczak B. (50%), Nalepa J. (20%), Kotowski K. (20%), Evans D. (10%),The OPS-SAT benchmark for detecting anomalies in satellite telemetry. Scientific Data, vol. 12, no. 1, pp.1-14, Article number:710, Springer, 2025, DOI: 10.1038/s41597-025-05035-3
Indywidualny wkład B. Ruszczaka: implementacja algorytmów nienadzorowanych i nadzorowanych detektorów anomalii, zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych i przygotowanie ich publikacji, przygotowanie wstępnej wersji artykułu i współautorstwo finalnej wersji artykułu, oraz zgłoszenie artykułu i odpowiedzi na recenzje.

Krzysztof Kotowski

Opole, dn. 09.05.2025

Dr Sabina Kubiciel-Lodzińska
Katedra Polityki Regionalnej i Rynku Pracy,
Wydział Ekonomii i Zarządzania,
Politechnika Opolska

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Dr Sabina Kubiciel-Lodzińska

Dotyczy publikacji:

Ruszczak B. (60%), Kubiciel-Lodzińska S. (2.5%), Rudnik K. (5%), Bougriche Z. (2.5%), Wijata A. M. (10%), Nalepa J. (20%), Single- and multi- objective genetic optimization of hyperspectral image preprocessing, Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Malaga, Hiszpania, ACM, 2025, DOI: 10.1145/3712255.3726774

Indywidualny wkład B. Ruszczaka: opracowanie metody preprocesingu obrazów hoperspektralnych z użyciem algorytmów genetycznych dla jedno i wielo kryterialnej optymalizacji, opracowanie procedury ewaluzacji algorytmów, wykonanie i ocena eksperymentów, wizualizacja wyników, przygotowanie, publikacja i utrzymanie zbioru danych, implementacja algorytmów uczenia nadzorowanego dla zadania regresji zawartości chlorofilu, prezentacja artykułu na konferencji GECCO, współautorstwo i edycja artykułu.

Ventspils, 06.05.2025

Mr Rodrigo Laurinovičs
Ventspils International Radio Astronomy Centre
Ventspils University of Applied Sciences

Co-author statement

I hereby declare, that I have no objections to the percentages and contribution descriptions of our joint publications included in the statement of research, teaching and organizational achievements of **dr Bogdan Ruszczak**, prepared for the purpose of the Bogdan Ruszczak's habilitation procedure (postdoctoral degree).

Mr Rodrigo Laurinovičs

Applies to publication:

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics L. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21

Individual contribution of B. Ruszczak: supervised anomaly detectors implementation, algorithm evaluation procedure development, conducting and evaluation of experiments, results visualization, dataset development, method of intermediate telemetry segments features generation, development, and matching the detector features, presenting the paper at the ICCS conference, and manuscript co-writing.

Opole, dn. 20.01.2025

Mgr Alicja Musiał
KP Labs sp z o.o.
Gliwice

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Mgr Alicja Musiał

Dotyczy publikacji:

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics L. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21

Indywidualny wkład B. Ruszczaka: implementacja nadzorowanych detektorów anomalii, zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych, opracowanie metody generowania pośrednich syntetycznych cech danych i doboru cech modelu detektora, prezentacja pracy na konferencji międzynarodowej, współautorstwo artykułu.

Opole, dn. 12.11.2025

Dr hab. inż. Jakub Nalepa, prof. ucz.
Katedra Algorytmiki i Oprogramowania,
Wydział Automatyki, Elektroniki i Informatyki,
Politechnika Śląska

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.



Dr hab. inż. Jakub Nalepa, prof. ucz.

Dotyczy publikacji:

Ruszczak B. (60%), Wijata A. M. (15%), Nalepa J. (25%), Unbiasing the estimation of chlorophyll from hyperspectral images: A benchmark dataset, validation procedure and baseline results. Remote Sensing, 14(21), 2022, DOI: 10.3390/rs14215526
Indywidualny wkład B. Ruszczaka: opracowanie zbioru danych, przygotowanie i oprogramowanie metod i modeli oraz wizualizacji danych, autorstwo wstępnego artykułu oraz współautorstwo końcowej wersji, formalnej analizy oraz przeglądu literatury.

Ruszczak B. (60%), Wijata A. M. (15%), Nalepa J. (25%), Estimating chlorophyll content from hyperspectral data using gradient features. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part II, wolumen 14074 Lecture Notes in Computer Science, Springer, 2023, DOI: 10.1007/978-3-031-36021-3_18
Indywidualny wkład B. Ruszczaka: opracowanie zbioru danych, przygotowanie i oprogramowanie metod i modeli oraz wizualizacji danych, wykonanie i ewaluacja eksperymentów, propozycja i imlementacja metody automatycznego preprocesingu z użyciem filtrowania Savitzkiego-Golaya, autorstwo wstępnego artykułu oraz współautorstwo końcowej wersji artykułu, formalnej analizy oraz przeglądu literatury, zaprezentowanie pracy na konferencji międzynarodowej.

Ruszczak B. (30%), Kotowski K. (20%), Andrzejewski J. (20%), Haskamp C. (10%), Nalepa J. (20%), OXI: an online tool for visualization and annotation of satellite time series data. SoftwareX, Vol. 23:101476, 202, ISSN: 2352-7110, DOI: 10.1016/j.softx.2023.101476
Indywidualny wkład B. Ruszczaka: opracowanie testów metody, współautorstwo założeń metody, implementacji, wizualizacji i finalnej wersji artykułu.

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics R. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21
Indywidualny wkład B. Ruszczaka: implementacja nadzorowanych detektorów anomalii, zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych, opracowanie metody generowania pośrednich syntetycznych cech danych i doboru cech modelu detektora, prezentacja pracy na konferencji międzynarodowej, współautorstwo artykułu.

Ruszczak B. (50%), Nalepa J. (20%), Kotowski K. (20%), Evans D. (10%),The OPS-SAT benchmark for

detecting anomalies in satellite telemetry. Scientific Data, vol. 12, no. 1, pp.1-14, Article number:710, Springer, 2025, DOI: 10.1038/s41597-025-05035-3

Indywidualny wkład B. Ruszczaka: implementacja algorytmów nienadzorowanych i nadzorowanych detektorów anomalii, zaproponowanie procedury ewaluacji algorytmów, wykonanie i ewaluacja eksperymentów, wizualizacja wyników, opracowanie zbioru danych i przygotowanie ich publikacji, przygotowanie wstępnej wersji artykułu i współautorstwo finalnej wersji artykułu, oraz zgłoszenie artykułu i odpowiedzi na recenzje.

Ruszczak B. (60%), Kubiciel-Lodzińska S. (2.5%), Rudnik K. (5%), Bougriche Z. (2.5%), Wijata A. M. (10%), Nalepa J. (20%), Single- and multi- objective genetic optimization of hyperspectral image preprocessing, Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Malaga, Hiszpania, ACM, 2025, DOI: 10.1145/3712255.3726774

Indywidualny wkład B. Ruszczaka: opracowanie metody preprocesingu obrazów hoperspektralnych z użyciem algorytmów genetycznych dla jedno i wielo kryterialnej optymalizacji, opracowanie procedury ewaluacji algorytmów, wykonanie i ocena eksperymentów, wizualizacja wyników, przygotowanie, publikacja i utrzymanie zbioru danych, implementacja algorytmów uczenia nadzorowanego dla zadania regresji zawartości chlorofilu, prezentacja artykułu na konferencji GECCO, współautorstwo i edycja artykułu.

Cartagena, d. 10.06.2024

Prof dr Pedro Javier Navarro Lorente
Escuela Técnica Superior de Ingeniería de Telecomunicación (DSIE),
Campus Muralla del Mar,
Universidad Politécnica de Cartagena

Co-author statement

I hereby declare, that I have no objections to the percentages and contribution descriptions of our joint publications included in the statement of research, teaching and organizational achievements of **dr Bogdan Ruszczak**, prepared for the purpose of the Bogdan Ruszczak's habilitation procedure (postdoctoral degree).

Prof dr Pedro Javier Navarro Lorente

PEDRO JAVIER| NAVARRO| LORENTE

Firmado digitalmente por PEDRO JAVIER| NAVARRO| LORENTE

Applies to publication:

Ruszczak B. (40%), Smykała, K. (25%), Tomaszewski, M. (20%), Navarro Lorente P. J. (15%), Various tomato infection discrimination using spectroscopy. Signal, Image and Video Processing, May 2024, DOI: 10.1007/s11760-024-03247-5

Individual contribution of B. Ruszczak: participation in study conceptualization, methodology, computational algorithms implementation, results interpretation, original draft writing, manuscript review and editing, and correspondence to the journal and the reviewers.

Opole, dn. 09.05.2025

Dr inż. Katarzyna Rudnik
Katedra Zarządzania i Inżynierii Produkcji,
Wydział Inżynierii Produkcji i Logistyki,
Politechnika Opolska

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Dr inż. Katarzyna Rudnik

Dotyczy publikacji:

Ruszczak B. (60%), Kubiciel-Lodzińska S. (2.5%), Rudnik K. (5%), Bougriche Z. (2.5%), Wijata A. M. (10%), Nalepa J. (20%), Single- and multi- objective genetic optimization of hyperspectral image preprocessing, Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Malaga, Hiszpania, ACM, 2025, DOI: 10.1145/3712255.3726774

Indywidualny wkład B. Ruszczaka: opracowanie metody preprocesingu obrazów hoperspektralnych z użyciem algorytmów genetycznych dla jedno i wielo kryterialnej optymalizacji, opracowanie procedury ewaluzacji algorytmów, wykonanie i ocena eksperymentów, wizualizacja wyników, przygotowanie, publikacja i utrzymanie zbioru danych, implementacja algorytmów uczenia nadzorowanego dla zadania regresji zawartości chlorofilu, prezentacja artykułu na konferencji GECCO, współautorstwo i edycja artykułu.

Opole, dn. 05.06.2024

Dr inż. Krzysztof Smykała
Katedra Informatyki,
Wydział Elektrotechniki, Automatyki i Informatyki,
Politechnika Opolska

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Dr inż. Krzysztof Smykała

Dotyczy publikacji:

Ruszczak B. (70%), Smykała K. (20%), Dziubański K. (10%), The detection of alternaria solani infection on tomatoes using ensemble learning. Journal of Ambient Intelligence and Smart Environments, 12:407–418, 2020, DOI: 10.3233/AIS-200573

Indywidualny wkład B. Ruszczaka: opracowanie koncepcji badania, manuskryptu, implementacja algorytmów obliczeniowych, zgłoszenie artykułu i odpowiedzi na recenzje, współudział w procesie gromadzenia i przetwarzania danych, analizie wyników, poprawianiu manuskryptu, i pozyskaniu finansowania na publikację.

Ruszczak B. (40%), Smykała, K. (25%), Tomaszewski, M. (20%), Navarro Lorente P. J. (15%), Various tomato infection discrimination using spectroscopy. Signal, Image and Video Processing, May 2024, DOI: 10.1007/s11760-024-03247-5

Indywidualny wkład B. Ruszczaka: współpraca przy opracowaniu koncepcji badania, opracowaniu metodologii, implementacji algorytmów obliczeniowych, współudział w analizie wyników, pisaniu i poprawianiu manuskryptu, oraz zgłoszenie artykułu i odpowiedzi na recenzje.

Opole, dn. 04.06.2024

Dr hab. inż. Michał Tomaszewski, prof. ucz.
Katedra Informatyki,
Wydział Elektrotechniki, Automatyki i Informatyki,
Politechnika Opolska

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Dr hab. inż. Michał Tomaszewski, prof. ucz.

Dotyczy publikacji:

Ruszczak B. (40%), Smykała, K. (25%), Tomaszewski, M. (20%), Navarro Lorente P. J.. (15%), Various tomato infection discrimination using spectroscopy. Signal, Image and Video Processing, Ma / 2024, DOI: 10.1007/s11760-024-03247-5

Indywidualny wkład B. Ruszczaka: współpraca przy opracowaniu koncepcji badania, opracow / iu metodologii, implementacji algorytmów obliczeniowych, współudział w analizie wyników, pisaniu i poprawianiu / anuskryptu, oraz zgłoszenie artykułu i odpowiedzi na recenzje.

Opole, dn. 20.05.2025

Dr inż. Agata M. Wijata
Katedra Informatyki Medycznej i Sztucznej Inteligencji,
Wydział Inżynierii Biomedycznej,
Politechnika Śląska

Oświadczenie współautora

Oświadczam, że nie zgłaszam zastrzeżeń do udziałów procentowych i opisu wkładu merytorycznego we wspólnych publikacjach zawartych w zestawieniu dorobku naukowo-badawczego, dydaktycznego i organizacyjnego **dr. inż. Bogdana Ruszczaka** sporządzonym na potrzeby procedury o nadanie tytułu doktora habilitowanego.

Dr inż. Agata M. Wijata

Dotyczy publikacji:

Ruszczak B. (60%), Wijata A. M. (15%), Nalepa J. (25%), Estimating chlorophyll content from hyperspectral data using gradient features. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part II, wolumen 14074 Lecture Notes in Computer Science, Springer, 2023, DOI: 10.1007/978-3-031-36021-3_18
Indywidualny wkład B. Ruszczaka: opracowanie zbioru danych, przygotowanie i oprogramowanie metod i modeli oraz wizualizacji danych, wykonanie i ewaluacja eksperymentów, propozycja i imlementacja metody automatycznego preprocesingu z użyciem filtrowania Savitzkiego-Golaya, autorstwo wstępnego artykułu oraz współautorstwo końcowej wersji artykułu, formalnej analizy oraz przeglądu literatury, zaprezentowanie pracy na konferencji międzynarodowej.

Ruszczak B. (60%), Wijata A. M. (15%), Nalepa J. (25%), Unbiasing the estimation of chlorophyll from hyperspectral images: A benchmark dataset, validation procedure and baseline results. Remote Sensing, 14(21), 2022, DOI: 10.3390/rs14215526
Indywidualny wkład B. Ruszczaka: opracowanie zbioru danych, przygotowanie i oprogramowanie metod i modeli oraz wizualizacji danych, autorstwo wstępnego artykułu oraz współautorstwo końcowej wersji, formalnej analizy oraz przeglądu literatury.

Ruszczak B. (60%), Kubiciel-Lodzińska S. (2.5%), Rudnik K. (5%), Bougriche Z. (2.5%), Wijata A. M. (10%), Nalepa J. (20%), Single- and multi- objective genetic optimization of hyperspectral image preprocessing, Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO, Malaga, Hiszpania, ACM, 2025, DOI: 10.1145/3712255.3726774
Indywidualny wkład B. Ruszczaka: opracowanie metody preprocesingu obrazów hoperspektralnych z użyciem algorytmów genetycznych dla jedno i wielo kryterialnej optymalizacji, opracowanie procedury ewaluacji algorytmów, wykonanie i ocena eksperymentów, wizualizacja wyników, przygotowanie, publikacja i utrzymanie zbioru danych, implementacja algorytmów uczenia nadzorowanego dla zadania regresji zawartości chlorofilu, prezentacja artykułu na konferencji GECCO, współautorstwo i edycja artykułu.

Oxford, 05.05.2025

Mr Vladimir Zelenevskiy
Spacecraft Operations Engineer,
Astroscale, Oxford, United Kingdom

Co-author statement

I hereby declare, that I have no objections to the percentages and contribution descriptions of our joint publications included in the statement of research, teaching and organizational achievements of **dr Bogdan Ruszczak**, prepared for the purpose of the Bogdan Ruszczak's habilitation procedure (postdoctoral degree).

Mr Vladimir Zelenevskiy



Applies to publication:

Ruszczak B. (50%), Kotowski K. (20%), Andrzejewski J. (5%), Musiał A. (1%), Evans D. (1%), Zelenevskiy V. (1%), Bammens S. (1%), Laurinovics L. (1%), Nalepa J. (20%), Machine learning detects anomalies in OPS-SAT telemetry. J. Mikyska, C. Mulatier, M. Paszynski, V. Krzhizhanovskaya, J. Dongarra, P. Sloot, red., Computational Science - ICCS 2023 - 23rd International Conference, Prague, July 3-5, 2023, Proceedings, Part I, wolumen 14073 serii Lecture Notes in Computer Science, strony 295–306. Springer, 2023, DOI: 10.1007/978-3-031-35995-8_21

Individual contribution of B. Ruszczak: supervised anomaly detectors implementation, algorithm evaluation procedure development, conducting and evaluation of experiments, results visualization, dataset development, method of intermediate telemetry segments features generation, development, and matching the detector features, presenting the paper at the ICCS conference, and manuscript co-writing.