



Kierunek studiów: Geoinformacja

Nr albumu: 461791

Filip Ratajszczak

**Wykrywanie farm fotowoltaicznych na podstawie
danych teledetekcyjnych**

*Detection of photovoltaic farms based on remote
sensing data*

Praca inżynierska napisana
w Instytucie Geoekologii i Geoinformacji
pod kierunkiem
dr. hab. Jakuba Nowosada

Streszczenie

Abstrakt

Energia pozyskiwana z odnawialnych źródeł odgrywa kluczową rolę w ograniczaniu emisji gazów cieplarnianych do atmosfery. Szybki wzrost liczby wielkoskalowych instalacji fotowoltaicznych, czyli farm fotowoltaicznych, powoduje konieczność monitorowania ich ilości i lokalizacji w celu analizy postępów w tym sektorze energii odnawialnej. Celem pracy było określenie optymalnych metod detekcji farm fotowoltaicznych na podstawie danych teledetekcyjnych. W procesie detekcji wykorzystano dane radarowe z misji Sentinel-1 oraz dane multispektralne z misji Sentinel-2, wraz z ich pochodnymi, takimi jak wskaźniki teledetekcyjne i tekstury obrazu. Na podstawie tych zmiennych dokonano uczenia i walidacji kilku modeli lasów losowych (ang. *Random Forest*) w celu wskazania optymalnej konfiguracji predyktorów. Uzyskane wyniki potwierdziły możliwość detekcji farm fotowoltaicznych na podstawie danych teledetekcyjnych. Najlepsze rezultaty w końcowej klasyfikacji po procesie przetwarzania końcowego osiągnął model oparty na kanałach Sentinel-2 oraz wykorzystanych wskaźnikach teledetekcyjnych (NDVI, NDBI i mNDWI). Po przetwarzaniu końcowym model ten wykazał precyję na poziomie 87% i czułość na poziomie 84%, co przełożyło się na ostateczny wynik 85% dla miary F1-score. Dodatkowo badanie wykazało, że dane radarowe z misji Sentinel-1 i ich pochodne nie były istotne w procesie detekcji farm fotowoltaicznych. Wyniki badania mogą stanowić podstawę dla przyszłych prac nad monitorowaniem rozwoju farm fotowoltaicznych w Polsce.

Słowa kluczowe: odnawialne źródła energii, dane satelitarne, pokrycie terenu i użytkowania ziemi, uczenie maszynowe, lasy losowe, klasyfikacja obrazu

Abstract

The abstract must be consistent with the above text.

Keywords: (as stated before)

Spis treści

Streszczenie	3
1 Wprowadzenie	7
2 Materiały	11
2.1 Lokalizacje farm fotowoltaicznych	12
2.2 Zdjęcia satelitarne	16
3 Metody	21
3.1 Przygotowanie danych	22
3.2 Próbki treningowe i testowe	29
3.3 Uczenie maszynowe	30
3.4 Ważność zmiennych	36
3.5 Przetwarzanie końcowe	39
3.6 Oprogramowanie	40
4 Wyniki	43
4.1 Ocena jakości modeli	43
4.2 Ważność zmiennych	46
4.3 Wyniki klasyfikacji po przetwarzaniu końcowym	49
5 Podsumowanie	57
Bibliografia	61

Rozdział 1

Wprowadzenie

Energia elektryczna odgrywa kluczową rolę w życiu większości ludzi, stanowiąc podstawę funkcjonowania społeczeństwa i napędzając rozwój gospodarczy [ref]. Istnieje wiele metod pozyskiwania energii elektrycznej, włączając w to tradycyjne, konwencjonalne sposoby oparte na spalaniu paliw kopalnych, a także energię atomową oraz odnawialne źródła energii (OZE). Odnawialne źródła energii, określane inaczej jako zielona energia, wykorzystują naturalne procesy, takie jak promieniowanie słoneczne, siła wiatru czy ciepło wnętrza Ziemi, w celu generowania energii elektrycznej.

Zasoby surowców wykorzystywanych do produkcji energii konwencjonalnej są ograniczone. Statystyczny przegląd światowej energetyki (*Statistical Review of World Energy*) opracowany przez firmę BP w 2021 roku wskazuje, że obecne zasoby paliw kopalnych, takich jak ropa naftowa, gaz ziemny i węgiel, wystarczą odpowiednio na 50, 48,8 i 139 lat obecnej produkcji (BP, 2021). Dodatkowo, rosnące zainteresowanie ochroną środowiska i negatywnymi skutkami zmian klimatu oraz wprowadzane przepisy prawne zobowiązują do podejmowania działań mających na celu ograniczenie emisji szkodliwych gazów do atmosfery. W obliczu globalnych zmian klimatycznych i kryzysu energetycznego, spowodowanego między innymi inwazją Rosji na Ukrainę w 2022 roku, społeczeństwo jest zmuszone do przejścia z produkcji energii opartej na paliwach kopalnych na rzecz odnawialnych źródeł energii i energetyki jądrowej [ref].

ROZDZIAŁ 1. WPROWADZENIE

Ambitne cele klimatyczne Unii Europejskiej, których podstawę stanowi Europejski Zielony Ład (ang. *European Green Deal*), sprawiają, że energia elektryczna pochodząca z odnawialnych źródeł energii zaczyna odgrywać kluczową rolę w procesie transformacji energetycznej Unii Europejskiej. Europejski Zielony Ład to zestaw inicjatyw politycznych mających na celu przeprowadzenie transformacji ekologicznej i osiągnięcie neutralności klimatycznej w UE do 2050 roku (Rada Europejska, Rada Unii Europejskiej, 2023b). Inicjatywy takie jak Gotowi na 55 (ang. *Fit For 55*) oraz rozporządzenie o europejskim prawie klimatycznym mają zapewnić redukcję emisji gazów cieplarnianych o co najmniej 55% do 2030 roku (Rada Europejska, Rada Unii Europejskiej, 2023b). Jednocześnie, zgodnie z ustaleniami Rady Unii Europejskiej, Parlamentu Europejskiego oraz Komisji Europejskiej, udział energii odnawialnej w ogólnym zużyciu energii w Unii Europejskiej ma wynosić 42,5% do 2030 roku (Rada Europejska, Rada Unii Europejskiej, 2023a). Inną inicjatywą organów Unii Europejskiej jest plan REPowerEU, czyli unijny plan transformacji energetycznej i uniezależnienia od rosyjskich paliw kopalnych przedstawiony przez Komisję Europejską w maju 2022 roku (Komisja Europejska, 2022). Przedsięwzięcie to skupia się na inwestycjach w odnawialne źródła energii, zakładając uruchomienie niemal 600 GW mocy nowych instalacji fotowoltaicznych w całej Unii Europejskiej do 2030 roku (Komisja Europejska, 2022). Działania te uznaje się za kluczowe z punktu widzenia interesu publicznego.

Osiągnięcie wymienionych celów będzie możliwe jedynie poprzez dokładne monitorowanie wyników działań oraz posiadanie aktualnych informacji dotyczących odnawialnych źródeł energii. Transformacja energetyczna jest dynamicznym procesem, którego monitorowanie będzie miało znaczący wpływ na rozwój kolejnych projektów z zakresu energetyki odnawialnej. Niejednolita organizacja działania władz samorządowych w Polsce oraz występowanie na terenie kraju różnych operatorów energetycznych utrudnia jednak śledzenie rozwoju polskiej branży OZE. W oficjalnych, dostępnych w Polsce źródłach danych brakuje aktualnych informacji na temat lokalizacji elektrowni słonecznych. Wśród nieoficjalnych źródeł danych informacje na ten temat można znaleźć m.in. w projekcie społeczności internetowej OpenStreetMap (OSM contributors, 2024), gdzie są one oznaczone jako tag:plant:source=solar, jednak dane te są nieformalne, niepełne

i zazwyczaj nieaktualne. W związku z brakiem jednolitych danych na temat lokalizacji farm fotowoltaicznych w oficjalnych źródłach w Polsce, wykorzystanie danych satelitarnych staje się potencjalnym narzędziem do ich detekcji, umożliwiając pozyskiwanie aktualnych informacji na ten temat. Obserwacje satelitarne pozwalają również regularnie monitorować rozwój farm fotowoltaicznych, co umożliwia analizę postępów w tym sektorze energetyki odnawialnej.

Na przestrzeni ostatnich kilku lat podejmowano kilka prób detekcji farm fotowoltaicznych na podstawie danych teledetekcyjnych przy wykorzystaniu metod uczenia maszynowego. Kruitwagen et al. (2021), wykorzystując zdjęcia satelitarne SPOT-6/7 i Sentinel-2 w połączeniu z metodami uczenia maszynowego, wskazał istniejące farmy fotowoltaiczne na świecie na dzień 30 września 2018 roku. Inne badania, takie jak te przeprowadzone przez Zhang et al. (2021), Plakman et al. (2022) i Wang et al. (2022), również koncentrowały się na detekcji instalacji fotowoltaicznych, lecz nie obejmowały obszaru Polski.

Celem niniejszej pracy jest określenie optymalnych metod uczenia maszynowego do wykrywania farm fotowoltaicznych na podstawie danych teledetekcyjnych. W pracy wykorzystano dane radarowe z misji Sentinel-1 oraz dane multispektralne z misji Sentinel-2, wraz z ich pochodnymi, takimi jak wskaźniki teledetekcyjne i tekstury obrazu. W procesie klasyfikacji wykorzystano metodę lasów losowych (ang. *Random Forest*) zaproponowaną przez Breimana (2001). Cały proces oparty został na otwartym oprogramowaniu GIS, takim jak SNAP (ESA, 2023n), S1TBX (ESA, 2023a), QGIS (QGIS Development Team, 2023) oraz środowisku programistycznym języka R (R Core Team, 2023). Zgodnie z wiedzą autora niniejsza praca jest pierwszym opracowaniem w Polsce dotyczącym detekcji farm fotowoltaicznych przy wykorzystaniu danych satelitarnych i metod uczenia maszynowego.

Rozdział 2

Materiały

W ramach badania wykorzystano zarówno radarowe, jak i multispektralne dane satelitarne z programu Copernicus, pochodzące odpowiednio z misji Sentinel-1 i Sentinel-2. Program Copernicus to unijna inicjatywa obserwacji Ziemi, koordynowana i zarządzana przez Komisję Europejską we współpracy z różnymi partnerami, w tym Europejską Agencją Kosmiczną (ang. *European Space Agency*, ESA). Zarówno dane radarowe z misji Sentinel-1 oraz multispektralne zdjęcia satelitarne z misji Sentinel-2 zostały pozyskane 8 maja 2023 roku, gdy współczynnik zachmurzenia na obszarze badań wynosił mniej niż 1%. W badaniu wykorzystano dane Sentinel-1 Ground Range Detected (GRD) oraz skorygowane atmosferycznie dane Sentinel-2 L2A, reprezentujące współczynnik odbicia na powierzchni Ziemi.

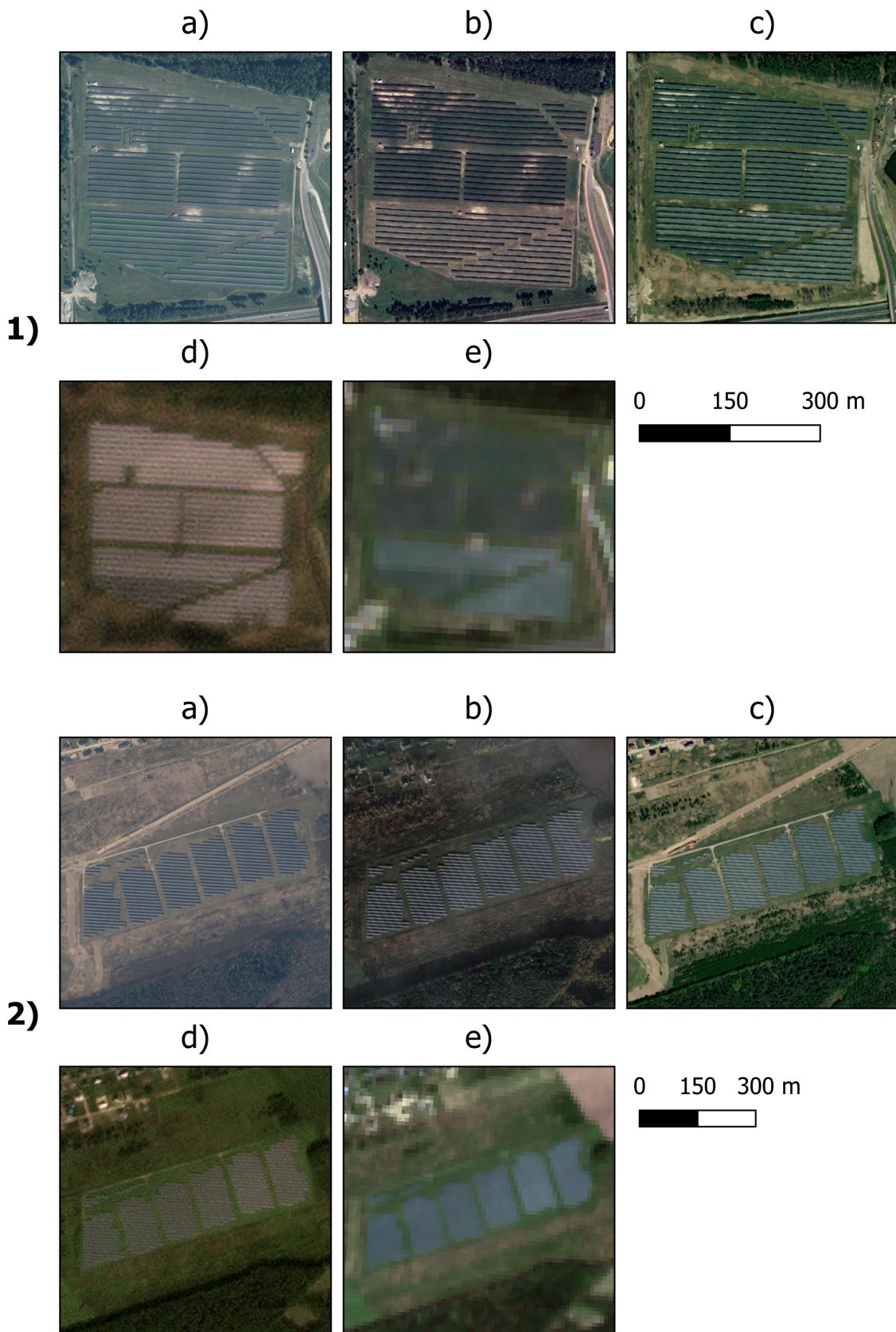
Obszar analizy stanowi kafel Sentinel-2 o oznaczeniu 33UWV, obejmujący głównie obszar województwa zachodniopomorskiego oraz fragment województwa wielkopolskiego. Powierzchnia obszaru analizy wynosi w przybliżeniu 12000 km^2 , a jego lokalizację przedstawia rycina 2.1.



Rycina 2.1: Obszar badań

2.1 Lokalizacje farm fotowoltaicznych

Jednym z zastosowań mozaik obrazów satelitarnych jest tworzenie zestawów danych referencyjnych poprzez interpretację wizualną, np. w celu walidacji wyników klasyfikacji produktów pokrycia terenu (Lesiv et al., 2018). Do lokalizacji oraz digitalizacji istniejących farm fotowoltaicznych wykorzystano ortofotomapę udostępnianą przez Główny Urząd Geodezji i Kartografii (GUGiK) oraz mozaiki zdjęć satelitarnych dostarczane przez podmioty komercyjne. W celu stworzenia zbioru danych testowych i treningowych użyto ortomozaik Google Satellite, Bing Aerial oraz Planet Basemaps (Planet Labs PBC, 2024), udostępnianych w formie usług sieciowych (WMS, WMTS, XYZ Tiles). Mozaiki te są tworzone na podstawie komercyjnych zdjęć satelitarnych wykonywanych przez podmioty takie jak Maxar Technologies, Airbus czy Planet Labs.



Rycina 2.2: Porównanie wyglądu farm fotowoltaicznych w Kulicach k. Nowogardu (1) i Wałczu (2) na ortofotomapie udostępnianej przez GUGiK (a), Google Satellite (b), Bing Aerial (c), Planet Basemaps (d) oraz w kompozycji RGB Sentinel-2 (e)

Większość ortofotomap dostarczanych przez GUGiK jest zrealizowana w standardzie 25 x 25 cm, jednakże na obszarach miejskich charakteryzują się one rozdzielczością przestrzenną wynoszącą 10 cm lub nawet wyższą (GUGiK, 2023). Rozdzielcość przestrzenna wykorzystanych mozaik obrazów satelitarnych (oprócz Planet Basemaps) jest wyższa niż 1 m, na przykład mozaika Bing Aerial dostarczana przez firmę Microsoft cechuje się rozmiarem komórki od 30 do 60 cm (Microsoft, 2023). Często jednak nie jest możliwe ustalenie dat wykonania zdjęć satelitarnych, które posłużyły do stworzenia konkretnej mozaiki zobrazowań satelitarnych (Lesiv et al., 2018). Mozaika tworzona przez Planet na podstawie zdjęć satelitarnych wykonywanych przez konstelację satelitów PlanetScope charakteryzuje się niższą rozdzielczością przestrzenną (4,77 m na równiku), jednak w porównaniu do pozostałych wymienionych produktów jest tworzona z miesięczną oraz kwartalną częstotliwością (Planet Labs PBC, 2019). Pozwala to, pomimo niższej rozdzielczości przestrzennej, na stworzenie zbioru danych referencyjnych dla konkretnego zakresu czasu. Mozaiki Planet Basemaps (Planet Labs PBC, 2024) są tworzone na podstawie najlepszych obrazów z katalogu Planet w określonym przedziale czasowym, co umożliwia generowanie wysokorozdzielczych mozaik, które są dokładne radiometrycznie i przestrzennie, a także charakteryzują się zminimalizowanym wpływem czynników atmosferycznych (Planet Labs PBC, 2019).

Na podstawie wyżej wymienionych danych dokonano digitalizacji farm fotowoltaicznych na obszarze badania w celu stworzenia zestawu danych referencyjnych. Digitalizacja polega na tworzeniu obiektów wektorowych, takich jak punkty, linie czy poligony, poprzez rysownie figur geometrycznych uwzględniając ich współrzędne geograficzne na podstawie obrazów satelitarnych czy zeskanowanych map papierowych. Do tego celu wykorzystano narzędzia digitalizacji z oprogramowania QGIS (QGIS Development Team, 2023), które umożliwiają rysowanie kształtów i edycję danych wektorowych.

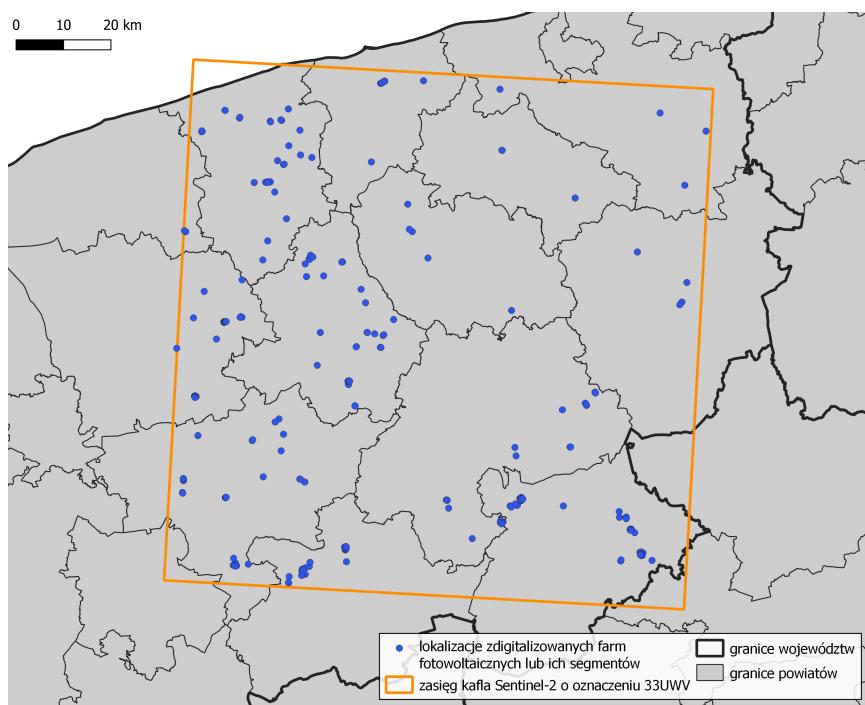
Na podstawie ortofotomapy oraz mozaik satelitarnych zdigitalizowane zostały wszystkie znalezione farmy fotowoltaiczne na obszarze kafla 33UWV Sentinel-2, istniejące w czasie wykonywania wykorzystanych zobrazowań (8 maja 2023 roku, sekcja 2.2). Użyte do digitalizacji dane pochodzą z różnych okresów, co spowodowało sytuacje, gdzie farmy obecne na najbardziej aktualnej mapie podkładowej (Planet Basemaps) nie występowały

Tabela 2.1: Statystki zdigitalizowanych farm fotowoltaicznych na obszarze kafla Sentinel-2 o oznaczeniu 33UWV

Liczba zdigitalizowanych farm PV lub ich segmentów	226
Suma powierzchni farm fotowoltaicznych [ha]	346.47
Minimalna powierzchnia farmy PV lub jej segmentu	0.01
Średnia powierzchnia farmy PV lub jej segmentu [ha]	1.53
Maksymalna powierzchnia farmy PV lub jej segmentu [ha]	24.89

na mozaikach o wyższej rozdzielczości przestrzennej (ortofotomapa, Google Satellite, Bing Aerial). Podczas digitalizacji przyjęto, że granice farm fotowoltaicznych lub ich segmentów będą ustalane na podstawie mapy podkładowej o najwyższej rozdzielczości, na której dana farma fotowoltaiczna występuje.

Podstawowe statystki zdigitalizowanych obiektów przedstawia tabela 2.1, a lokalizacje oznaczonych farm fotowoltaicznych lub ich segmentów przedstawia rycina 2.3. Zauważać można duże zróżnicowanie w powierzchni farm fotowoltaicznych lub ich segmentów, od małych instalacji o powierzchni 0,01 ha do znacznie większych konstrukcji o powierzchni prawie 25 ha. Rozmieszczenie farm fotowoltaicznych na obszarze kafla 33UWV Sentinel-2 skupia się głównie w zachodniej i południowej części badanego obszaru (rycina 2.3).

**Rycina 2.3:** Rozmieszczenie zdigitalizowanych farm fotowoltaicznych lub ich segmentów

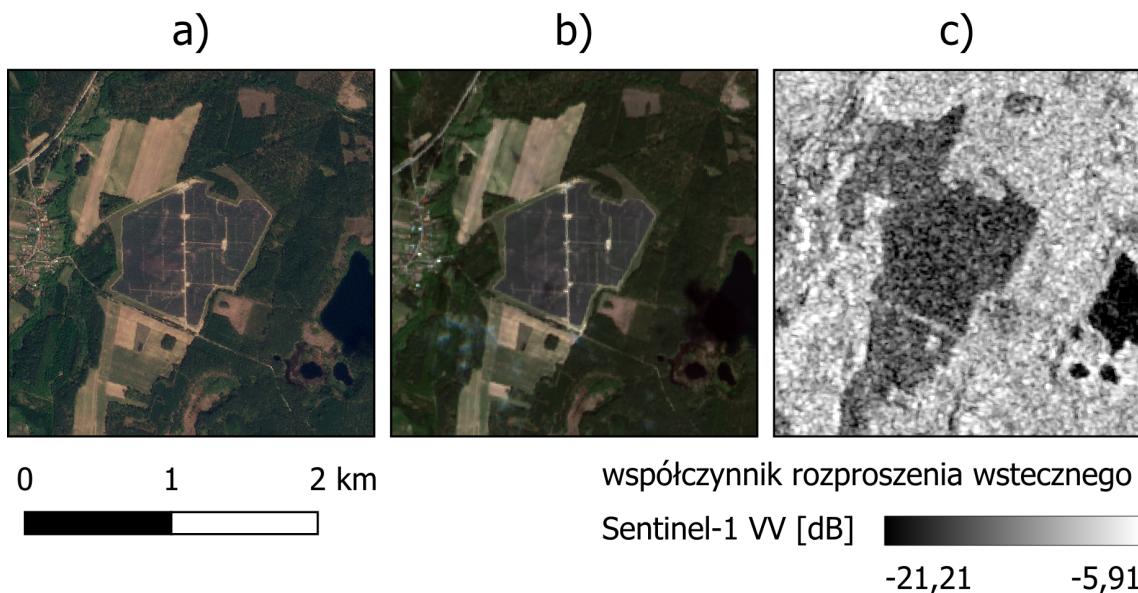
2.2 Zdjęcia satelitarne

2.2.1 Sentinel-1

Misja Sentinel-1 to wspólna inicjatywa Komisji Europejskiej i Europejskiej Agencji Kosmicznej w ramach programu Copernicus, mająca na celu dostarczanie danych radarowych obejmujących powierzchnię Ziemi, w tym lądów, europejskich stref przybrzeżnych, tras żeglugowych, stref lodu morskiego, mórz i oceanów (Hejmanowska et al., 2020; ESA, 2023g). Misja odpowiada na potrzeby monitorowania obszarów morskich i lądowych, w tym przemieszczania kry lodowej, transportu morskiego, deformacji i przemieszczeń terenu, a także obserwacji zmian klimatu i klęsk żywiołowych (Hejmanowska et al., 2020; ESA, 2023g).

Sentinel-1 posiada pojedynczy radar z syntetyzowaną aperturą (ang. *Synthetic Aperture Radar*, SAR), który działa na jednej częstotliwości środkowej 5,405 GHz, odpowiadającej długości fali 5,6 cm (Braun, 2020; ESA, 2023e). Częstotliwość radaru Sentinel-1, mieszcząca się w pasmie C (od 4 do 8 GHz), determinuje jego zdolność do penetracji jedynie górnych warstw koron drzew, roślinności oraz gleby (ESA, 2023j). Radar jest najczęściej wykorzystywany w trybie podwójnej polaryzacji, emitując fale pionowe i mierząc zarówno fale pionowe (ang. *vertical*), jak i poziome (ang. *horizontal*) po ich powrocie do czujnika, dzięki czemu otrzymujemy dane o intensywności rozproszenia wstecznego VV i VH (Braun, 2020). Radar z syntetyzowaną aperturą umożliwia prowadzenie obserwacji zarówno w nocy, jak i w dzień, niezależnie od warunków pogodowych, przez co czas rewizyty konstelacji dwóch satelitów dla obszaru Polski wynosi około 2 dni (Attema et al., 2008; ESA, 2023h). W grudniu 2021 roku w satelicie Sentinel-1B wystąpiła awaria układu zasilania elektroniki radaru, uniemożliwiająca dalsze dostarczanie danych radarowych (ESA, 2023b). W wyniku awarii satelita Sentinel-1B został wyłączony z użytku, a wystrzelenie platformy Sentinel-1C zaplanowano na marzec 2024 roku (ESA, 2023b; eoPortal, 2023).

Dane radarowe Sentinel-1 są dostępne na trzech poziomach przetworzenia. Dane poziomu 0 to surowe, nieobrazowe dane, z których generowane są produkty wyższych poziomów przetworzenia. Poziom 1 obejmuje produkty Single Look Complex (SLC) oraz



Rycina 2.4: Farma fotowoltaiczna na wysokorozdzielczym obrazie satelitarnym PlanetScope (a), kompozycji RGB Sentinel-2 (b) oraz obrazie rozproszenia wstecznego w polaryzacji VV Sentinel-1 (c)

Ground Range Detected (GRD) przeznaczone do monitorowania Ziemi, klasyfikacji pokrycia terenu oraz aplikacji interferometrycznych¹ (Hejmanowska et al., 2020). Produkty SLC to obrazy SAR w geometrii ukośnej, posiadające informację o wartości amplitudy i fazie sygnału, co czyni je odpowiednimi do interferometrii radarowej (Hejmanowska et al., 2020). Produkty GRD są rezultatem przepróbkowania obrazów SLC do jednolitej rozdzielczości przestrzennej (10×10 m) i rzutowania na powierzchnię elipsoidy odniesienia (Hejmanowska et al., 2020). Przy konwersji SLC do GRD tracimy informację fazową sygnału, co wyklucza zastosowanie produktów GRD w interferometrii radarowej (ESA, 2023d). Produkty GRD nie posiadają cech ortofotomapy i wymagają dodatkowego przepróbkowania z użyciem cyfrowego modelu wysokości przed wykorzystaniem w systemach GIS (Hejmanowska et al., 2020). Produkty oceaniczne poziomu drugiego (Level-2 Ocean (OCN)) służą natomiast do zastosowań związanych z pomiarem wiatrów, fal i prądów morskich (ESA, 2024). Produkty przeznaczone dla użytkowników końcowych to produkty pierwszego i drugiego poziomu przetworzenia.

¹Interferometria radarowa (ang. *Interferometric Synthetic Aperture Radar*, InSAR) to technika umożliwiająca generowanie cyfrowych modeli wysokościowych (ang. *digital elevation model*, DEM) oraz pomiary deformacji terenu na podstawie informacji o fazie sygnału (Hanssen, 2001; Hejmanowska et al., 2020).

Tabela 2.2: Tryby radaru Sentinel-1 (ESA, 2023). Pogrubiony wiersz wskazuje tryb wykorzystany w niniejszym badaniu.

Tryb	Kąt padania wiązki [°]	Rozdzielcość przestrzenna [m]	Szerokość rejestrowanego pasa [km]	Polaryzacje (H=pozioma, V=pionowa)
Stripmap	20 - 45	5 x 5	80	HH+HV, VH+VV, HH, VV
Interferometric Wide Swath	29 - 46	5 x 20	250	HH+HV, VH+VV, HH, VV
Extra Wide Swath	19 - 47	20 x 40	400	HH+HV, VH+VV, HH, VV
Wave	22 - 35 35 - 38	5 x 5	20 x 20	HH, VV

Dane Sentinel-1 rejestrowane są w czterech trybach: Stripmap (SM) - używany do obrazowania małych wysp i na potrzeby zarządzania kryzysowego, Interferometric Wide Swath (IW) - podstawowy tryb obrazowania dla obszarów lądowych, Extra-Wide Swath (EW) - tryb do monitorowania stref polarnych i niektórych obszarów morskich oraz Wave (WV) - tryb obrazowania oceanów (Hejmanowska et al., 2020; ESA, 2023e,c). Charakterystykę konkretnych trybów akwizycji radaru Sentinel-1 przedstawia tabela 2.2 (ESA, 2023f).

W pracy wykorzystano dwie polaryzacje (VV i VH) pochodzące z produktów Ground Range Detected, zarejestrowanych w trybie Interferometric Wide Swath przez platformę Sentinel-1A w dniu 8 maja 2023 roku. Podczas rejestracji danych platforma znajdowała się na orbicie wznoszącej (ang. *ascending orbit*), co oznacza, że satelita przemieszczał się z południa w kierunku bieguna północnego. Użyty zestaw danych został utworzony poprzez połączenie dwóch sąsiednich produktów Sentinel-1 GRD, a następnie ograniczenie obszaru analizy do kafla Sentinel-2 o oznaczeniu 33UWV.

2.2.2 Sentinel-2

Misja Sentinel-2 stanowi inicjatywę Komisji Europejskiej, która jest operacyjnie prowadzona przez Europejską Agencję Kosmiczną w ramach programu Copernicus. Celem

tej misji jest dostarczanie obrazów satelitarnych, obejmujących trzynaście zakresów spektralnych o różnych rozdzielczościach przestrzennych: 10, 20 lub 60 metrów, zależnie od rejestrowanego kanału. Czas rewizyty konstelacji dwóch satelitów wzrasta z 5 dni nad równikiem do 2-3 dni na średnich szerokościach geograficznych, obejmujących również obszar Polski (Hejmanowska et al., 2020; ESA, 2023k).

Dane pozyskiwane przez satelity Sentinel-2 są dostępne na różnych poziomach przetworzenia, lecz najczęściej używane przy tworzeniu map pokrycia terenu i użytkowania ziemi (ang. *Land Use/Land Cover*, LULC) są produkty 1C (współczynnik odbicia na poziomie górnej części atmosfery; ang. *Top-of-Atmospheric reflectance*, TOA) oraz 2A (współczynnik odbicia na powierzchni Ziemi; ang. *Bottom-of-Atmospheric reflectance*, BOA) (Phiri et al., 2020).

Produkty poziomu 1C to dane poddane korekcjom radiometrycznym i geometrycznym, prezentowane jako sceny (ang. *tile*, *granule*) o zasięgu 100 x 100 km w projekcji UTM/WGS84 (ESA, 2015). Skuteczne wykorzystanie tych danych w zastosowaniach związanych z terenami lądowymi wymaga precyzyjnej korekcji zdjęć satelitarnych pod kątem efektów atmosferycznych (Main-Knorn et al., 2017). Produkty poziomu 2A powstają poprzez zastosowanie dodatkowej korekcji atmosferycznej dla danych poziomu 1C za pomocą procesora korekcji atmosferycznej Sen2Cor (Main-Knorn et al., 2017).

Dane wykorzystane w analizie pochodzą z dnia 8 maja 2023 roku i zostały dostarczone przez satelitę Sentinel-2B. Obszar analizy obejmuje kafel (ang. *tile*) o oznaczeniu 33UWV, dla którego współczynnik zachmurzenia w tym dniu wynosił 0,7%. Użyte zostały dane na poziomie przetworzenia L2A. Z dostępnych kanałów spektralnych (tabela 2.3 (ESA, 2023l)) wykorzystano 10 zakresów, ponieważ pasma rejestrowane w rozdzielczości 60 metrów są przeznaczone głównie do korekcji atmosferycznych i detekcji chmur. Kanał 1 (443 nm) służy do korekcji wpływu aerozoli, kanał 9 (945 nm) do korekcji wpływu pary wodnej, a kanał 10 (1374 nm) do wykrywania chmur typu cirrus (Drusch et al., 2012).

Tabela 2.3: Kanały spektralne satelitów Sentinel-2 (ESA, 2023). Pogrubione wiersze wskazują kanały wykorzystane w niniejszym badaniu.

Kanał	Nazwa kanału	Centralna długość fali [nm]	Zakres spektralny [nm]	Rozdzielcość przestrzenna [m]
B01	Coastal Aerosol	443	433–453	60
B02	Blue	493	458–523	10
B03	Green	560	543–578	10
B04	Red	665	650–680	10
B05	Vegetation RedEdge	704	698–713	20
B06	Vegetation RedEdge	740	733–748	20
B07	Vegetation RedEdge	783	773–793	20
B08	NIR	833	785–900	10
B8A	NIR	865	855–875	20
B09	Water Vapour	945	935–955	60
B10	Cirrus	1374	1360–1390	60
B11	SWIR	1610	1565–1655	20
B12	SWIR	2190	2100–2280	20

Rozdział 3

Metody

Badanie obejmowało kilka etapów, rozpoczynając od wstępного przetwarzania danych satelitarnych, opisanego w sekcjach 3.1.1 i 3.1.2. Następnie przeprowadzono obliczenia produktów pochodnych, takich jak wskaźniki spektralne oraz tekstury obrazu, opisane odpowiednio w sekcjach 3.1.3 i 3.1.4. Kolejnym krokiem było połączenie przetworzonych danych źródłowych i ich produktów pochodnych w kilka wariantów zbiorów danych, stworzonych na potrzeby detekcji farm fotowoltaicznych. Warianty zbiorów danych przedstawia tabela 3.1, a przebieg łączenia danych w wielowarstwowe rastry został opisany w sekcji 3.1.5. Na podstawie uzyskanych rastrów, zgodnie z sekcją 3.2, utworzono zbiory danych uczących, które zostały wykorzystane do tworzenia modeli predykcyjnych. W procesie klasyfikacji wykorzystano metodę lasów losowych (ang. *Random Forest*, RF) zaproponowaną przez Breimana ([2001](#)), która została dokładniej przedstawiona w sekcji 3.3.1. Następnie zoptymalizowano parametry tych modeli, oceniono ich jakość oraz ważność zmiennych dla każdego z nich, co opisano odpowiednio w sekcjach 3.3.2, 3.3.3 oraz 3.4. Ostatecznie, przeprowadzono predykcje, a uzyskane wyniki poddano procesowi przetwarzania końcowego, przedstawionemu w sekcji 3.5. Wizualną reprezentację przebiegu badania przedstawia rycina **?@fig-rycina-workflow**.

Przetwarzanie danych Sentinel-1 GRD zostało przeprowadzane przy użyciu oprogramowania Sentinel-1 Toolbox (S1TBX) ([ESA, 2023a](#)) i SNAP ([ESA, 2023n](#)), udostępnianego przez Europejską Agencję Kosmiczną. Każdy z kolejnych etapów został wykonany przy

użyciu języka R (R Core Team, 2023) oraz pakietów rozszerzających funkcjonalności tego języka. Końcowe wizualizacje zostały stworzone w oprogramowaniu QGIS (QGIS Development Team, 2023). Zarówno oprogramowanie GIS (SNAP, S1TBX, QGIS), jak i środowisko programistyczne R stanowią oprogramowanie o otwartym kodzie źródłowym (ang. *open-source software*).

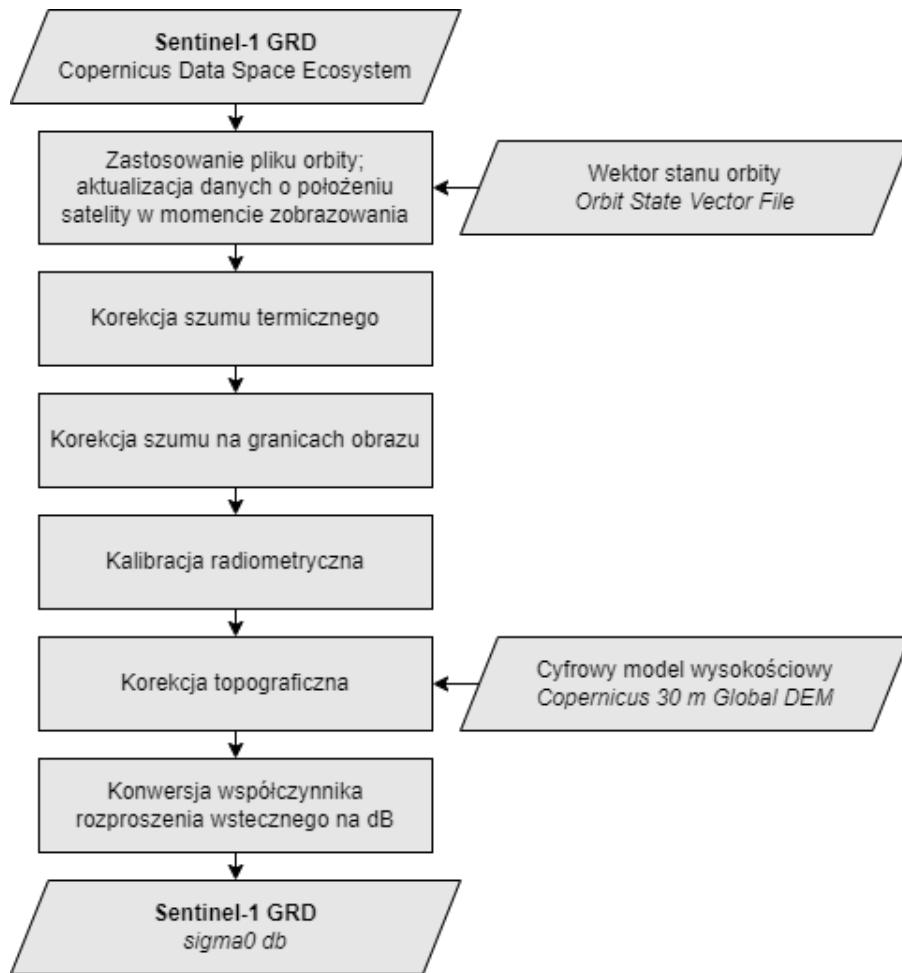
RYCINA - OGÓLNY SCHEMAT PRZEBIEGU BADANIA (WORKFLOW)

3.1 Przygotowanie danych

3.1.1 Sentinel-1

Korzystanie z danych radarowych wymaga wcześniejszego przygotowania danych poprzez proces kalibracji, aby zapewnić poprawne wyniki analizy. Procesy te mogą różnić się w zależności od konkretnego zastosowania, mając na celu dostosowanie danych do specyficznych potrzeb. W pracy wykorzystano schemat przetwarzania danych Sentinel-1 GRD, który zaproponował Filippone (2019), obejmujący:

1. aktualizację informacji o położeniu satelity w momencie zobrazowania poprzez pobranie dokładnych wektorów stanu orbity dla produktu zapewniając precyzyjne informacje o pozycji i prędkości satelity podczas akwizycji;
2. korekcję szumów termicznych;
3. korekcję szumów na granicach obrazów;
4. obliczenie współczynnika rozproszenia wstecznego (ang. *backscatter coefficient*) sigma0 za pomocą kalibracji radiometrycznej;
5. korekcję topograficzną (ortorektyfikacja za pomocą Copernicus 30 m Global DEM);
6. konwersję współczynnika rozproszenia wstecznego na dB za pomocą transformacji logarytmicznej



Rycina 3.1: Przebieg wstępniego przetwarzania danych Sentinel-1 Ground Range Detected (GRD)

Podczas przetwarzania danych radarowych pominięto etap filtrowania plamek (ang. *speckle filtering*), ponieważ ze względu na możliwość utraty istotnych informacji nie jest to zalecane w kontekście identyfikacji tekstur obrazu (Filipponi, 2019).

Dane Sentinel-1 GRD dla obu polaryzacji (VV i VH) przygotowano w sposób przedstawiony na rycinie 3.1 przy użyciu narzędzi dostępnych w ESA Sentinel-1 Toolbox (ESA, 2023a), a proces ten został ułożony w sekwencję operacji przetwarzania za pomocą narzędzia GraphBuilder w oprogramowaniu SNAP (ESA, 2023n). Kolejne etapy przygotowania danych zostały zrealizowane przy wykorzystaniu języka R (R Core Team, 2023) oraz pakietu *terra* (Hijmans, 2023). Obszar analizy, będący kaflem Sentinel-2 o oznaczeniu 33UWV, znajduje się na granicy dwóch sąsiednich produktów Sentinel-1 GRD. Na potrzeby dalszego przetwarzania, sąsiadujące produkty zostały połączone i odpowiednio ograniczone do obszaru zainteresowania. Na granicy sąsiednich produktów



Rycina 3.2: Obszar badań i zasięg wykorzystanych produktów Sentinel-1

Sentinel-1 GRD występowała przestrzeń bez danych o szerokości jednej komórki, co wymagało wypełnienia tego obszaru danymi przy użyciu funkcji *focal* z pakietu *terra* (Hijmans, 2023).

3.1.2 Sentinel-2

Przetwarzanie danych Sentinel-2 polegało na sprowadzeniu kanałów o rozdzielczości przestrzennej 20 m do rozdzielczości 10 m. Przepróbkowanie (ang. *resampling*) zostało przeprowadzone przy pomocy funkcji *resample* z pakietu *terra* (Hijmans, 2023), stosując interpolację dwuliniową (ang. *bilinear interpolation*).

W teledetekcji miarą opisującą wielkość odbicia jest rzeczywisty współczynnik odbicia, nazywany również reflektancją, którego wartości mieszczą się w przedziale od 0 do 1 (Hejmanowska et al., 2020). Informacje dotyczące odbitego promieniowania w przypadku danych Sentinel-2 są udostępniane za pomocą współczynnika odbicia, wyrażonego w

zakresie od 0 do 10000, gdzie wartości te reprezentują jasność komórki, określana w języku angielskim jako *Digital Number* (DN). Po standaryzacji wszystkich kanałów Sentinel-2 do wspólnej rozdzielczości, dokonano przeliczenia wartości kanałów w celu uzyskania rzeczywistego współczynnika odbicia, poprzez podzielenie wartości DN przez 10000.

3.1.3 Wskaźniki spektralne

Oprócz surowych współczynników odbicia, w niektórych wariantach predykcji zastosowano również wskaźniki spektralne (ang. *spectral indices*), które były wykorzystywane w poprzednich badaniach dotyczących detekcji farm fotowoltaicznych na podstawie danych teledetekcyjnych przez Zhang et al. (2021), Plakman et al. (2022), Wang et al. (2022) i innych, takie jak:

- znormalizowany różnicowy wskaźnik wegetacji (ang. *Normalized Difference Vegetation Index*, NDVI) (Tucker, 1979), monitorujący zawartość biomasy i kondycję roślinności na danym obszarze:

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

, gdzie *NIR* – reflektancja w kanale bliskiej podczerwieni, *Red* – reflektancja w kanale czerwonym

- znormalizowany różnicowy wskaźnik obszarów zabudowanych (ang. *Normalized Difference Built-up Index*, NDBI) (Zha et al., 2003), przeznaczony do kartowania obszarów zabudowanych:

$$NDBI = \frac{SWIR1 - NIR}{SWIR1 + NIR}$$

, gdzie *SWIR1* – reflektancja w kanale średniej podczerwieni, *NIR* – reflektancja w kanale bliskiej podczerwieni

- znormalizowany zmodyfikowany różnicowy wskaźnik wody (ang. *Modified Normalized Difference Water Index*, mNDWI) (Xu, 2006), który skutecznie identyfikuje

obszary wodne na zdjęciach satelitarnych, mając możliwości tłumienia zakłóceń spowodowanych przez zabudowę, roślinność i gleby:

$$mNDWI = \frac{Green - SWIR1}{Green + SWIR1}$$

, gdzie *Green* – reflektancja w kanale zielonym, *SWIR1* – reflektancja w kanale średniej podczerwieni

3.1.4 Tekstury obrazu

Tekstura stanowi istotną cechę wykorzystywaną do identyfikacji obiektów i obszarów zainteresowania na obrazie (Haralick et al., 1973) oraz odgrywa dużą rolę w interpretacji wizualnej zdjęć lotniczych i satelitarnych (Lewiński, 2012). Gdy różnice widmowe pomiędzy klasami są niewielkie, tekstura umożliwia rozróżnienie odmiennych typów obiektów na podstawie ich ułożenia w przestrzeni, często kontrastując obszary naturalne z antropogenicznymi (Lennert et al., 2023). W zależności od zastosowanej funkcji wybrane cechy obrazu zostają uwidoczone w porównaniu z pierwotnym obrazem wejściowym (Lewiński, 2012). Informacja o teksturze może stanowić dodatkową, przydatną zmienną wejściową w procesach klasyfikacji lub segmentacji obrazu (Gong et al., 1992; Mumby et al., 2002). Tekstura obejmuje różnice poziomów szarości (kontrast), obecność lub brak kierunkowości, regularne wzory i zdefiniowany obszar, na którym występują zmiany, określony przez rozmiar okna (Hall-Beyer, 2017; Lennert et al., 2023). Można ją opisać za pomocą tonu (intensywność poziomu szarości) i struktury (relacje przestrzenne) (Lennert et al., 2023).

Model oparty na macierzy współwystępowania poziomów szarości (ang. *Gray Level Co-Occurrence Matrix*, GLCM), zaproponowany przez Haralicka et al. (1973), jest często używany do określania tekstur obrazu. Ta metoda polega na tworzeniu macierzy opisującej częstotliwość występowania par wartości w określonym fragmencie obrazu, uwzględniając określone sąsiedztwo, kierunki i odstępy między komórkami (Kupidura, 2019).

Dane teledetekcyjne zazwyczaj reprezentują dane ciągłe o szerokim zakresie, które niejednokrotnie mogą przyjmować zarówno wartości dodatnie, jak i ujemne, niekoniecznie ograniczając się do liczb całkowitych (Illich, 2023). Obliczanie tekstur obrazu wymaga uprzedniego zredukowania zakresu wartości obrazów rastrowych do dyskretnej liczby poziomów szarości, co określa się mianem kwantyzacji (ang. *quantization*). Zalecana przez Haralicka et al. (1973) metoda kwantyzacji danych opiera się na równym prawdopodobieństwie (ang. *equal probability*), dokonując redukcji poziomów szarości wykorzystując kwantyle do utworzenia przedziałów zawierających w przybliżeniu równą liczbę próbek (Illich, 2023).

Przydatność i wykorzystanie tekstury w dużym stopniu zależy od rozdzielnosci przestrzennej zdjęć satelitarnych i wielkości zjawiska, które ukształtowało teksturę (Lennert et al., 2023). Badanie, które przeprowadził Zhang et al. (2021) dotyczące wykorzystania filtracji teksturalnych w identyfikacji elektrowni fotowoltaicznych z użyciem metody Random Forest i danych z Landsata-8 wykazało pozytywny wpływ tekstur na skuteczność modelu. Według wyników badania najlepiej dopasowany model wykorzystywał tekstury GLCM o sąsiedztwie 30 pikseli (co odpowiada wymiarom ruchomego okna o wymiarach 1830 m na 1830 m), natomiast tekstura o rozmiarze jednego sasiada miała niewielki wpływ na poprawę dokładności modelu (Zhang et al., 2021).

Obliczanie tekstur obrazu może być procesem wymagającym obliczeniowo, dlatego w pracy wykorzystano jedynie teksturę średniej sumy (ang. *Sum Average*, SA lub SAVG), wskazaną przez Zhang et al. (2021) i Wanga et al. (2022) jako teksturę niosącą najwięcej informacji w kontekście detekcji farm fotowoltaicznych na obrazach satelitarnych. Przeprowadzono obliczenia sześciu tekstur obrazów dla: dwóch kanałów Sentinel-2 (B02 i B8A), dwóch wskaźników teledetekcyjnych (NDBI i nMDWI) oraz dla obu polaryzacji Sentinel-1 (VV i VH), zgodnie ze wskazaniami Wanga et al. (2022). Z uwagi na rosnący czas obliczeń wraz ze zwiększeniem liczby poziomów szarości oraz rozmiarów ruchomego okna, zdecydowano się na kwantyzację wartości rastrów do 32 przedziałów oraz zastosowanie ruchomego okna o sąsiedztwie 9 komórek. Odpowiada to wymiarom ruchomego okna o wymiarach 19 pikseli na 19 pikseli lub kwadratowi o wymiarach 190 m na 190 m.

Tabela 3.1: Zestawienie różnych wariantów zbiorów danych stworzonych na potrzeby detekcji farm fotowoltaicznych na podstawie danych teledetekcyjnych

Wariant	Liczba zmiennych	Zmienne ^a
1	10	S2: B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12
2	13	S2: B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12; wskaźniki spektralne: NDVI, NDBI, mNDWI
3	16	S2: B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12; wskaźniki spektralne: NDVI, NDBI, mNDWI; tekstury: B02_SAVG, B8A_SAVG, NDBI_SAVG, mNDWI_SAVG
4	12	S2: B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12; S1: VV, VH
5	16	S2: B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12; S1: VV, VH; tekstury: B02_SAVG, B8A_SAVG, VV_SAVG, VH_SAVG
6	21	S2: B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12; wskaźniki spektralne: NDVI, NDBI, mNDWI; S1:VV, VH; tekstury: B02_SAVG, B8A_SAVG, NDBI_SAVG, mNDWI_SAVG, VV_SAVG, VH_SAVG

^a Uwaga: S2 oznacza Sentinel-2, podczas gdy S1 oznacza Sentinel-1

3.1.5 Łączenie danych

W celu uzyskania spójnych wielowarstwowych rastrów wszystkie zbiory danych zostały sprowadzone do wspólnej rozdzielczości i siatki. Rozdzielcość przestrzenna danych Sentinel-1 GRD, analogicznie do danych Sentinel-2 przygotowanych w sposób przedstawiony w sekcji 3.1.2 wynosi 10 m. Mimo że rozdzielcość przestrzenna danych Sentinel-1 GRD jest identyczna z rozdzielcością danych Sentinel-2, siatki przestrzenne obu zbiorów różniły się od siebie, przez co wymagana była transformacja danych Sentinel-1 do siatki danych Sentinel-2 w celu zachowania ich zgodności. Przetransformowane dane zostały wykorzystane do obliczeń tekstur obrazu oraz wskaźników teledetekcyjnych. Po uzyskaniu produktów pochodnych, w zależności od wariantu, dane zostały scalone w formie kilku wielowarstwowych rastrów, które posłużyły do wyodrębnienia próbek obserwacji, niezbędnych do stworzenia modelu. Złączone rastry zostały również bezpośrednio wykorzystane do przeprowadzenia predykcji w późniejszym etapie analizy. Warianty zestawów danych zostały szczegółowo przedstawione w tabeli 3.1.

3.2 Próbki treningowe i testowe

Na podstawie ortofotomapy oraz mozaik satelitarnych wskazanych w sekcji 2.1 zdigitalizowane zostały prawdopodobnie wszystkie farmy fotowoltaiczne na obszarze kafla Sentinel-2 o oznaczeniu 33UWV istniejące w czasie wykonywania wykorzystanych zobrażeń (8 maja 2023 roku). Z każdego zdigitalizowanego poligona, reprezentującego obszar pod panelami fotowoltaicznymi pozyskano dwie losowo zlokalizowane próbki, stanowiące obserwacje pozytywne. Lokalizacje próbek negatywnych, znajdujących się poza obszarami oznaczonymi jako farmy fotowoltaiczne na obszarze kafla Sentinel-2 o oznaczeniu 33UWV zostały wylosowane, wykorzystując próbkowanie losowe stratyfikowane (ang. *stratified*). Próbkowanie losowe stratyfikowane polega na podziale obszaru analizy na regularne komórki, a następnie losowaniu lokalizacji punktu w każdej komórce (Wang et al., 2012).

Po wykonaniu testowych predykcji zauważono, że modele przeuczały się na niektórych typach pokrycia terenu i użytkowania ziemi, wskazując farmy fotowoltaiczne w miejscach, gdzie faktycznie nie występowały. W celu poprawy wyników predykcji dodatkowe lokalizacje negatywnych próbek zostały wylosowane na obszarach niepoprawnie sklasyfikowanych, wykorzystując dane z OpenStreetMap (OSM contributors, 2024).

Obszary niepoprawnie sklasyfikowane przez testowe modele to plaże, budynki oraz drogi. W celu poprawy wyników predykcji pobrano z bazy danych OpenStreetMap dane przestrzenne o plażach (`tag:natural=beach`), budynkach (`key:building`) oraz drogach (`key:highway`) na obszarze kafla Sentinel-2 o oznaczeniu 33UWV, z których wylosowano lokalizacje kolejnych negatywnych próbek. Dodatkowo, lokalizacje negatywnych obserwacji zostały wylosowane na zbiornikach wodnych (jeziorach; `tag:water=lake`).

Przy tworzeniu kolejnych testowych predykcji zauważono również skłonność do regularnego przeuczania się kolejnych, poprawionych modeli na terenach oznaczonych w OSM jako kopalnie torfu. W celu eliminacji tych błędnie sklasyfikowanych terenów, przy tworzeniu ostatecznych modeli, wykorzystano również negatywne próbki wylosowane na terenach kopalni odkrywkowych oznaczonych w bazie OpenStreetMap jako `tag:landuse=quarry`.

Dla każdej wylosowanej próbki zostały wyekstraktowane wartości pochodzące z danych teledetekcyjnych i ich pochodnych, przygotowanych zgodnie z opisem przedstawionym w sekcjach 3.1.1, 3.1.2, 3.1.3 oraz 3.1.4. Tak przygotowane próbki były następnie wykorzystane przy tworzeniu modeli uczenia maszynowego umożliwiających wykrywanie farm fotowoltaicznych na podstawie danych teledetekcyjnych i ich pochodnych.

3.3 Uczenie maszynowe

Klasyfikacja obrazów w teledetekcji polega na grupowaniu komórek w niewielkie zestawy klas, aby komórki w tych samych klasach miały podobne właściwości (Ismail et al., 2009). Istnieje wiele różnych metod klasyfikacji danych teledetekcyjnych. Stosunkowo nowymi podejściami wykorzystywanymi w tym kontekście są metody oparte na sztucznej inteligencji, takie jak uczenie maszynowe (ang. *Machine Learning*, ML) lub uczenie głębokie (ang. *Deep Learning*, DL) (Hejmanowska et al., 2020).

Uczenie maszynowe stanowi obszar sztucznej inteligencji, koncentrujący się na opracowywaniu algorytmów i modeli statystycznych zapewniających systemom komputerowym możliwość automatycznego uczenia się z danych i wykonywania określonych zadań bez konieczności bezpośredniego programowania. W przypadku skomplikowanych i złożonych zestawów danych nie jesteśmy w stanie odpowiednio ich zinterpretować oraz wydobyć poprawnych informacji po wizualnym przejrzeniu danych (Mahesh, 2019). Uczenie maszynowe jest wykorzystywane w celu doskonalenia efektywnego przetwarzania danych przez maszyny (Sindayigaya et al., 2022). Algorytmy uczenia maszynowego można podzielić na cztery główne podejścia: uczenie nienadzorowane (ang. *unsupervised learning*), uczenie nadzorowane (ang. *supervised learning*), uczenie częściowo nadzorowane (ang. *semi-supervised learning*) oraz uczenie przez wzmacnianie (uczenie posiłkowane, ang. *reinforcement learning*) (Sarker, 2021).

W poniższym badaniu do klasyfikacji wykorzystano nadzorowaną metodę lasów losowych (ang. *Random Forest*, RF) (Breiman, 2001).

Nadzorowane algorytmy uczenia maszynowego wykorzystują oznaczone dane treningowe do znajdywania powiązań pomiędzy różnymi zmiennymi. Proces uczenia nadzorowanego zachodzi, gdy określone cele mają zostać osiągnięte na podstawie konkretnego zestawu danych wejściowych (treningowych). W uczeniu nadzorowanym wyróżniamy klasyfikację, która dzieli dane na klasy, oraz regresję, która umożliwia oszacowanie określonych wartości (Sarker, 2021).

3.3.1 Metoda lasów losowych

Random Forest stał się jednym z najpopularniejszych klasyfikatorów uczenia maszynowego wykorzystywanych w dziedzinie teledetekcji ze względu na wysoką dokładność klasyfikacji i efektywność obliczeniową (Belgiu et al., 2016; Sheykhou et al., 2020). Metoda lasów losowych charakteryzuje się pewną odpornością na szумy (ang. *noise*) i przeuczenie (ang. *overfitting*), ponieważ nie bazuje na ważeniu (Gislason et al., 2006).

Algorytm Random Forest rozwija koncepcję drzew decyzyjnych, operując na zasadzie uczenia zespołowego (ang. *ensemble learning*), czyli łączenia wielu słabszych modeli (indywidualnych drzew decyzyjnych) w jeden silniejszy model (Maxwell et al., 2018; Sekulić et al., 2020). Procedura generuje liczne drzewa decyzyjne, opierając się na losowo wybranym zestawie danych ze zbioru danych uczących oraz losowo wyselekcjonowanych zmiennych klasyfikacyjnych (Breiman, 2001). Pojedyncze drzewo korzysta ze zredukowanej liczby danych treningowych i zmiennych, co sprawia, że drzewa różnią się od siebie i są mniej dokładne, ale jednocześnie są też mniej skorelowane, przez co model złożony z wielu drzew będzie bardziej niezawodny (Sekulić et al., 2020). W fazie predykcji każde z drzew dokonuje prognozy, a ostateczna decyzja jest formułowana na podstawie głosowania większościowego. W przypadku klasyfikacji klasa wybierana jest na podstawie największej liczby głosów (Breiman, 2001).

3.3.2 Dostrajanie modeli

Większość algorytmów uczenia maszynowego wymaga ustalenia wartości pewnych zmiennych konfiguracyjnych zwanych hiperparametrami (Król-Nowak et al., 2022). Hiperparametry są ustawiane przed rozpoczęciem procesu uczenia, nie uzyskuje się ich w

wyniku trenowania modelu ani nie są modyfikowane w trakcie wykonywania algorytmu (Król-Nowak et al., 2022). Optymalna konfiguracja hiperparametrów jest zwykle znajdowana w określonej przestrzeni poszukiwań (ang. *search space*) i może być ustalana np. na podstawie zagnieżdzonej walidacji krzyżowej (ang. *nested cross-validation*) (Lovelace et al., 2019). Optymalizacja hiperparametrów (ang. *hyperparameters*) odgrywa kluczową rolę w osiągnięciu najwyższej mocy predykcyjnej i jakości modelu (Schratz et al., 2019), a jej celem jest znalezienie optymalnej konfiguracji hiperparametrów algorytmu uczenia maszynowego dla danego zadania (Bischl et al., 2024).

Wykorzystanie tych samych danych do oceny jakości i dostrajania mogłoby potencjalnie prowadzić do nadmiernie wysokich wyników oceny jakości, czego można uniknąć, stosując zagnieżdżoną przestrzenną kroswalidację. Metoda przestrzennej zagnieżdżonej walidacji krzyżowej to rozwinięcie walidacji krzyżowej (sekcja 3.3.3), przeznaczone do procesu strojenia hiperparametrów. Każdy podzbiór (*fold*) utworzony w przestrzennej walidacji krzyżowej jest dzielony na kolejne, zewnętrzne i wewnętrzne podzbiory. W zewnętrznym obiegu model jest trenowany na danych z jednego obszaru przestrzennego i testowany na innym, podczas gdy wewnętrzny obieg służy do doboru optymalnych hiperparametrów. Cały proces jest następnie powtarzany na każdym z k zewnętrznych podzbiorów, co prowadzi do określenia najlepszego ustawienia hiperparametrów.

Hiperparametry `mtry`, `sample.fraction` i `min.node.size` są parametrami określającymi stopień losowości lasu losowego i powinny zostać odpowiednio dostrojone (Probst et al., 2019). Liczba losowo wybranych zmiennych, `mtry`, wskazuje, ile predyktorów powinno zostać użytych w każdym drzewie, a parametr `sample.fraction` odnosi się do wielkości próby, czyli ułamka obserwacji użytego w każdym drzewie (Lovelace et al., 2019). Mniejsza wielkość próby prowadzi do większej różnorodności drzew, a tym samym do mniejszej korelacji między nimi, co pozytywnie wpływa na dokładność predykcji przy agregacji drzew (Probst et al., 2019). Minimalna wielkość węzła `min.node.size` określa minimalną liczbę obserwacji w węźle końcowym (Probst et al., 2019). W ramach optymalizacji uwzględniono również parametry `num.trees` oraz `max.depth`, odnoszące się odpowiednio do liczby drzew w lesie oraz maksymalnej głębokości pojedynczego drzewa.

Tabela 3.2: Optymalne hiperparametry otrzymane w wyniku dostrajania modeli RF

Wariant ^a	Optymalizowane parametry						AUC
	mtry	sample.fraction	min.node.size	num.trees	max.depth		
1	6	0.6689	2	264	34	0.9842	
2	7	0.7373	1	372	23	0.9914	
3	7	0.8847	5	388	75	0.9904	
4	6	0.8758	5	291	54	0.9881	
5	5	0.6898	1	234	99	0.9850	
6	8	0.8847	5	388	75	0.9905	

^a Patrz: tabela 3.1

Kombinacje hiperparametrów zostały wybrane losowo w ramach określonych granic strojenia ustalonych za pomocą pakietu R *paradox* (Lang et al., 2023b). Zasięg przestrzeni strojenia został wybrany zgodnie z wartościami zalecanymi w dedykowanym do tego pakietie R *mlr3tuningspaces* (Becker, 2023) oraz literaturze (Probst et al., 2019; Schratz et al., 2019). `mtry` powinno przyjmować wartości z przedziału od 1 do liczby predyktorów, `sample.fraction` powinno mieścić się w zakresie od 0,2 do 0,9, a `min.node.size` powinno przybierać wartości z przedziału od 1 do 10. Zgodnie z pakietem R *mlr3tuningspaces* (Becker, 2023), hiperparametr `num.trees` powinien być ustawiony w zakresie od 1 do 2000, jednak ograniczono jego wartości do przedziału od 50 do 500 w celu zwiększenia wydajności obliczeń.

Optymalne hiperparametry uzyskane w wyniku dostrajania modeli lasów losowych dla poszczególnych wariantów (zbiorów danych) razem z oceną jakości AUC¹ zostały przedstawione w tabeli 3.2. Warto zauważyć, że warianty 3 i 6 wykazują identyczne wartości dla parametrów `sample.fraction`, `min.node.size`, `num.trees` i `max.depth`. Podczas losowania hiperparametrów dla każdego wariantu wykorzystano to samo ziarno losowości (ang. *random seed*), ustawione za pomocą funkcji `set.seed()`. Dla dwóch wspomnianych wariantów najlepsze wyniki zostały osiągnięte przy wykorzystaniu identycznego zestawu czterech z pięciu optymalizowanych hiperparametrów.

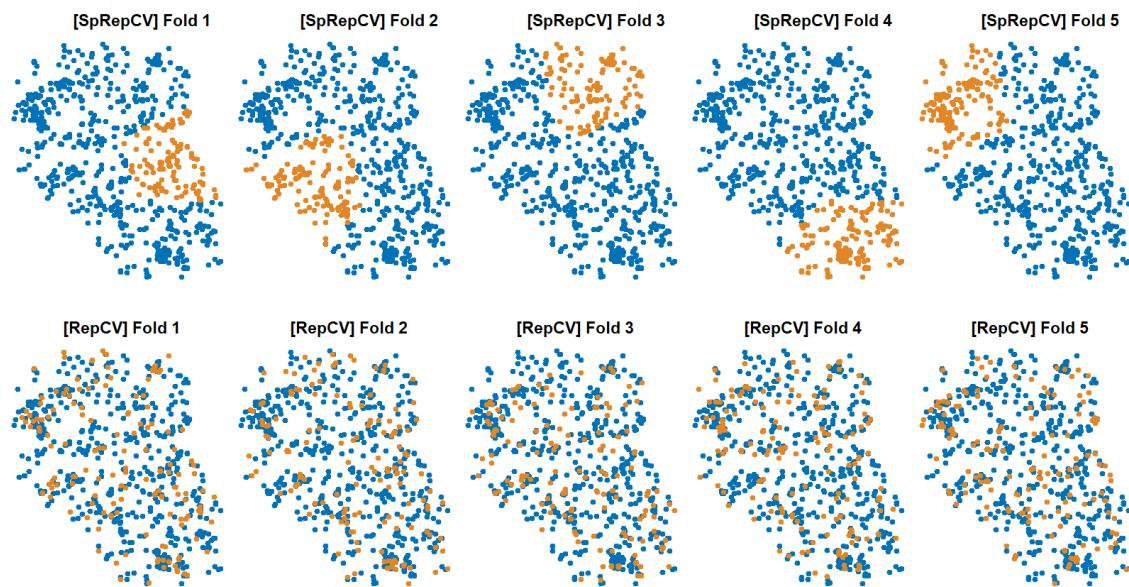
¹Pole powierzchni pod krzywą (ang. *Area Under Curve* lub *Area Under the ROC Curve*, AUC lub AUROC) to miara jakości modelu, obliczająca obszar pod krzywą ROC (ang. *receiver operating characteristic curve*), która graficznie przedstawia zależność pomiędzy czułością (*true positive rate*) a specyficznością (*false positive rate*) (Jaworski, 2013). AUC reprezentuje zdolność klasyfikatora binarnego do oddzielania klas pozytywnych od klas negatywnych. Wartości AUC mieszczą się w zakresie od 0 do 1, gdzie wartość 0,5 lub niższa oznacza model nie lepszy od losowego, a 1,0 – doskonałe przewidywanie obu klas.

Warto dodać, że lasy losowe często wykazują satysfakcjonujące wyniki nawet z domyślnymi wartościami hiperparametrów, co może być jednym z powodów ich dużej popularności (Lovelace et al., 2019). Chociaż dostrojenie lasów losowych powinno poprawiać jakość modeli, korzyści ze strojenia są znacznie mniejsze w porównaniu do innych algorytmów uczenia maszynowego, takich jak maszyny wektorów nośnych (ang. *Support Vector Machines*, SVM (Boser et al., 1992)) (Probst et al., 2019) czy XGBoost (Chen et al., 2016).

3.3.3 Ocena jakości modeli

Ważnym krokiem w procesie uczenia maszynowego jest ocena jakości modelu. W tym celu można zastosować k -krotną walidację krzyżową (ang. *k -fold cross-validation*, CV), która zakłada, że obserwacje są od siebie niezależne (Pohjankukka et al., 2017). W standardowej k -krotnej walidacji krzyżowej dostępny zbiór uczący jest dzielony na k podzbiorów podobnej wielkości, gdzie *fold* odnosi się do liczby powstałych podzbiorów. Podział ten przeprowadza się poprzez losowe próbkowanie obserwacji ze zbioru uczącego się bez zastępowania. Model jest uczony na $k - 1$ podzbiorach, które razem tworzą zbiór uczący. Następnie model jest testowany na pozostałym podzbiorze, określonym jako zbiór walidacyjny, po czym mierzona jest jego jakość. Procedurę tę powtarza się, aż każdy z k podzbiorów zostanie użyty jako zbiór walidacyjny (Berrar, 2019). Kroswalidacja pozwala na ocenę ogólnej jakości modelu, uwzględniając różnorodność danych i pomaga uniknąć sytuacji, w której wyniki są mocno uzależnione od konkretnego podziału danych.

Obserwacje geograficzne, posiadające określone współrzędne, nie spełniają założenia niezależności danych ze względu na autokorelację przestrzenną (ang. *spatial autocorrelation*) (Pohjankukka et al., 2017). Ogólnie rzecz biorąc, dane przestrzenne wykazują autokorelację przestrzenną zgodnie z pierwszym prawem geografii, będącym jednocześnie podstawowym założeniem analizy geostatystycznej, według którego „Wszystko jest powiązane ze wszystkim innym, ale rzeczy bliskie są bardziej powiązane niż rzeczy odległe” (Tobler, 1970). Traktowanie zbiorów danych przestrzennych jak nieprzestrzennych prowadzi do zbyt optymistycznych wyników oceny jakości modeli (Brenning, 2005). Konsekwencją autokorelacji przestrzennej dla oceny jakości jest nadmierne dopasowanie



Rycina 3.3: Porównanie przestrzennego i losowego podziału zbioru danych na potrzeby walidacji krzyżowej jednego powtórzenia. Podział przestrzenny (górnny rząd) i losowy (dolny rząd). Niebieskie punkty reprezentują dane treningowe, a pomarańczowe dane testowe. Opracowanie własne na podstawie Schratz, 2021

klasyfikatorów do obserwacji uczących, jeśli obserwacje testowe (lub walidacyjne) nie są niezależne od zbioru uczącego (Brenning, 2012).

Przestrzenna walidacja krzyżowa (ang. *spatial cross-validation*) jest modyfikacją standardowej kroswalidacji, która zapobiega błędom w ocenie jakości modelu wynikającym z bliskości danych testowych i treningowych. Aby ograniczyć stronniczość w wynikach oceny dokładności predykcyjnej, w ramach przestrzennej walidacji krzyżowej wykorzystuje się przestrzennie odseparowane podzbiory danych, wprowadzając przestrzenną odległość pomiędzy zbiorem treningowym a testowym (Pohjankukka et al., 2017). Przykład takiego podejścia przedstawia rycina 3.3.

Macierz błędów (ang. *error matrix* lub *confusion matrix*) to tablica prezentująca wyniki klasyfikatora binarnego, podając informację o liczbie obiektów przypisanych do każdej z klas. Macierz błędów jest wynikiem porównania prognozy (klasyfikacji) z rzeczywistymi danymi, składającą się z czterech wartości reprezentujących różne kombinacje przewidywanych i rzeczywistych klas.

W macierzy błędów dwie wartości reprezentują predykcje pozytywne: przypadki prawdziwie pozytywne (ang. *true positive*, TP) opisują sytuacje, w których klasyfikator poprawnie przewidział daną klasę jako pozytywną, natomiast przypadki prawdziwie negatywne (ang. *true negative*, TN) opisują poprawne przewidywanie klasy negatywnej. Dwie pozostałe wartości informują o błędnie sklasyfikowanych przypadkach: przypadki fałszywie pozytywne (ang. *false positive*, FP) opisują sytuacje, w których klasyfikator błędnie wskazał klasę pozytywną, podczas gdy rzeczywiście była ona negatywna. Przypadki fałszywie negatywne (ang. *false negative*, FN) to natomiast sytuacje, w których klasyfikator błędnie przewidział klasę jako negatywną, gdy faktycznie była ona pozytywna.

Ocenę klasyfikatorów stworzonych w niniejszym badaniu przeprowadzono przy użyciu przestrzennej walidacji krzyżowej, macierzy błędów oraz trzech miar dostosowanych do klasyfikatorów binarnych:

- Precyza (ang. *precision*, inaczej *positive predictive value*) - określa jaką część wyników wskazanych przez klasyfikator jako pozytywne jest faktycznie pozytywna (Jaworski, 2013)
- Czułość (ang. *sensitivity*, inaczej *recall* lub *true positive rate*) - określa jaką część prawdziwie pozytywnych wyników wykrył klasyfikator (Jaworski, 2013)
- F1-score - średnia harmoniczna pomiędzy precyzją i czułością, umożliwiająca ocenę równowagi między tymi miarami, która w pewnym stopniu opisuje całościowo wynik. Miara ta nie uwzględnia wyników prawdziwie negatywnych (Żygierewicz, 2021).

Wartości każdej z wymienionych miar jakości modelu zawierają się w zakresie od 0 do 1, gdzie wartość 0 reprezentuje niską jakość modelu, natomiast wartość 1 odzwierciedla wysoki poziom jakości.

3.4 Ważność zmiennych

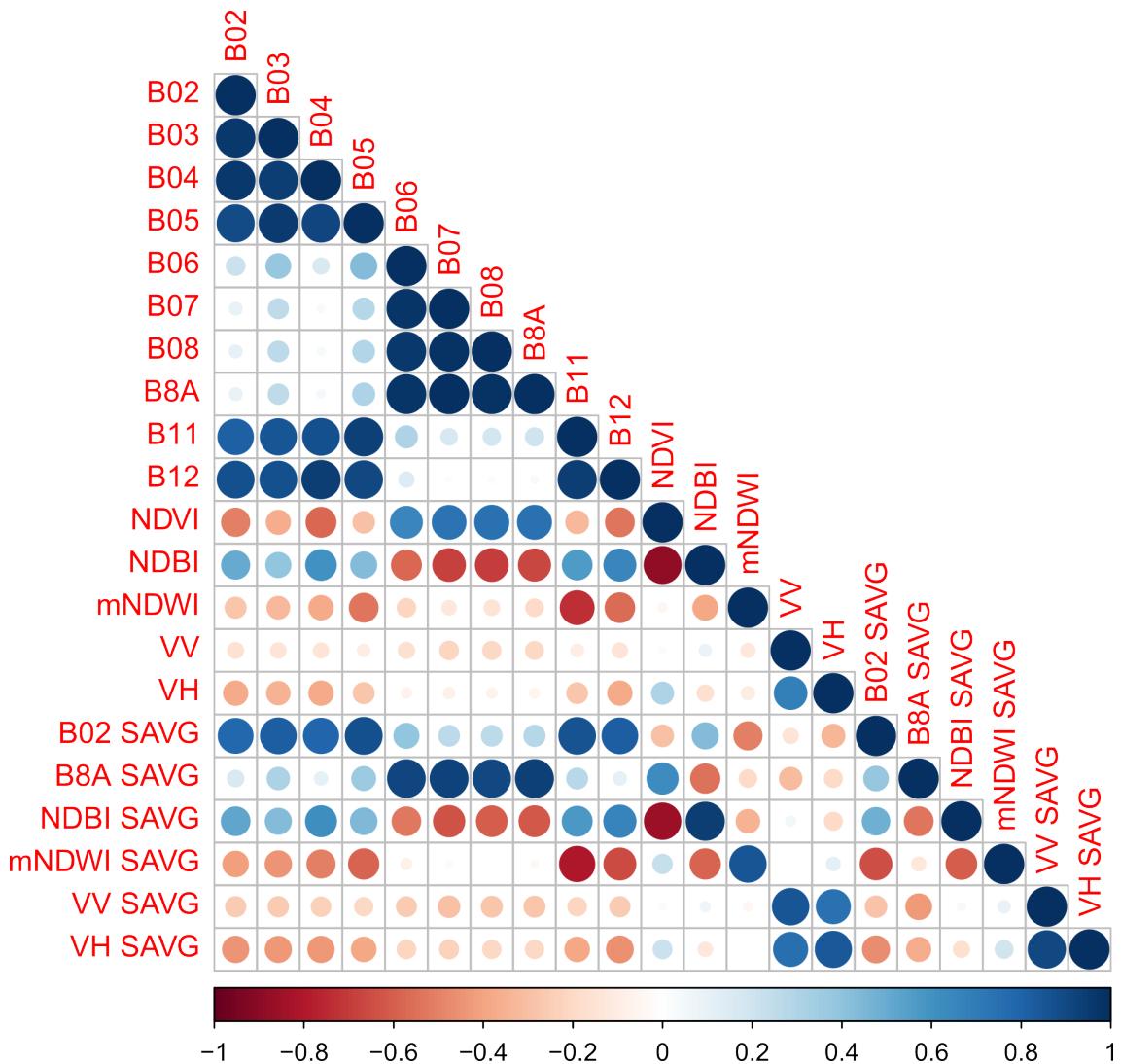
Ocena ważności zmiennych (ang. *variable importance*) jest elementem oceny jakości stworzonych modeli. Analiza wpływu poszczególnych zmiennych na dokładność modelu

pozwala ocenić, które zmienne są istotne dla przewidywań. Stosowanie zmiennych o niskiej mocy predykcyjnej może prowadzić do nadmiernego dopasowania (ang. *overfitting*) modelu lub obniżenia jego jakości. Dlatego istotny jest wybór odpowiednich zmiennych do trenowania modelu, unikając na przykład kolinearności predyktorów, czyli wysokiej korelacji między zmiennymi.

W lasach losowych ważność zmiennych można ocenić różnymi metodami, z których dwie najpopularniejsze to miara zanieczyszczenia Giniego (ang. *Gini impurity*) oraz metoda oparta na permutacji (Biecek, 2017). W celu określenia wartości zmiennych wykorzystywanych do identyfikacji farm fotowoltaicznych na podstawie danych teledetekcyjnych zastosowano metodę permutacji, która może być również używana do upraszczania i eksploracji modeli lub generowania wiedzy (Biecek et al., 2021).

Główna ideą metody opartej na permutacji jest pomiar tego, jak bardzo zmieni się dopasowanie modelu, gdy usunięty zostanie wpływ wybranej zmiennej lub grupy zmiennych. Jeśli zmienna jest istotna, permutacja jej wartości skutkuje pogorszeniem jakości modelu. Im większa zmiana dopasowania modelu, tym istotniejsza jest permutowana zmienna (Biecek et al., 2021). Metoda oparta na permutacji została pierwotnie zaproponowana przez Breimana (2001) dla lasów losowych, jednak jej prostota umożliwia zastosowanie permutacji do dowolnego modelu, a także porównywanie ważności zmiennych pomiędzy modelami o różnych strukturach (Biecek et al., 2021).

W przypadku silnie skorelowanych ze sobą zmiennych, brak jednej z nich niekoniecznie będzie negatywnie wpływać na jakość modelu, ponieważ inna, silnie skorelowana zmienna może zastąpić tę brakującą informację (Biecek et al., 2021). Macierz korelacji zmiennych użytych w tym badaniu (rycina 3.4) ukazuje silne związki między niektórymi zmiennymi. Wysoka korelacja występuje szczególnie między pasmami Sentinel-2 w zakresie widzialnym (kanały B02-04) a pasmami średniej podczerwieni (kanały B11-12). Zauważalna jest również wysoka korelacja między pasmami czerwieni krawędziowej (kanały B05-07) oraz bliskiej podczerwieni (kanały B08 i B8A). Wykorzystane polaryzacje Sentinel-1 również są ze sobą skorelowane. Wskaźniki teledetekcyjne wykazują zazwyczaj znaczną dodatnią lub ujemną korelację z pasmami, które zostały wykorzystane do



Rycina 3.4: Macierz korelacji wszystkich zmiennych wykorzystanych w badaniu

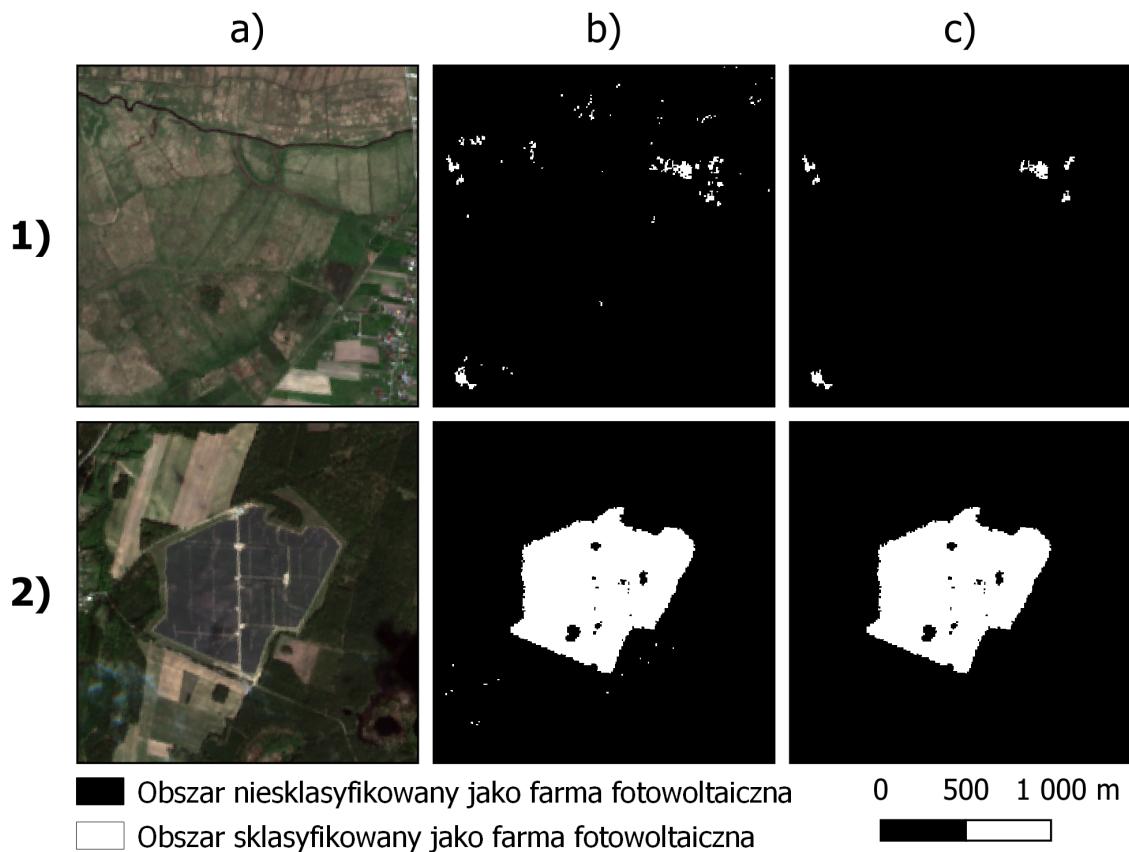
ich obliczenia. Tekstury obrazu charakteryzują się natomiast wysoką dodatnią korelacją ze zmiennymi, dla których zostały one określone.

3.5 Przetwarzanie końcowe

Ostatnim etapem procesu detekcji farm fotowoltaicznych było określenie metod przetwarzania końcowego (ang. *post-processing*), mających na celu udoskonalenie wyników wykrywania farm fotowoltaicznych. Z teoretycznego punktu widzenia, obszar zajmowany przez farmę fotowoltaiczną lub jej fragmenty powinien być większy niż powierzchnia pojedynczej komórki lub grupy kilku komórek. W celu usunięcia pozytywnych predykcji, które reprezentowały pojedyncze komórki lub obszary składające się z 10 lub mniej komórek (czyli o powierzchni mniejszej/równiej 1000 m^2) zastosowano sekwencję procesów przetwarzania, dzięki której udało się pozbyć tzw. efektu soli i pieprzu (ang. *salt-and-pepper effect*).

Przetwarzanie końcowe obejmowało wektoryzację rastrowych wyników predykcji dla każdego wariantu, a następnie selekcję pozytywnych predykcji oraz obliczenie powierzchni każdego wybranego obszaru. Na podstawie obliczonej powierzchni usunięto wszystkie pozytywne predykcje reprezentujące pojedyncze komórki lub obszary składające się z 10 lub mniej komórek, dzięki czemu wiele błędnych predykcji zostało usuniętych. W rezultacie uzyskano końcowy produkt wykrywania farm fotowoltaicznych oparty na danych teledetekcyjnych w formie wektorowej.

W celu stworzenia wizualizacji oraz porównania produktów końcowych z predykcjami niepoddanymi etapowi przetwarzania końcowego produkty wektorowe zostały ponownie przetworzone do formy rastrowej. Efekty zastosowanego przetwarzania końcowego przedstawia rycina 3.5.



Rycina 3.5: Kompozycja RGB Sentinel-2 (a), wstępne wyniki detekcji PV przed przetwarzaniem końcowym (b), PV wykryte po przetwarzaniu końcowym (c) w przypadku niepoprawnej (1) i poprawnej (2) predykcji dla wariantu nr 1

3.6 Oprogramowanie

3.6.1 QGIS

QGIS (QGIS Development Team, 2023), to wieloplatformowe i wolne oprogramowanie o otwartym kodzie źródłowym przeznaczone do przetwarzania danych przestrzennych, rozwijane od 2002 roku (Hejmanowska et al., 2020; Flenniken et al., 2020). Algorytmy przetwarzania danych przestrzennych zebrane w oprogramowaniu QGIS umożliwiają manipulację danymi rastrowymi oraz wektorowymi, a także prowadzenie analiz i wizualizację wyników (Hejmanowska et al., 2020). Oprogramowanie QGIS oferuje również możliwość korzystania z wielu zewnętrznych programów, tzw. wtyczek (ang. *plug-in*) rozszerzających jego funkcjonalność (Hejmanowska et al., 2020). W repozytorium wtyczek znaleźć można narzędzia do zarządzania danymi, przetwarzania obrazów, wizualizacji,

czy wykonania dodatkowych zadań, takich jak np. nadawanie georeferencji czy klasyfikacja zobrazowań satelitarnych (Hejmanowska et al., 2020). QGIS dostarcza także zaawansowane narzędzia do digitalizacji, umożliwiające rysowanie i edytowanie obiektów wektorowych oraz pozwala na przeglądanie danych przestrzennych dostępnych w Internecie za pomocą usług sieciowych, takich jak WMS, WMTS czy XYZ Tiles.

Oprogramowanie QGIS zostało wykorzystane w tej pracy do stworzenia zestawu danych referencyjnych poprzez wizualną interpretację ortofotomapy oraz mozaik satelitarnych, na podstawie których dokonano digitalizacji farm fotowoltaicznych istniejących na obszarze badań. QGIS posłużył również do stworzenia końcowych wizualizacji (map) prezentowanych w niniejszej pracy.

3.6.2 Sentinel-1 Toolbox i SNAP

Przetwarzanie danych pochodzących z misji Sentinel-1 umożliwia zestaw narzędzi Sentinel-1 Toolbox (S1TBX) (ESA, 2023a), przeznaczony do obsługi danych radarowych. Zestaw narzędzi S1TBX zawiera narzędzia do kalibracji, filtrowania plamek (tzw. efektu pieprzu i soli), koregistracji, ortorektyfikacji, mozaikowania, konwersji danych, polarymetrii czy interferometrii (ESA, 2023i). Sentinel-1 Toolbox jest opracowywany dla ESA przez firmę Array we współpracy z DLR, Brockmann Consult i OceanDataLab (ESA, 2023i).

Snap (ESA, 2023n), czyli Sentinel Application Platform to platforma oprogramowania rozwijana wspólnie przez firmy Brockmann Consult, SkyWatch i C-S na zlecenie Europejskiej Agencji Kosmicznej, przeznaczona do naukowego wykorzystania misji optycznych i mikrofalowych Sentinel (ESA, 2023o,m). Oprogramowanie SNAP zawiera zestawy narzędzi do wizualizacji, przetwarzania oraz analizy danych teledetekcyjnych, a także umożliwia tworzenie łańcuchów procesów przetwarzania danych zdefiniowanych przez użytkownika (Hejmanowska et al., 2020; Moskolaï et al., 2022).

3.6.3 Środowisko języka R

Czynności związane z końcowym przygotowaniem danych wejściowych oraz bezpośrednio z uczeniem maszynowym zostały wykonane z wykorzystaniem środowiska języka

R (R Core Team, 2023). R to wieloplatformowy język programowania o otwartym kodzie źródłowym do obliczeń statystycznych i wizualizacji danych. Dzięki dużej liczbie pakietów R obsługuje również statystki geoprzestrzenne, modelowanie oraz wizualizację danych przestrzennych (Lovelace et al., 2019). W pracy wykorzystane zostało zintegrowane środowisko programistyczne (ang. *Integrated Development Environment*, IDE) RStudio (RStudio Team, 2020) przeznaczone dla języka R. Poza standardowymi możliwościami środowiska R, w procesie pracy wykorzystane zostały pakiety stworzone przez społeczność R w celu rozszerzenia funkcjonalności tego języka. Do operacji na danych rastrowych zastosowano pakiety *terra* (Hijmans, 2023) i *stars* (Pebesma, 2023b), natomiast do przetwarzania danych wektorowych używany był pakiet *sf* (Pebesma, 2023a). Obliczanie tekstury obrazu Sum Average wyprowadzonej z macierzy współwystępowania poziomu szarości (ang. *gray-level co-occurrence matrix*, GLCM) zostało wykonane przy pomocy pakietu *GLCMTextures* (Illich, 2023). Losowe generowanie danych przestrzennych umożliwił pakiet *spatstat.random* (Baddeley et al., 2023b) z rodziny pakietów *spatstat* (Baddeley et al., 2023a). Do przeprowadzenia analizy oraz predykcji opartej o elementy uczenia maszynowego wykorzystano pakiet *mlr3* (Lang et al., 2023a), w ramach którego użyty został algorytm lasów losowych zaimplementowany w pakiecie *ranger* (Wright et al., 2023). Podczas strojenia parametrów algorytmów uczenia maszynowego, korzystano również z pakietu *paradox* (Lang et al., 2023b), umożliwiającego definiowanie granic przestrzeni szukania optymalnych wartości hiperparametrów. Do obliczeń związanych z teksturami obrazu oraz uczeniem maszynowym wykorzystano pakiet *future* (Bengtsson, 2023), umożliwiający równoległe (wielowątkowe) przetwarzanie wyrażeń R, skracające czas realizacji zadań w stosunku do przetwarzania sekwencyjnego. Wizualizacje przedstawione w niniejszej pracy zostały utworzone za pomocą pakietów *ggplot2* (Wickham et al., 2023), *spectralR* (Prylutskyi, 2023), *corrplot* (Wei et al., 2021) oraz *cowplot* (Wilke, 2024).

Rozdział 4

Wyniki

4.1 Ocena jakości modeli

Dla każdego z sześciu wariantów (zestawów danych) przeprowadzono osobną zagnieżdżoną k -krotną przestrzenną walidację krzyżową, zgodnie z procedurą przedstawioną w sekcji 3.3.3. Warianty odnoszą się do różnych zestawów zmiennych stworzonych w celu ustalenia optymalnego zestawu danych na potrzeby detekcji farm fotowoltaicznych przy wykorzystaniu danych teledetekcyjnych (tabela 3.1). Zagnieżdżona przestrzenna walidacja krzyżowa dla każdego z wariantów składała się z kilku etapów, obejmujących optymalizację hiperparametrów oraz ocenę jakości.

Proces strojenia został skonfigurowany tak, aby generować 1 000 modeli dla jednego podzbioru w celu określenia optymalnych hiperparametrów. Powtarzając tę procedurę dla każdego z pięciu ustalonych podzbiorów, uzyskano łącznie 5 000 modeli w ramach jednego powtórzenia. W celu identyfikacji optymalnych hiperparametrów założono dwadzieścia iteracji wymienionych powyżej działań, co doprowadziło do stworzenia łącznie 100 000 modeli.

Zoptymalizowane parametry modelu zostały następnie wykorzystane do oszacowania jakości modelu, co wymagało dopasowania dodatkowych 100 modeli (5 podzbiorów * 20 powtórzeń). W rezultacie całkowita liczba stworzonych modeli wykorzystanych do oceny jakości i dostrajania hiperparametrów dla jednego wariantu wyniosła 100 100.

Tabela 4.1: Średnie wyniki oceny jakości modeli uzyskane podczas przestrzennej walidacji krzyżowej

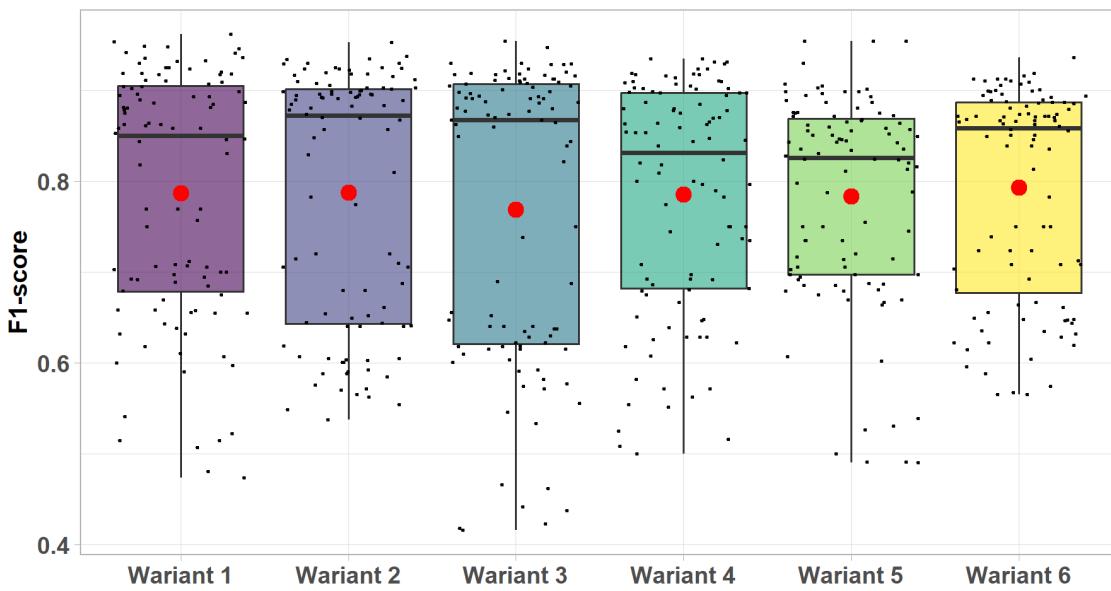
	Wariant 1	Wariant 2	Wariant 3	Wariant 4	Wariant 5	Wariant 6
Precyza	0.8852	0.9001	0.8746	0.9112	0.9034	0.9180
Czułość	0.7413	0.7296	0.7181	0.7134	0.7081	0.7151
F1-score	0.7872	0.7875	0.7691	0.7855	0.7837	0.7929

Ocenę klasyfikatorów stworzonych dla każdego z sześciu wariantów (tabela 3.1) przeprowadzono przy użyciu trzech miar jakości opisanych w sekcji 3.3.3. Średnie wyniki tych miar, oszacowane na podstawie przestrzennej walidacji krzyżowej, przedstawia tabela 4.1.

Ogólnie średnie wartości miar jakości dla każdego z wariantów są zbliżone. Prawdopodobnie wynika to z faktu, że w każdym wariantie dziesięć predyktorów uwzględniało dane dotyczące reflektancji Sentinel-2 w jego poszczególnych kanałach. Każdy wariant różnił się jednak od pozostałych zestawem dodatkowych zmiennych.

Na podstawie średnich wartości miar jakości przedstawionych w tabeli 4.1, najlepsze dopasowanie uzyskano, wykorzystując wszystkie predyktory (wariant nr 6). Ten wariant osiągnął najwyższe średnie oceny w dwóch z trzech zastosowanych miar (precyza i F1-score). Najlepszy wynik czułości uzyskał natomiast wariant nr 1. W przypadku dwóch z trzech miar najniższe wyniki odnotowano dla wariantu nr 3 (precyza i F1-score), a największą czułością cechuje się natomiast wariant nr 5.

Rozrzut pomiędzy najlepszym a najgorszym wynikiem dla precyzji wynosi 0,0434, dla czułości 0,0332, a dla F1-score 0,0238. Średnie wyniki precyzji czterech z sześciu klasyfikatorów przekroczyły wartość 0,90 (0,8746 dla najgorszego wariantu nr 3, 0,9180 dla najlepszego wariantu nr 6). Wszystkie klasyfikatory uzyskały wyniki czułości na poziomie wyższym niż 0,70, kształtuje się w zakresie od 0,7081 do 0,7413. Niemniej jednak, takie wyniki czułości wskazują na dość przeciętne radzenie sobie z wykrywaniem wyników prawdziwie pozytywnych (ang. *true positive*). Średnie wyniki miary F1-score opisującej całkowitą jakość modelu poprzez ocenę balansu pomiędzy precyją a czułością przekroczyły wartość 0,75 dla każdego wariantu, mieszcząc się w zakresie pomiędzy 0,7691, a 0,7929.



Rycina 4.1: Rozkład wyników miary F1-score na podstawie 100 modeli dla każdego wariantu. Czerwony punkt reprezentuje średnią wyników F1-score dla danego wariantu, a czarne punkty wyniki dla poszczególnych modeli

Stabilność klasyfikatora można ocenić na podstawie rozrzutu wyników oceny jakości poszczególnych modeli. Im mniej różnorodne są wyniki oceny jakości modeli w klasyfikatorze, tym bardziej jest on stabilny. Wykres pudełkowy (rycina 4.1) wskazuje, że wariant o najniższym średnim wyniku F1-score charakteryzuje się również największym rozrzutem wyników dla poszczególnych modeli. Największą stabilnością przewiduyań charakteryzuje się natomiast wariant nr 6.

Wykres pudełkowy przedstawiający rozkład wyników miary F1-score na podstawie 100 modeli dla każdego wariantu (rycina 4.1) ujawnia jej bimodalny rozkład wyników w każdym z wariantów. Charakteryzuje się on grupowaniem wyników wokół dwóch wartości oddzielonych od siebie, przy czym grupy te nie są równe pod względem liczebności, ale są wyraźnie większe od pozostałych. Miara F1-score, będąc średnią harmoniczną precyzji i czułości modelu, jest silnie zależna od tych dwóch niezależnych miar.

Analiza wyników precyzji ujawniła, że wartości odstające w wynikach tej miary pochodzą z modeli, gdzie podzbiór testowy składał się głównie z obserwacji negatywnych, przy niewielkiej liczbie obserwacji pozytywnych. W przypadku gdy odsetek obserwacji

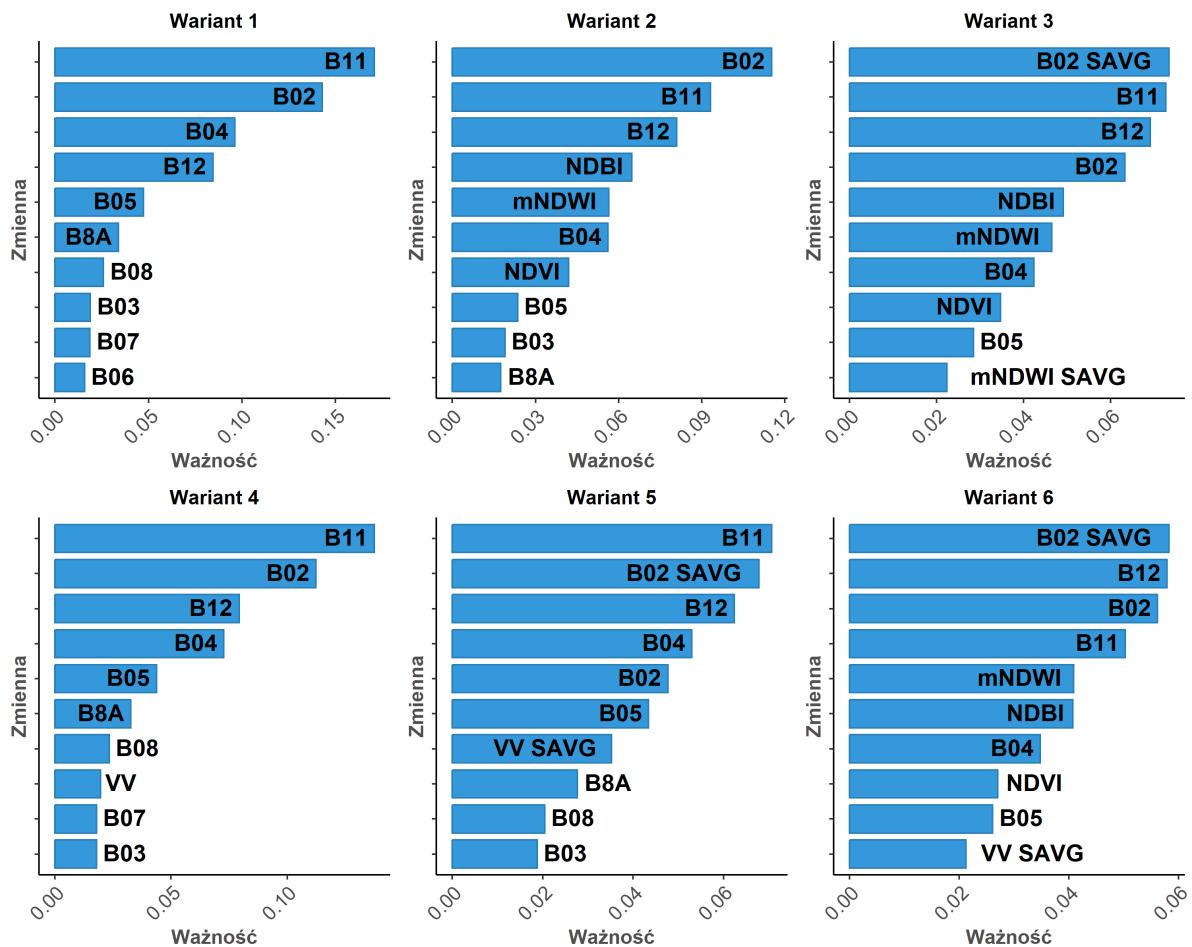
pozytywnych był wyższy, wyniki precyzji dla danego modelu były bliskie lub równe maksymalnej wartości jaką może osiągnąć ta miara.

W przeciwnieństwie do precyzji, czułość wydaje się być mniej podatna na wpływ proporcji obserwacji pozytywnych do negatywnych w zbiorze treningowym. Wyniki czułości dla poszczególnych modeli w każdym z wariantów, podobnie jak w przypadku F1-score, mają rozkład bimodalny. Na ogół rozkład wyników czułości jest bardzo zbliżony do rozkładu wyników miary F1-score, a największą różnicą jest obecność wartości odstających w wynikach czułości dla dwóch wariantów, co nie ma miejsca w przypadku F1-score. Oznacza to prawdopodobnie, że bimodalny rozkład wyników F1-score jest definiowany przez rozkład wyników czułości.

4.2 Ważność zmiennych

Dla każdego z sześciu wariantów (tabela 3.1) przeprowadzono ocenę ważności zmiennych z wykorzystaniem metody opartej na permutacji, szczegółowo opisanej w sekcji 3.4. Ważność zmiennych dla każdego wariantu została posortowana w porządku malejącym i przedstawiona na rycinie 4.2, która prezentuje moc predykcyjną różnych zmiennych wejściowych. Na potrzeby wizualizacji przedstawiono 10 najważniejszych zmiennych dla każdego wariantu, co odpowiada liczbie predyktorów w wariantie o najmniejszej ilości zmiennych (wariant nr 1). Rycina 4.3 ilustruje permutowaną ważność zmiennych dla wariantu nr 6, obejmującego wszystkie dostępne zmienne (21 predyktorów).

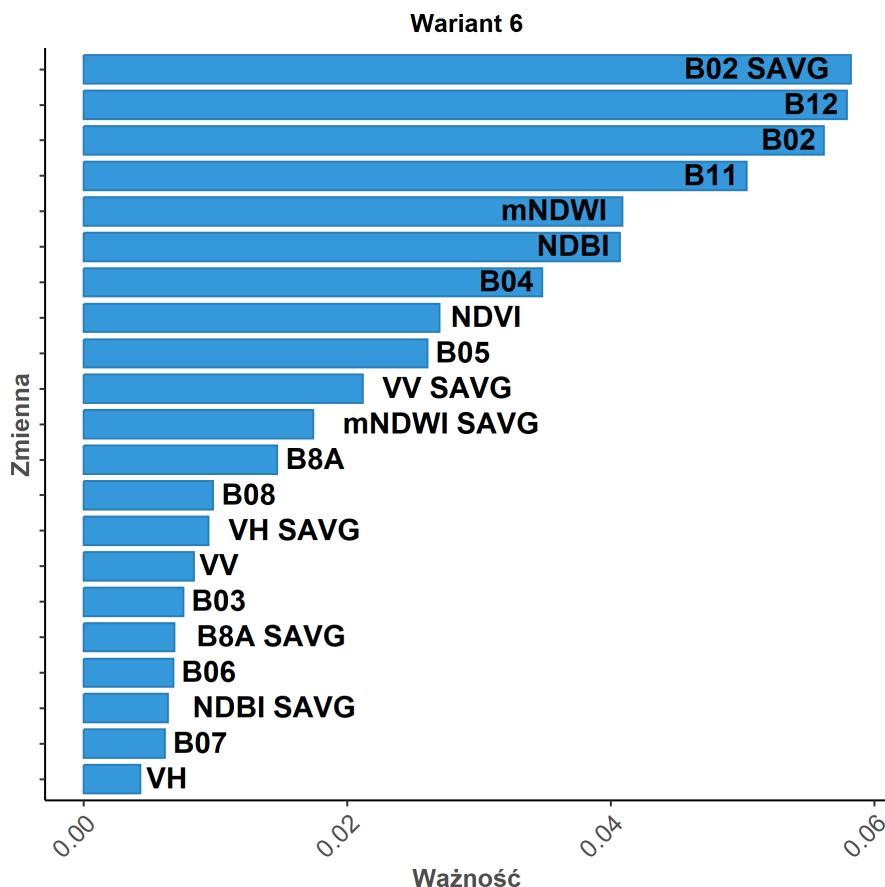
Pomijając aspekt związków i interakcji pomiędzy zmiennymi, które powodowały subtelne różnice w kolejności znaczenia zmiennych dla poszczególnych wariantów, zauważalny jest podział zmiennych na pięć grup według ich istotności w detekcji farm fotowoltaicznych. Z rycin 4.2 i 4.3 wynika, że największe znaczenie spośród wszystkich predyktorów miała grupa czterech zmiennych: tekstura średniej sumy kanału niebieskiego (B02 SAVG), kanały średniej podczerwieni (SWIR1 (B11) i SWIR2 (B12)) oraz kanał niebieski (B02). Trochę niższą istotność w kontekście wykrywania farm fotowoltaicznych miały trzy kolejne zmienne: znormalizowany zmodyfikowany różnicowy wskaźnik wody (mNDVI), znormalizowany różnicowy wskaźnik obszarów zabudowanych (NDBI) oraz kanał czerwony (B04). Trzecią grupę stanowiły dwie zmienne: znormalizowany różnicowy



Rycina 4.2: Permutowana ważność 10 najważniejszych zmiennych dla każdego wariantu

wskaźnik wegetacji (NDVI) oraz jeden z kanałów czerwieni krawędziowej (tzw. RedEdge, B05), a do czwartej grupy zaliczyć można tekstury średniej sumy dla polaryzacji VV i wskaźnika mNDWI (VV SAVG, mNDWI SAVG) oraz kanał bliskiej podczerwieni (B8A). Najmniejsze znaczenie przy detekcji farm fotowoltaicznych miały kanał zielony (B03), pozostałe kanały czerwieni krawędziowej (B06 i B07), kanał bliskiej podczerwieni (B08), obie wykorzystane polaryzacje (VV i VH) oraz tekstury średniej sumy dla kanału B8A, polaryzacji VH oraz wskaźnika NDBI (B8A SAVG, VH SAVG i NDBI SAVG).

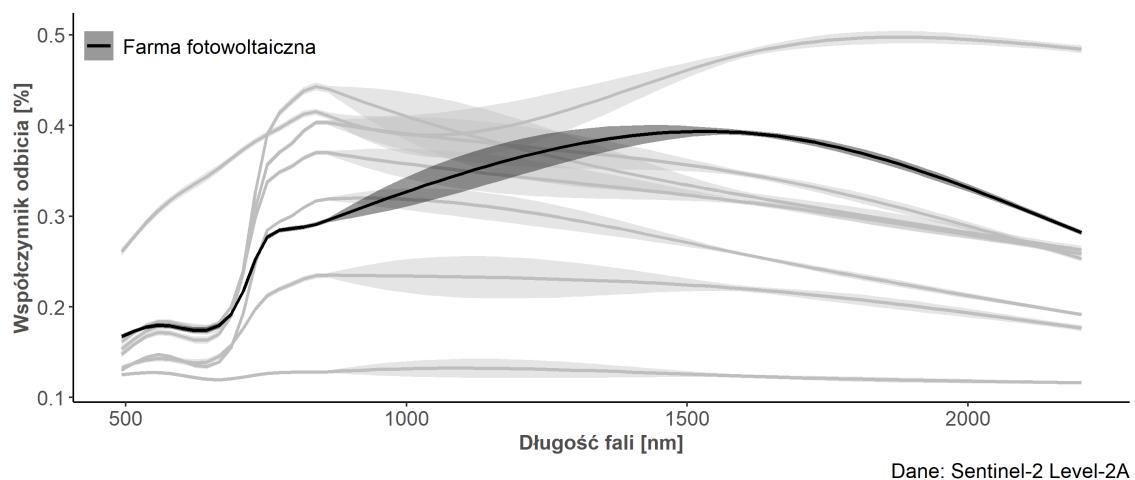
Uzyskane wyniki oceny ważności zmiennych wskazują na dość spore znacznie wskaźników spektralnych, głównie mNDWI oraz NDBI w kontekście wykrywania farm fotowoltaicznych. Niskie znaczenie w tym zastosowaniu wykazują natomiast dane radarowe pochodzące z misji Sentinel-1 oraz tekstury średniej sumy dla tych zmiennych.



Rycina 4.3: Permutowana ważność zmiennych dla wariantu nr 6

Spośród sześciu obliczonych tekstur średniej sumy wskazanych przez Wang et al. (2022) jako istotne przy detekcji farm fotowoltaicznych na podstawie danych Sentinel-1, Sentinel-2 i algorytmu Random Forest, jedynie tekstura średniej sumy dla kanału niebieskiego wykazywała się znaczącym wpływem na wynik klasyfikacji. Pozostałe tekstury wskazywały przeciętną lub niską istotność w tym konkretnym zadaniu. Ważność trzech z obliczonych tekstur (B02 SAVG, VV SAVG i VH SAVG) była wyższa niż ocena ważności odpowiadających im danych pierwotnych. W przypadku pozostałych trzech tekstur (B8A SAVG, mNDWI SAVG i NDBI SAVG), uzyskane wyniki były niższe niż wyniki pierwotnych danych teledetekcyjnych.

Charakterystyka spektralna farm fotowoltaicznych (rycina 4.4) wyjaśnia znaczenie kanałów B02, B11 i B12 w kontekście detekcji farm fotowoltaicznych na podstawie danych teledetekcyjnych. Krzywa spektralna instalacji fotowoltaicznych wskazuje na spektralne zróżnicowanie tej powierzchni względem innych typów pokrycia terenu w zakresie pasm



Dane: Sentinel-2 Level-2A

Rycina 4.4: Krzywa odbicia spektralnego farm fotowoltaicznych na tle innych typów pokrycia terenu

średniej podczerwieni, do których należą kanały B11 (1610 nm) i B12 (2190 nm). Farmy fotowoltaiczne charakteryzują się wyższym współczynnikiem odbicia od większości innych powierzchni w zakresie fal niebieskich (kanał B02, 493 nm). Oznacza to, że odpowiedź spektralna instalacji fotowoltaicznych w tym zakresie różnicuje je od innych typów pokrycia terenu.

4.3 Wyniki klasyfikacji po przetwarzaniu końcowym

Sekcja 4.1 omawiała wyniki jakości klasyfikacji na podstawie próby, a wyniki oceny jakości modelu uzyskane na niewielkiej części populacji mogą znacznie różnić się od rezultatów uzyskanych podczas klasyfikacji całej populacji, obejmującej wszystkie obserwacje na badanym obszarze.

W celu oceny rzeczywistej jakości klasyfikacji dla całej populacji po procesie przetwarzania końcowego dokonano porównania wyników predykcji każdego wariantu (tabela 3.1) ze zbiorem referencyjnym. Referencyjny zbiór danych obejmował wszystkie farmy fotowoltaiczne, które udało się określić na obszarze kafla Sentinel-2 o oznaczeniu 33UWV w czasie wykonywania użytych zobrazowań (8 maja 2023 roku). Dane te zostały zdigitalizowane na podstawie ortofotomapy oraz mozaik satelitarnych. Porównanie wymagało przekształcenia zdigitalizowanych farm fotowoltaicznych do postaci rastrowej, przyjmując siatkę 10-metrowych kanałów wykorzystanej sceny Sentinel-2.

Tabela 4.2: Wyniki klasyfikacji uzyskane po procesie przetwarzania końcowego

Wariant ^a	Liczba wykrytych poligonów	Suma wykrytej powierzchni [ha]	Poprawnie wykryta powierzchnia farm na podstawie macierzy pomyłek [ha]
Zbiór referencyjny ^b	210	345.04	-
1	650	516.57	290.86
2	263	335.71	291.01
3	294	344.28	289.30
4	733	487.05	287.24
5	505	406.26	286.29
6	323	394.95	291.10

^a Patrz: tabela 4.1

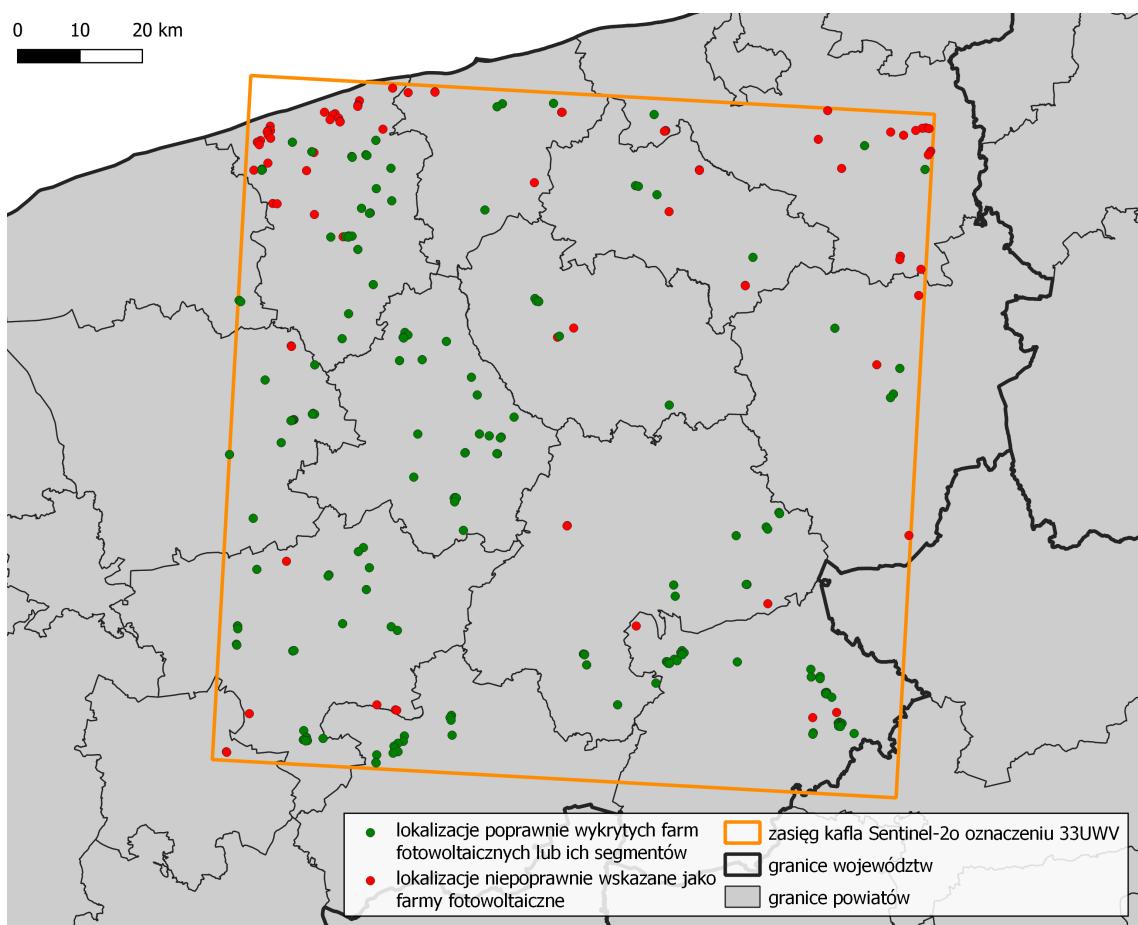
^b Zdigitalizowane farmy fotowoltaiczne na podstawie ortofotomapy i mozaik satelitarnych

Wyniki klasyfikacji dotyczące liczby wykrytych farm fotowoltaicznych lub ich oddzielnych segmentów oraz sumy wykrytej powierzchni po procesie przetwarzania końcowego zostały przedstawione w tabeli 4.2. Pierwszy wiersz odnosi się do zbioru referencyjnego, który obejmował wszystkie farmy fotowoltaiczne określone na obszarze kafla Sentinel-2 o oznaczeniu 33UWV w czasie wykonania użytych zobrazowań (8 maja 2023 roku).

Analiza wyników w tabeli wskazuje na znaczne zróżnicowanie zarówno pod względem liczby wykrytych poligonów (odpowiadających poszczególnym segmentom farm fotowoltaicznych), jak i sumy wykrytej powierzchni. Warianty nr 2 i 3 osiągnęły wyniki najbardziej zbliżone do rzeczywistych, prezentując zbliżoną sumę powierzchni wskazanej jako farmy fotowoltaiczne. Niemniej jednak oba warianty różnią się znacząco pod względem liczby wykrytych poligonów.

Pod względem poprawnie wykrytej powierzchni elektrowni fotowoltaicznych, ustalonej na podstawie macierzy pomyłek, wyniki poszczególnych wariantów są stosunkowo zbliżone i mieszczą się w zakresie od 286,29 ha do 291,10 ha. Wariant nr 6 okazał się najlepszy pod względem wykrytej powierzchni, natomiast najgorszy wynik uzyskał wariant nr 5.

Warto jednak zaznaczyć, że wyniki predykcji po etapie przetwarzania końcowego są dość mocno niekompletne, ponieważ każdemu z wariantów brakuje około 55 ha powierzchni farm fotowoltaicznych względem zbioru referencyjnego. Wartość ta stanowi ponad



Rycina 4.5: Lokalizacje poprawnie i niepoprawnie wykrytych farm fotowoltaicznych lub ich segmentów dla wariantu nr 2

15% powierzchni elektrowni fotowoltaicznych na obszarze badań i sygnalizuje pewne ograniczenia w skuteczności klasyfikacji.

Rozmieszczenie poprawnie i niepoprawnie wykrytych farm fotowoltaicznych dla wariantu drugiego przedstawia rycina 4.5. Z ilustracji wynika, że większość elektrowni fotowoltaicznych lub ich segmentów została właściwie zidentyfikowana, a błędnie sklasyfikowane lokalizacje farm fotowoltaicznych występują głównie w północnej części badanego obszaru.

4.3.1 Ocena jakości klasyfikacji dla populacji

Ocena jakości klasyfikacji dla całego obszaru badań (populacji) została przeprowadzona zgodnie z podejściem opisany w sekcji 3.3.3. Wykorzystane podejście, podobnie jak

Tabela 4.3: Wyniki oceny jakości klasyfikacji uzyskane dla całej populacji

Wariant ^a	Precyzja	Czułość	F1-score
1	0.5631	0.8414	0.6747
2	0.8668	0.8419	0.8542
3	0.8403	0.8369	0.8386
4	0.5898	0.8310	0.6899
5	0.7047	0.8282	0.7615
6	0.7371	0.8421	0.7861

^a Patrz: tabela 4.1

ocena jakości modeli, opiera się na analizie macierzy błędów, która umożliwiła obliczenie trzech miar jakości: precyzji, czułości oraz F1-score.

Wyniki oceny jakości klasyfikacji dla populacji każdego wariantu przedstawia tabela 4.3, która wskazuje na znaczne rozbieżności w kontekście precyzji. Precyzja ocenia w pewnym sensie skłonność modelu do przeuczania się, określając jaka część wyników wskazanych przez klasyfikator jako pozytywne jest faktycznie pozytywne w rzeczywistości. Przeuczanie występuje w sytuacji, gdy klasyfikator wskazuje farmy fotowoltaiczne w miejscowościach, gdzie faktycznie one nie występowały. Najlepszą precyzję osiągnęły warianty nr 2 i 3, gdzie wartości tej miary wynoszą odpowiednio 0,8668 i 0,8403. Warianty nr 4 i 1 prezentują natomiast niską precyzję, wynoszącą odpowiednio 0,5898 i 0,5631, co sugeruje duże przeuczenie tych dwóch klasyfikatorów. Wyniki precyzji poniżej wartości 0,60 wskazują, że ponad 40% obszarów wskazanych jako farmy fotowoltaiczne w rzeczywistości nimi nie jest. Oba warianty z najniższymi wynikami precyzji oparte były wyłącznie na pierwotnych danych teledetekcyjnych. Wariant nr 1 składał się wyłącznie ze zmiennych będących reflektancją kanałów Sentinel-2, natomiast wariant nr 4 oprócz reflektancji zawierał surowe dane o współczynniku rozproszenia wstecznego dla obu polaryzacji Sentinel-1. Wyniki wariantów zawierających pochodne danych teledetekcyjnych (wskaźniki teledetekcyjne, tekstury obrazu) uzyskały znacznie wyższe wyniki precyzji, co wskazuje na duże znaczenie informacji pochodnej w kontekście wykrywania farm fotowoltaicznych na podstawie danych teledetekcyjnych. Sugeruje to również, że wykorzystanie danych pochodnych zmniejsza tendencję modeli do przeuczania się.

W zakresie czułości wyniki są znacznie mniej zróżnicowane, utrzymując się między 0,8282 a 0,8421. Czułość określa, jaką część rzeczywistych przypadków wykrył klasyfikator, czyli jaki ułamek farm znajdujących się w referencyjnym zbiorze danych został wykryty. Najwyższy wynik czułości osiągnął wariant nr 6, a najniższy - wariant nr 5.

Jak wspomniano w sekcji 3.3.3, miara F1-score jest średnią harmoniczną precyzyji i czułości, używaną, gdy obie te miary są równie istotne. Miara F1-score opisuje całkowicie wynik, a ponieważ wyniki czułości dla poszczególnych wariantów są zbliżone do siebie, to precyza będzie miała kluczowy wpływ na ostateczną ocenę jakości klasyfikacji. Najwyższym wynikiem miary F1-score charakteryzuje się wariant nr 2 (0,8542), a próg wartości 0,80 przekroczył również wariant nr 3. Stosując wyłącznie surowe dane teledetekcyjne, warianty nr 1 i 4 nie przekraczają progu 0,70 dla miary F1-score, podczas gdy pozostałe warianty, zawierające dane pochodne, uzyskały znacznie wyższe wartości.

Ogólnie rzecz biorąc, wyniki sugerują, że korzystanie z pochodnych danych teledetekcyjnych istotnie redukuje skłonność modelu do przeuczenia, poprawiając jednocześnie jego jakość w kontekście detekcji farm fotowoltaicznych.

4.3.2 Wizualna kontrola wyników klasyfikacji

Zgodnie z sekcją dotyczącą przestrzennej oceny jakości (sekcja 4.3.1), modele stworzone w ramach niniejszego badania wykazują dobrą skuteczność w identyfikacji farm fotowoltaicznych, co ilustruje również rycina 4.6, przedstawiająca przykłady poprawnych przewidywań dla wariantu nr 2. Dla porównania, na rycinie przedstawiono także instalacje fotowoltaiczne generujące energię elektryczną, pochodzące z badania przeprowadzonego przez Kruitwagena et al. (2021), wskazującego istniejące konstrukcje fotowoltaiczne na świecie na dzień 30 września 2018 roku. Wysokorozdzielcze obrazy satelitarne na rycinie 4.6 (kolumna a) pochodzą z różnych okresów. Obraz satelitarny 2a został pozyskany później niż dane użyte do detekcji farm fotowoltaicznych w niniejszym badaniu (8 maja 2023 roku), natomiast obrazy 4a i 4b pochodzą sprzed tego okresu.

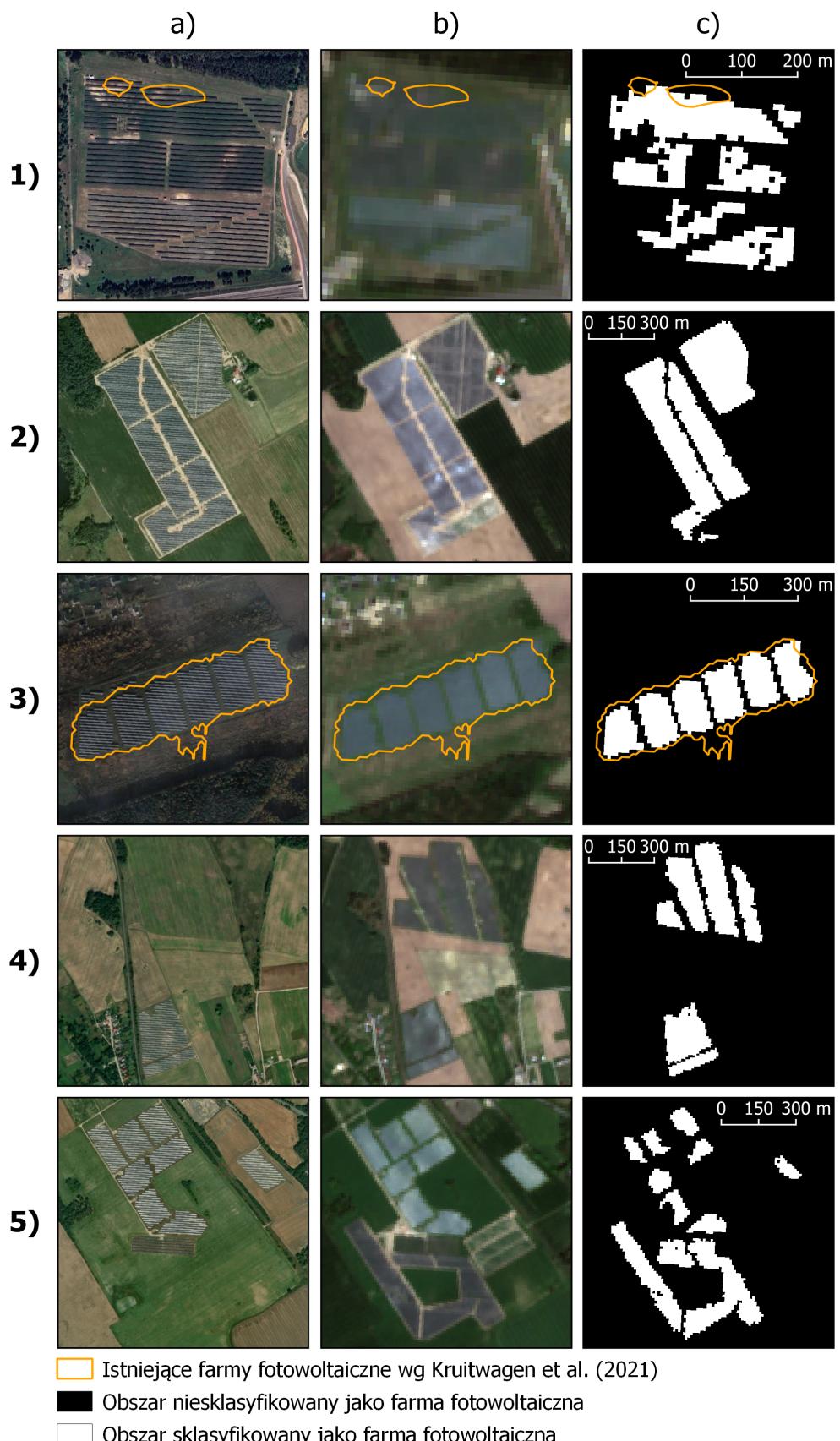
W niektórych przypadkach wewnętrz wykrytych instalacji fotowoltaicznych pojawiły się fałszywie negatywne predykcje (ang. *false negative*), co przedstawiają ryciny 4.6 1c i 3.5 2c. Na kompozycji RGB Sentinel-2 możemy zaobserwować w miejscach tych błędnych

wskazań różnice w jasności komórek względem otaczającej instalacji fotowoltaicznej lub zróżnicowanie powierzchni pod panelami fotowoltaicznymi. Rycina 4.6 5c pokazuje, że w niektórych przypadkach poprawne wskazania powierzchni farm fotowoltaicznych są niepełne, pomimo jednolitego wyglądu instalacji na kompozycji RGB Sentinel-2.

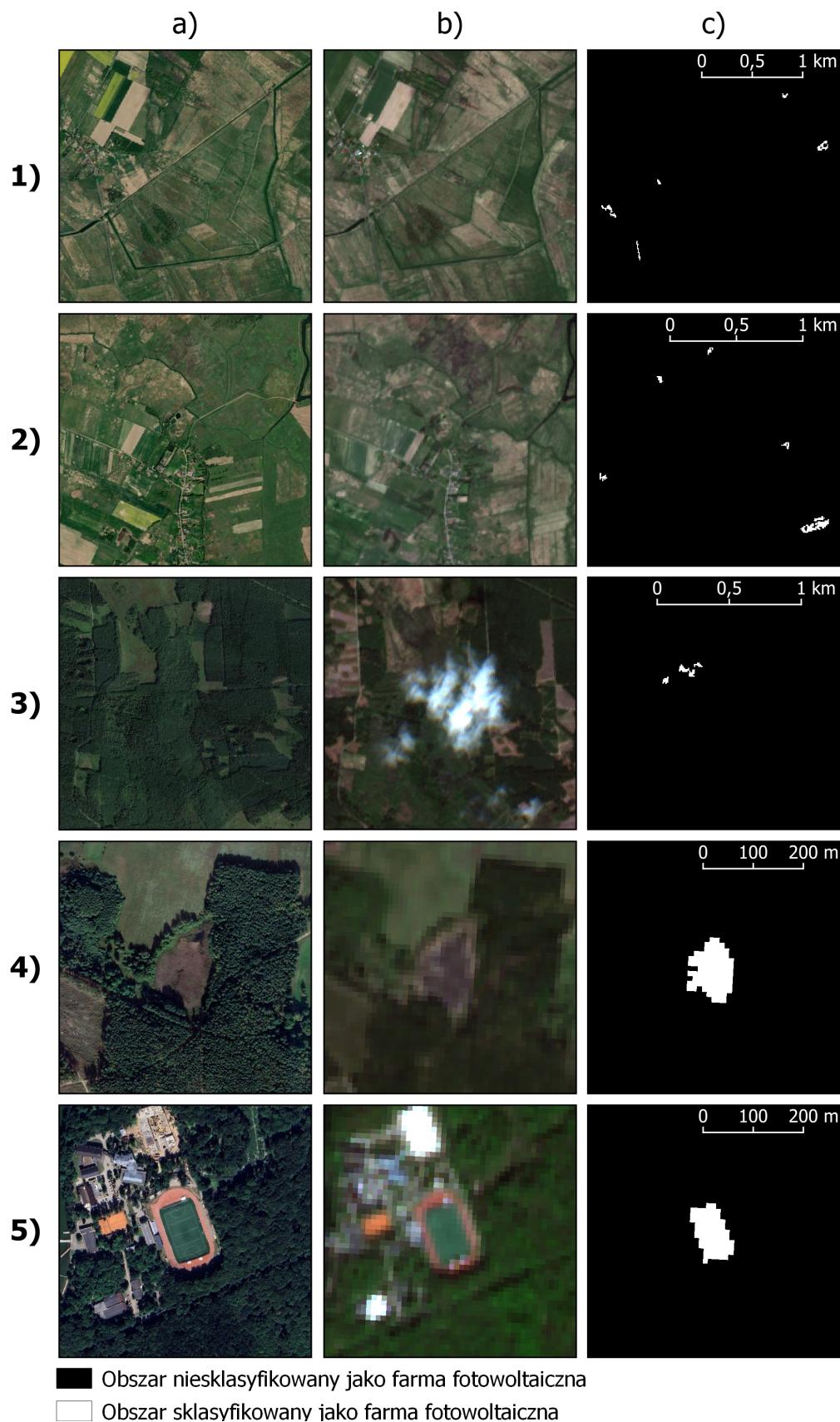
Problemem stworzonych modeli jest ich tendencja do przeuczania się na niektórych typach pokrycia terenu i użytkowania ziemi, co zostało szerzej opisane w sekcji dotyczącej losowania próbek (sekcja 3.2). Zaproponowane metody przetwarzania końcowego, omówione w sekcji 3.5, poprawiają wyniki predykcji. Niemniej jednak, w zależności od wariantu, nadal występują mniejsze lub większe błędy w klasyfikacji. Przykłady fałszywie pozytywnych przewidywań (ang. *false positive*) zostały przedstawione na rycinach 4.7 (dla wariantu nr 2) oraz 3.5 (dla wariantu nr 1). Wybrane lokalizacje ilustrują typowe błędy modeli na różnych typach pokrycia terenu i użytkowania ziemi.

Mimo zastosowania dodatkowych próbek negatywnych na drogach, jak wskazano w sekcji 3.2, błędne pozytywne przewidywania w tych miejscach nadal występują, co ilustruje pierwszy rząd ryciny 4.7. Pierwszy, drugi oraz czwarty rząd ryciny 4.7 oraz pierwszy rząd ryciny 3.5 pokazują natomiast, że błędne predykcje obejmują także obszary użytków rolnych (pola uprawne, łąki i pastwiska) oraz nieużytków, szczególnie w miejscach, gdzie istnieje gęsta sieć melioracyjna. W wyniku klasyfikacji wariantu nr 2 pojawił się nietypowy przypadek błędного sklasyfikowania boiska sportowego jako instalacji fotowoltaicznej, które ze względu na powierzchnię większą niż 1000 m² nie zostało skorygowane przez zastosowane metody przetwarzania końcowego. Trzeci rząd ryciny 4.7 oraz drugi rząd ryciny 3.5 sugerują, że błędne predykcje mogą występować również na obszarach zachmurzonych.

Należy zaznaczyć, że pomimo wykorzystania próbek z tych samych lokalizacji do trenowania każdego z wariantów, różne modele wykazują zróżnicowaną skuteczność klasyfikacji w zależności od typów pokrycia terenu i użytkowania ziemi. Jest to prawdopodobnie rezultat wpływów różnych zmiennych, gdzie niektóre z nich wspomagały decyzje w przypadku konkretnego typu pokrycia terenu, podczas gdy inne miały przeciwny efekt.



Rycina 4.6: Wyniki klasifikacji wariantu nr 2 po procesie przetwarzania końcowego. Porównanie wysokorozdzielczych obrazów satelitarnych (a) oraz kompozycji RGB Sentinel-2 (b) z przykładami prawdziwie pozytywnych przewidywań (ang. true positive)



Rycina 4.7: Wyniki klasyfikacji wariantu nr 2 po procesie przetwarzania końcowego. Porównanie wysokorozdzielczych obrazów satelitarnych (a) oraz kompozycji RGB Sentinel-2 (b) z przykładami fałszywie pozytywnych przewidywań (ang. false positive) (c)

Rozdział 5

Podsumowanie

Badanie wykazało, że możliwe jest wykrywanie farm fotowoltaicznych na podstawie danych teledetekcyjnych, wykorzystując do tego celu dane satelitarne z misji Sentinel-1 i Sentinel-2. Wyniki uzyskane z wariantów zawierających tylko pierwotne dane teledetekcyjne z tych misji uzyskały dość przeciętne wyniki detekcji farm fotowoltaicznych. Wykorzystanie dodatkowych zmiennych, takich jak wskaźniki teledetekcyjne czy tekstury obrazu, istotnie poprawiło wyniki klasyfikacji. Dane radarowe z misji Sentinel-1 i ich pochodne nie okazały się istotne w procesie detekcji farm fotowoltaicznych. W odróżnieniu od tego, multispektralne dane z misji Sentinel-2, wraz z ich pochodnymi, wykazały się kluczowym elementem detekcji farm fotowoltaicznych przy wykorzystaniu danych satelitarnych.

Optymalnym wariantem okazał się wariant nr 2, składający się z kanałów Sentinel-2 oraz trzech wykorzystanych wskaźników teledetekcyjnych (NDVI, NDBI i mNDWI), który uzyskał najwyższe wyniki oceny jakości po ostatecznej klasyfikacji i przetwarzaniu końcowym. Zbliżone wyniki jakości uzyskał wariant nr 3, który był rozszerzeniem wariantu drugiego o dodatkowe zmienne w postaci tekstur obrazu dwóch kanałów Sentinel-2 i dwóch wskaźników teledetekcyjnych. Jeżeli dwa klasyfikatory charakteryzują się zbliżonymi wynikami jakości, to zazwyczaj lepszym wyborem jest ten prostszy, czyli zawierający mniej zmiennych. W tym kontekście wariant nr 2 również wypada lepiej od wariantu nr 3, ponieważ ma mniejszą liczbę predyktorów.

Warto zwrócić uwagę na istotne różnice między oceną jakości modeli na próbie a ostateczną oceną klasyfikacji na całej populacji po procesie przetwarzania końcowego. Chociaż wyniki oceny jakości modeli (tabela 4.1) sugerują bardzo zbliżoną skuteczność wszystkich modeli, to ocena przeprowadzona na pełnej populacji ujawnia duże różnice między poszczególnymi wariantami (tabela 4.3). Największe różnice zauważać można w wynikach precyzji, która ostatecznie warunkuje ostateczną ocenę jakości klasyfikatora po przeprowadzeniu predykcji i przetwarzaniu końcowym dla każdego wariantu. Jeśli ocena jakości modeli dla klasyfikacji nie zostałaby przeprowadzona, moglibyśmy sądzić, że dodatkowe zmienne, czyli produkty pochodne dostępne w poszczególnych wariantach, nie mają znaczącego wpływu na wyniki detekcji farm fotowoltaicznych.

Kwestią, której nie udało się rozwiązać w badaniu, jest obecność bimodalnego rozkładu wyników czułości w każdym z wariantów. Bimodalny rozkład wyników miary F1-score, będącej średnią harmoniczną precyzji i czułości modelu, prawdopodobnie wynika z obecności bimodalnego układu wyników czułości. Dalsze badania związane z detekcją farm fotowoltaicznych na podstawie danych satelitarnych będą wymagały wyjaśnienia tej kwestii, ponieważ może ona istotnie wpływać na otrzymane wyniki jakości modeli. W przypadku kontynuacji prac w tej tematyce ważne będzie również ustalenie minimalnego odsetka obserwacji pozytywnych w podzbiorze testowym, aby uniknąć występowania wartości odstających w wynikach precyzji.

Porównując wyniki najlepszych wariantów z istniejącymi bazami danych dotyczącymi elektrowni fotowoltaicznych, uzyskane wyniki są raczej zadowalające. Globalna baza danych elektrowni (*Global Power Plant Database*) (WRI, [2021](#)), stworzona przez Byersa et al. ([2018](#)) i udostępniona w czerwcu 2021 roku, wskazuje na istnienie zaledwie 9 elektrowni fotowoltaicznych w całej Polsce na wspomniany okres. Ujednolicone globalne zbiory danych dotyczące lokalizacji farm wiatrowych i słonecznych oraz mocy produkowanej (*Harmonised global datasets of wind and solar farm locations and power*), opracowane na podstawie danych OpenStreetMap przez Dunnetta et al. ([2020](#)), sugerują, że do końca roku 2018 na obszarze badania występowały jedynie 2 instalacje fotowoltaiczne, które zajmowały łącznie powierzchnię 1,99 hektara. Istnieje również jeden zbiór danych,

którego sposób stworzenia był najbardziej zbliżony do omawianego w niniejszym badaniu. Kruitwagen et al. (2021), wykorzystując zdjęcia satelitarne SPOT-6/7 i Sentinel-2 w połączeniu z metodami uczenia maszynowego, wskazał istniejące konstrukcje fotowoltaiczne na świecie na dzień 30 września 2018 roku. Według wyników tego badania, na obszarze niniejszego badania znajdowało się w tamtym momencie jedynie 10 poligonów reprezentujących segmenty farm fotowoltaicznych o łącznej powierzchni 14,67 hektara.

Wizualna kontrola obszarów wskazanych na rycinach potwierdziła, że modele w większości przypadków skutecznie dokonują rozróżnienia farm fotowoltaicznych i pozostałych obszarów podczas klasyfikacji. Niemniej jednak, w wynikach predykcji pojawiło się kilka powtarzających się błędów, które można w prosty sposób wyeliminować. W celu poprawy wyników identyfikacji farm fotowoltaicznych w kolejnych badaniach, zaleca się zastosowanie masek eliminujących obszary chmur i ich cieni. Dodatkowo, warto rozważyć wprowadzenie dodatkowego etapu w procesie przetwarzania końcowego, który wykluczałby pozytywne predykcje wzdłuż dróg, zgodnie z sugestią Ortiza et al. (2022).

W celu rozszerzenia obszaru detekcji farm fotowoltaicznych do obszaru całej Polski w przyszłych badaniach, zaleca się wykorzystanie platformy obliczeniowej Google Earth Engine (Gorelick et al., 2017), umożliwiającej przeprowadzanie analiz geoprzestrzennych w infrastrukturze Google. Platforma Google Earth Engine gromadzi ogólnodostępne dane satelitarne z ponad trzydziestu lat, a jej zasób jest codziennie aktualizowany i rozszerzany. Zaimplementowane w Google Earth Engine algorytmy dostarczają kompleksowy zestaw narzędzi, umożliwiając odtworzenie i rozwinięcie niniejszego badania na obszar całego kraju. Dodatkowo, wykorzystanie obliczeń w chmurze pozwoli na rozszerzenie badania o wykorzystanie dodatkowych tekstur obrazu, umożliwiając jednocześnie testowanie różnych konfiguracji rozmiarów ruchomego okna i liczby poziomów kwantyzacji w celu znalezienia optymalnych ustawień tych parametrów. Z uwagi na rosnący czas obliczeń wraz ze zwiększaniem liczby poziomów szarości oraz rozmiarów ruchomego okna, w niniejszej pracy nie zdecydowano się na sprawdzenie działania innych tekstur obrazu z różnymi konfiguracjami powyższych parametrów.

Bibliografia

- Attema, E, P Snoeij, M Davidson, N Flouri, G Levrini, B Rommen, and B Rosich (2008). The European GMES Sentinel-1 Radar Mission. In: vol. 1, pp.I–94.
- Baddeley, A, R Turner, and E Rubak (2023a). *spatstat: Spatial Point Pattern Analysis, Model-Fitting, Simulation, Tests*. R package version 3.0-7. <http://spatstat.org/>.
- Baddeley, A, R Turner, E Rubak, and T Davies (2023b). *spatstat.random: Random Generation Functionality for the spatstat Family*. R package version 3.2-2. <http://spatstat.org/>.
- Becker, M (2023). *mlr3tuningspaces: Search Spaces for mlr3*. R package version 0.4.0. <https://CRAN.R-project.org/package=mlr3tuningspaces>.
- Belgiu, M and L Drăguț (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* **114**, 24–31.
- Bengtsson, H (2023). *future: Unified Parallel and Distributed Processing in R for Everyone*. R package version 1.33.1. <https://CRAN.R-project.org/package=future>.
- Berrar, D (2019). “Cross-Validation”. In: *Encyclopedia of Bioinformatics and Computational Biology*. Ed. by S Ranganathan, M Gribkov, K Nakai, and C Schönbach. Oxford: Academic Press, pp.542–545. <https://www.sciencedirect.com/science/article/pii/B978012809633820349X>.
- Biecek, P (2017). *Przewodnik po pakiecie R*. Wydanie czwarte rozszerzone. Oficyna Wydawnicza GiS, p. 395. <https://pbiecek.github.io/Przewodnik/index.html>.
- Biecek, P and T Burzykowski (2021). *Explanatory Model Analysis*. Chapman and Hall/CRC, New York. <https://pbiecek.github.io/ema/>.
- Bischl, B, R Sonabend, L Kotthoff, and M Lang, eds. (2024). *Applied Machine Learning Using mlr3 in R*. CRC Press. <https://mlr3book.mlr-org.com>.

- Boser, BE, IM Guyon, and VN Vapnik (1992). A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp.144–152.
- BP p.l.c. (2021). *Full report: Statistical Review of World Energy*. Data dostępu: 02.02.2024.
<https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2021-full-report.pdf>.
- Braun, A (2020). *SAR-based Landcover Classification with Sentinel-1 GRD Products*. SkyWatch Space Applications Inc. <http://step.esa.int>.
- Breiman, L (2001). Random Forests. *Machine Learning* **45**, 5–32.
- Brenning, A (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences* **5**(6), 853–862.
- Brenning, A (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: *2012 IEEE International Geoscience and Remote Sensing Symposium*, pp.5372–5375.
- Byers, L, J Friedrich, R Hennig, A Kressig, X Li, C McCormick, and LM Valeri (2018). A global database of power plants. *World Resources Institute* **18**.
- Chen, T and C Guestrin (2016). Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp.785–794.
- Drusch, M, U Del Bello, S Carlier, O Colin, V Fernandez, F Gascon, B Hoersch, C Isola, P Laberinti, P Martimort, A Meygret, F Spoto, O Sy, F Marchese, and P Bargellini (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* **120**. The Sentinel Missions - New Opportunities for Science, 25–36.
- Dunnett, S, A Sorichetta, G Taylor, and F Eigenbrod (2020). Harmonised global datasets of wind and solar farm locations and power. *Scientific Data* **7**, 130.
- European Space Agency (2015). *Sentinel-2 User Handbook*. <https://sentinel-copernicus.eu/web/sentinel/user-guides/document-library/-/asset%5C%5Fpublisher/xlslt4309D5h/content/sentinel-2-user-handbook>.

- European Space Agency (2023). *Copernicus: Sentinel-1*. Accessed on 2023-12-12. <https://www.eoportal.org/satellite-missions/copernicus-sentinel-1#performance-specifications>.
- European Space Agency (5, 2023a). *ESA Sentinel-1 Toolbox*. Version 9.0.4. <http://step.esa.int/>.
- European Space Agency (2023b). *Mission ends for Copernicus Sentinel-1B satellite*. Accessed on 2023-12-12. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/Mission_ends_for_Copernicus_Sentinel-1B_satellite.
- European Space Agency (2023c). *Sentinel-1 Acquisition Modes: Stripmap*. Accessed on 2023-12-08. <https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar/acquisition-modes/stripmap>.
- European Space Agency (2023d). *Sentinel-1 Data Products*. Accessed on 2023-12-08. <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/data-products>.
- European Space Agency (2023e). *Sentinel-1 Instrument Payload*. Accessed on 2023-12-08. <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/instrument-payload>.
- European Space Agency (2023f). *Sentinel-1 Instrument Payload: Resolution and Swath*. Accessed on 2023-12-18. <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/instrument-payload/resolution-swath>.
- European Space Agency (2023g). *Sentinel-1 Mission Objectives*. Accessed on 2023-12-08. <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/mission-objectives>.
- European Space Agency (2023h). *Sentinel-1 Revisit and Coverage*. Accessed on 2023-12-12. <https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/revisit-and-coverage>.
- European Space Agency (5, 2023i). *Sentinel-1 Toolbox*. Accessed on 2023-12-05. <https://sentinel.esa.int/web/sentinel/toolboxes/sentinel-1>.

BIBLIOGRAFIA

- European Space Agency (2023j). *Sentinel-1 User Guide: Definitions*. Accessed on 2023-12-18.
<https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar/definitions>.
- European Space Agency (2023k). *Sentinel-2 Mission Guide*. Accessed on 2023-12-14. <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>.
- European Space Agency (2023l). *Sentinel-2 User Guide: Spectral Resolution*. Accessed on 2023-12-18. <https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/resolutions/spectral>.
- European Space Agency (2023m). *SNAP*. Accessed on 2023-12-05. <https://step.esa.int/main/toolboxes/snap/>.
- European Space Agency (5, 2023n). *SNAP - ESA Sentinel Application Platform*. Version 9.0.8. <http://step.esa.int>.
- European Space Agency (5, 2023o). *SNAP Desktop*. Accessed on 2023-12-05. <https://github.com/senbox-org/snap-desktop>.
- European Space Agency (2024). *Sentinel-1 Product Types and Processing Levels: Level-2*. Accessed on 2024-01-22. <https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar/product-types-processing-levels/level-2>.
- Filipponi, F (2019). Sentinel-1 GRD Preprocessing Workflow. *Proceedings* **18**(1).
- Flenningen, JM, S Stuglik, and BV Iannone (2020). Quantum GIS (QGIS): An introduction to a free alternative to more costly GIS platforms. *EDIS* **2020**(2), 7.
- Gislason, PO, JA Benediktsson, and JR Sveinsson (2006). Random Forests for land cover classification. *Pattern Recognition Letters* **27**(4). Pattern Recognition in Remote Sensing (PRRS 2004), 294–300.
- Global Energy Observatory and Google and KTH Royal Institute of Technology in Stockholm and Enipedia and World Resources Institute (2021). *Global Power Plant Database v1.3.0*. Published on Resource Watch (<http://resourcewatch.org/>) and Google Earth Engine (<https://earthengine.google.com/>). Version 1.3.0. Accessed through Resource Watch on 2024-01-19.
- Główny Urząd Geodezji i Kartografii (2023). *Ortofotomapa*. Accessed on 2023-12-09. <https://www.geoportal.gov.pl/pl/dane/ortofotomapa-ortho/>.

- Gong, P, DJ Marceau, and PJ Howarth (1992). A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data. *Remote Sensing of Environment* **40**(2), 137–151.
- Gorelick, N, M Hancher, M Dixon, S Ilyushchenko, D Thau, and R Moore (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.
- Hall-Beyer, M (2017). *GLCM Texture: A Tutorial v. 3.0 March 2017*. <https://doi.org/10.13140/RG.2.2.12424.21767>.
- Hanssen, RF (2001). *Radar Interferometry Data Interpretation and Error Analysis*. Vol. 2. Springer Science & Business Media.
- Haralick, RM, K Shanmugam, and IH Dinstein (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics* (6), 610–621.
- Hejmanowska, B and P Wezyk (2020). *Dane satelitarne dla administracji publicznej*. Polska Agencja Kosmiczna.
- Hijmans, RJ (2023). *terra: Spatial Data Analysis*. R package version 1.7-65. <https://rspatial.org/>.
- Ilich, A (2023). *GLCMTexels: GLCM Textures of Raster Layers*. R package version 0.3.9, <https://github.com/ailich/GLCMTexels>. <https://ailich.github.io/GLCMTexels/>.
- Ismail, MH, H Pakhriazad, and M Shahrin (2009). Evaluating supervised and unsupervised techniques for land cover mapping using remote sensing data. *Geografia : Malaysian Journal of Society and Space*.
- Jaworski, W (2013). *Miary jakości*. https://www.mimuw.edu.pl/~wjaworski/SU/SU04_miary_jakosci.pdf.
- Komisja Europejska (2022). *REPowerEU: Plan prowadzący do szybkiego ograniczenia zależności od rosyjskich paliw kopalnych oraz do szybkiej transformacji ekologicznej*. Data dostępu: 30.01.2024. https://ec.europa.eu/commission/presscorner/detail/pl_ip_22_3131.
- Król-Nowak, A and K Kotarba (2022). *Podstawy uczenia maszynowego*. Wydawnictwa AGH. <https://books.google.pl/books?id=DZXNzwEACAAJ>.

- Kruitwagen, L, KT Story, J Friedrich, L Byers, S Skillman, and C Hepburn (2021). A global inventory of photovoltaic solar energy generating units. *Nature* **598**, 604–610.
- Kupidura, P (2019). The Comparison of Different Methods of Texture Analysis for Their Efficacy for Land Use Classification in Satellite Imagery. *Remote Sensing* **11**(10).
- Lang, M, B Bischl, J Richter, P Schratz, M Binder, F Pfisterer, R Sonabend, and M Becker (2023a). *mlr3: Machine Learning in R - Next Generation*. R package version 0.17.1, <https://github.com/mlr-org/mlr3>. <https://mlr3.mlr-org.com>.
- Lang, M, B Bischl, J Richter, X Sun, and M Binder (2023b). *paradox: Define and Work with Parameter Spaces for Complex Algorithms*. R package version 0.11.1, <https://github.com/mlr-org/paradox>. <https://paradox.mlr-org.com>.
- Lennert, M, G Antoniol, C Basco, M Ceccarelli, M Metz, and GRASS Development Team (2023). *r.texture - GRASS GIS manual*. Accessed on 2023-12-11. <https://grass.osgeo.org/grass83/manuals/r.texture.html>.
- Lesiv, M, L See, JC Laso Bayas, T Sturn, D Schepaschenko, M Karner, I Moorthy, I McCalum, and S Fritz (2018). Characterizing the Spatial and Temporal Availability of Very High Resolution Satellite Imagery in Google Earth and Microsoft Bing Maps as a Source of Reference Data. *Land* **7**(4).
- Lewiński S.; Aleksandrowicz, S (2012). Ocena możliwości wykorzystania tekstury w rozpoznaniu podstawowych klas pokrycia terenu na zdjęciach satelitarnych różnej rozdzielczości. *Archiwum Fotogrametrii, Kartografii i Teledetekcji* **23**, 229–237.
- Lovelace, R, J Nowosad, and J Muenchow (2019). *Geocomputation with R*. CRC Press.
- Mahesh, B (2019). Machine Learning Algorithms - A Review. en. **9**(1).
- Main-Knorn, M, B Pflug, J Louis, V Debaecker, U Müller-Wilm, and F Gascon (2017). Sen2Cor for Sentinel-2. In: *Image and Signal Processing for Remote Sensing XXIII*. Ed. by L Bruzzone. Vol. 10427. International Society for Optics and Photonics. SPIE, pp.37–48. <https://doi.org/10.1117/12.2278218>.
- Maxwell, AE, TA Warner, and F Fang (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing* **39**(9), 2784–2817. eprint: <https://doi.org/10.1080/01431161.2018.1433343>.

- Microsoft (2023). *Maps and Geolocation Data. Bing Maps API, SDKs and Services*. Accessed on 2023-12-09. <https://www.microsoft.com/en-us/maps/bing-maps/location-data>.
- Moskolaï, WR, W Abdou, A Dipanda, and Kolyang (2022). A Workflow for Collecting and Preprocessing Sentinel-1 Images for Time Series Prediction Suitable for Deep Learning Algorithms. *Geomatics* **2**(4), 435–456.
- Mumby, PJ and AJ Edwards (2002). Mapping marine environments with IKONOS imagery: enhanced spatial resolution can deliver greater thematic accuracy. *Remote Sensing of Environment* **82**(2), 248–257.
- OpenStreetMap contributors (2024). *Planet dump retrieved from https://planet.osm.org*. <https://www.openstreetmap.org>.
- Ortiz, A, D Negandhi, S Mysorekar, S Nagaraju, J Kiesecker, C Robinson, P Bhatia, A Khurana, J Wang, F Oviedo, and J Lavista Ferres (2022). An Artificial Intelligence Dataset for Solar Energy Locations in India. *Scientific Data* **9**.
- Pebesma, E (2023a). *sf: Simple Features for R*. R package version 1.0-15, <https://github.com/r-spatial/sf>. <https://r-spatial.github.io/sf/>.
- Pebesma, E (2023b). *stars: Spatiotemporal Arrays, Raster and Vector Data Cubes*. R package version 0.6-4. <https://CRAN.R-project.org/package=stars>.
- Phiri, D, M Simwanda, S Salekin, VR Nyirenda, Y Murayama, and M Ranagalage (2020). Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sensing* **12**(14).
- Plakman, V, J Rosier, and J van Vliet (2022). Solar park detection from publicly available satellite imagery. *GIScience & Remote Sensing* **59**(1), 462–481. eprint: <https://doi.org/10.1080/15481603.2022.2036056>.
- Planet Labs PBC (2019). *Planet Basemaps Product Specification*. Accessed on 2023-12-05. Planet. <https://assets.planet.com/products/basemap/planet-basemaps-product-specifications.pdf>.
- Planet Labs PBC (2024). *Planet Application Program Interface: In Space for Life on Earth*. Planet. <https://api.planet.com>.

- Pohjankukka, J, T Pahikkala, P Nevalainen, and J Heikkonen (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science* **31**(10), 2001–2019. eprint: <https://doi.org/10.1080/13658816.2017.1346255>.
- Probst, P, MN Wright, and AL Boulesteix (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* **9**(3), e1301. eprint: <https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1301>.
- Prylutskyi, O (2023). *spectralR: Obtain and Visualize Spectral Reflectance Data for Earth Surface Polygons*. R package version 0.1.3. <https://github.com/olehprylutskyi/spectralR/>.
- QGIS Development Team (2023). *QGIS Geographic Information System*. QGIS Association. <https://www.qgis.org>.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>.
- Rada Europejska, Rada Unii Europejskiej (2023a). *Energia ze źródeł odnawialnych: Rada przyjmuje nowe przepisy*. Data dostępu: 30.01.2024. <https://www.consilium.europa.eu/pl/press/press-releases/2023/10/09/renewable-energy-council-adopts-new-rules/>.
- Rada Europejska, Rada Unii Europejskiej (2023b). *Europejski zielony ład*. Data dostępu: 30.01.2024. <https://www.consilium.europa.eu/pl/policies/green-deal/>.
- RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA. <http://www.rstudio.com/>.
- Sarker, I (2021). *Machine Learning: Algorithms, Real-World Applications and Research Directions*.
- Schratz, P, J Muenchow, E Iturritxa, J Richter, and A Brenning (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling* **406**, 109–120.
- Sekulić, A, M Kilibarda, GB Heuvelink, M Nikolić, and B Bajat (2020). Random Forest Spatial Interpolation. *Remote Sensing* **12**(10).

- Sheykhmousa, RM and M Mahdianpari (2020). Support Vector Machine vs. Random Forest for Remote Sensing Image Classification: A Meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Sindayigaya, L and A Dey (2022). Machine Learning Algorithms: A Review. *Information Systems Journal*, vol. ISJ-RA-3392, sp 6.
- Tobler, WR (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* **46**, 234–240. (Visited on 01/05/2024).
- Tucker, CJ (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment* **8**(2), 127–150.
- Wang, JF, A Stein, BB Gao, and Y Ge (2012). A review of spatial sampling. *Spatial Statistics* **2**, 1–14.
- Wang, J, J Liu, and L Li (2022). Detecting Photovoltaic Installations in Diverse Landscapes Using Open Multi-Source Remote Sensing Data. *Remote Sensing* **14**(24).
- Wei, T and V Simko (2021). *corrplot: Visualization of a Correlation Matrix*. R package version 0.92. <https://github.com/taiyun/corrplot>.
- Wickham, H, W Chang, L Henry, TL Pedersen, K Takahashi, C Wilke, K Woo, H Yutani, and D Dunnington (2023). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.4.4. <https://CRAN.R-project.org/package=ggplot2>.
- Wilke, CO (2024). *cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2*. R package version 1.1.3. <https://wilkelab.org/cowplot/>.
- Wright, MN, S Wager, and P Probst (2023). *ranger: A Fast Implementation of Random Forests*. R package version 0.16.0, <https://github.com/imbs-hl/ranger>. <http://imbs-hl.github.io/ranger/>.
- Xu, H (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing* **27**(14), 3025–3033. eprint: <https://doi.org/10.1080/01431160600589179>.
- Zha, Y, J Gao, and S Ni (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing* **24**(3), 583–594. eprint: <https://doi.org/10.1080/01431160304987>.

Zhang, X, M Zeraatpisheh, MM Rahman, S Wang, and M Xu (2021). Texture Is Important in Improving the Accuracy of Mapping Photovoltaic Power Plants: A Case Study of Ningxia Autonomous Region, China. *Remote Sensing* **13**(19).

Żygierewicz, J (2021). *Uczenie maszynowe i sztuczne sieci neuronowe: Ocena jakości klasyfikacji*. Accessed on 2024-01-04. https://brain.fuw.edu/index.php/Uczenie_maszynowe_i_sztuczne_sieci_neuronowe/Wyk%C5%82ad_Ocena_jako%C5%9Bci_klasyfikacji.