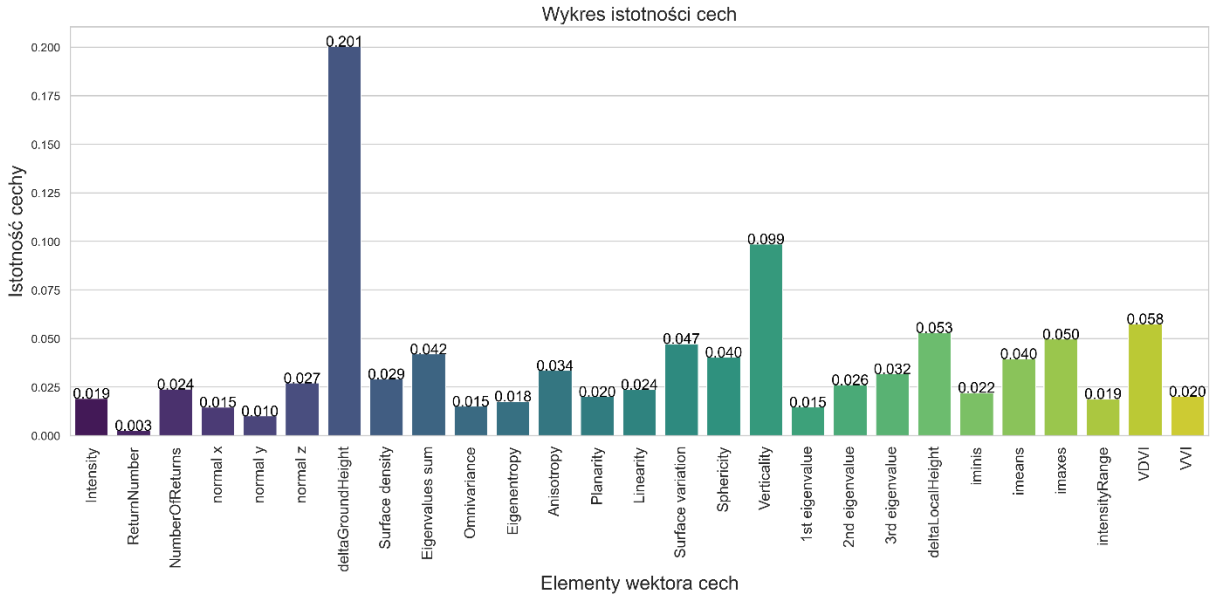
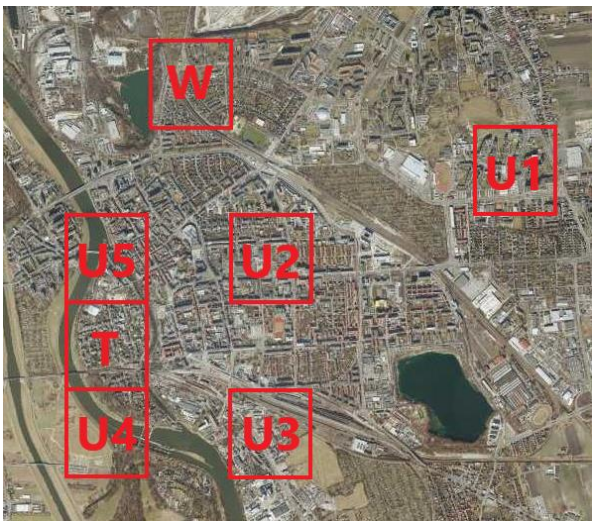


Rys. 21 Przykładowy wykres istotności cech.

4 Charakterystyka danych testowych

Dane testowe pochodzą z województwa opolskiego z miasta Opole. Zostały one pobrane ze strony geoportal.gov.pl z zakładki numeryczny model pokrycia terenu. Pochodzą one z 2022 roku i wykonane są w układzie PL-EVRF2007-NH. Zbiór danych stanowi 5 uczących obszarów, jeden walidacyjny oraz jeden testowy. Pola uczące obejmują odpowiednio 70,56%, 15,61%, 13,83% wszystkich pól. Każdy z obszarów ma powierzchnię około 32 ha, a nominalna gęstość chmury wynosi 12 pkt/m^2. Poniższy rysunek (Rys. 22) przedstawia rozmieszczenie wykorzystywanych obszarów.



Rys. 22 Rozmieszczenie danych uczących, walidacyjnych i testowych w mieście Opole. Odpowiednio symbole: U – dane uczące, W – dane walidacyjne oraz T – dane testowe.

4.1 Metodyka pracy

Dane były wstępnie sklasyfikowane, dodatkowo obszary uczące, walidacyjne i testowe zostały poddane ponownej obróbce w celu dokładniej klasyfikacji. Manualna korekta klasyfikacji ma na celu stworzenie dokładnego modelu, który posłuży do sklasyfikowania obszaru testowego. Dodatkowo, zostały wybrane klasy, jakie będzie posiadała każda chmura punktów. Poniższa Tabela 3 przedstawia wszystkie dane wraz z ilością punktów w danych klasach oraz procentowy udział danego pola w całym zbiorze w rozbiciu na klasy:

Tabela 3 Liczność punktów w poszczególnych danych uczących, walidacyjnych i testowych oraz procentowy udział danego pola w całym zbiorze w rozbiciu na klasy.

Id klasy	Nazwa klasy	Ucząca 1	Ucząca 2	Ucząca 3	Ucząca 4	Ucząca 5	Walidacyjna	Testowa
1	Unclassified	156926	193953	398074	69970	147565	111211	240002
	Udział [%]	11,91	14,72	30,21	5,31	11,20	8,44	18,21
2	Powierzchnia terenu	3788211	3165671	3559911	3589811	2466632	3507089	2834036
	Udział [%]	16,53	13,82	15,54	15,67	10,77	15,31	12,37
4	Wegetacja	608488	793935	564981	1063095	740250	1312800	1085610
	Udział [%]	9,86	12,87	9,16	17,23	12,00	21,28	17,60
6	Zabudowa	1077486	1506676	892049	200412	1253519	899831	1025750
	Udział [%]	15,72	21,98	13,01	2,92	18,28	13,13	14,96
7	Niskie szumy	1698	2247	1106	680	1754	1295	1 539
	Udział [%]	16,46	21,78	10,72	6,59	17,00	12,55	14,91
9	Woda	0	0	0	124303	234167	54069	27369
	Udział [%]	0,00	0,00	0,00	28,26	53,23	12,29	6,22
18	Wysokie szumy	185	121	5	37	313	26	31
	Udział [%]	25,77	16,85	0,70	5,15	43,59	3,62	4,32
	Łączna ilość punktów:	5632994	5662603	5416126	5048308	4844200	5886321	5214337

Klasy wegetacji zostały scalone w jedną klasę, ponieważ w większości przypadków wyróżnia się tylko jedną klasę wegetacji. Cała klasyfikacja została przeprowadzona w programie Agisoft Metashape Professional. Dane uczące zostały scalone do jednego pliku w formacie .laz i następnie został obliczony wektor cech. Utworzono go w programie

CloudCompare i Spyder w języku programowania Python. Dla danych walidacyjnych i testowych również został stworzony identyczny wektor cech.

Kolejnym krokiem była optymalizacja wzorca klasyfikatora, która polegała na odpowiednim doborze parametrów algorytmu. Następnie została przeprowadzona optymalizacja wektora cech opartego tylko i wyłącznie na elementach bazujących na sąsiedztwie. W celu ustawienia najlepszych parametrów zostały przeprowadzone testy, w których porównywano uzyskiwane dokładności. Po wyborze najlepszych parametrów, do wektora cech zostały dodane pozostałe cechy, które nie bazowały na sąsiedztwie. Ostatecznym krokiem optymalizacji było odrzucenie najmniej istotnych cech na podstawie ich istotności. Wytrenowany został model klasyfikatora na zoptymalizowanych parametrach algorytmu i zoptymalizowanym wektorze cech. Finalnie, zostały przeprowadzone testy poprawności klasyfikatora.

4.2 Narzędzia użyte podczas budowy klasyfikatora

Podczas tworzenia modelu klasyfikatora w oparciu o algorytm Random Forest użyto odpowiednich środowisk, programów oraz bibliotek wspomagających programy. Do wykonania manualnej klasyfikacji wykorzystano program Agisoft Metashape Professional 2.0.1. Podczas wyliczania cech wektora wykorzystano program CloudCompare 2.13alpha, Spyder 5.3.3 oraz język programowania Python 3.9.12. Do wyliczenia cech w środowisku Spyder wykorzystano następujące biblioteki: scipy 1.9.3 oraz NumPy 1.23.5. W tym samym środowisku została wykonana klasyfikacja oraz trenowanie modelu klasyfikatora. Wykorzystano wówczas dodatkowe biblioteki tj.: Pandas 1.4.4, time 3.9, pickle 3.14, matplotlib 3.6.2, scikit-learn 1.0.2 (Pedregosa i inni, 2011). Do stworzenia wykresów przedstawiających analizy dodatkowo została użyta biblioteka Seaborn 0.12.2.

Przetwarzanie danych, optymalizacja klasyfikatora oraz klasyfikacja ostatecznym modelem wykonana została na poniższym zestawie sprzętowym:

- procesor Intel Core i5-12400F,6 rdzeni, 12 wątków, częstotliwość 2,5GHz,
- pamięć RAM 32GB, DDR4, 3200MHz, 16CL.

4.3 Optymalizacja wzorca klasyfikatora

Podczas optymalizacji wzorca klasyfikatora wykorzystane zostały wszystkie elementy wektora cech. Wartość promienia bufora dookoła punktu i długość boku sześcianu przyjęto na

1,5 m. Wartości te zostały przyjęte na podstawie gęstości chmury punktów. Zestawienie tych cech zostało przedstawione w Tabeli 4.

Tabela 4 Wartości wstępne elementów wektora cech.

Lp.	Parametr	Cecha	Wartość [m]
1	Promień bufora dookoła punktu	Linearity	1,5
		Sphericity	
		Planarity	
		Sum of eigenvalues	
		Omnivariance	
		Eigenentropy	
		Anisotropy	
		Surface variation	
		Verticality	
		1st eigenvalue	
		2nd eigenvalue	
		3rd eigenvalue	
		Surface density	
		deltaLocalHeight	
2	Długość boku sześcianu	imins	1,5
		imeans	
		imaxes	
		intensityRange	

Dodatkowo, wstępnymi parametrami algorytmu Random Forest było 100 drzew decyzyjnych, głębokość pojedynczego drzewa równa 2 000 oraz maksymalna liczba próbek równa 5 000. Wartości te zostały wybrane na podstawie doświadczeń własnych podczas uczenia się działania algorytmu.

W niniejszej pracy zostały przeprowadzone kolejno trzy etapy optymalizacji:

- Optymalizacja parametrów algorytmu Random Forest, dobór odpowiedniej liczby drzew decyzyjnych, głębokości pojedynczego drzewa oraz maksymalnej liczby próbek.
- Optymalizacja elementów wektora cech, dobór odpowiednich wartości bufora dookoła punkt oraz wartości długości boku sześcianu.
- Optymalizacja wektora cech, odrzucenie najmniej istotnych elementów wektora cech.

4.3.1 Optymalizacja parametrów algorytmu Random Forest

Poniższa Tabela 5 przedstawia jakie wartości parametrów zostały przetestowane, aby wybrać najbardziej optymalne pod względem dokładności oraz czasu uczenia i klasyfikacji.

Tabela 5 Testowane parametry algorytmu Random Forest.

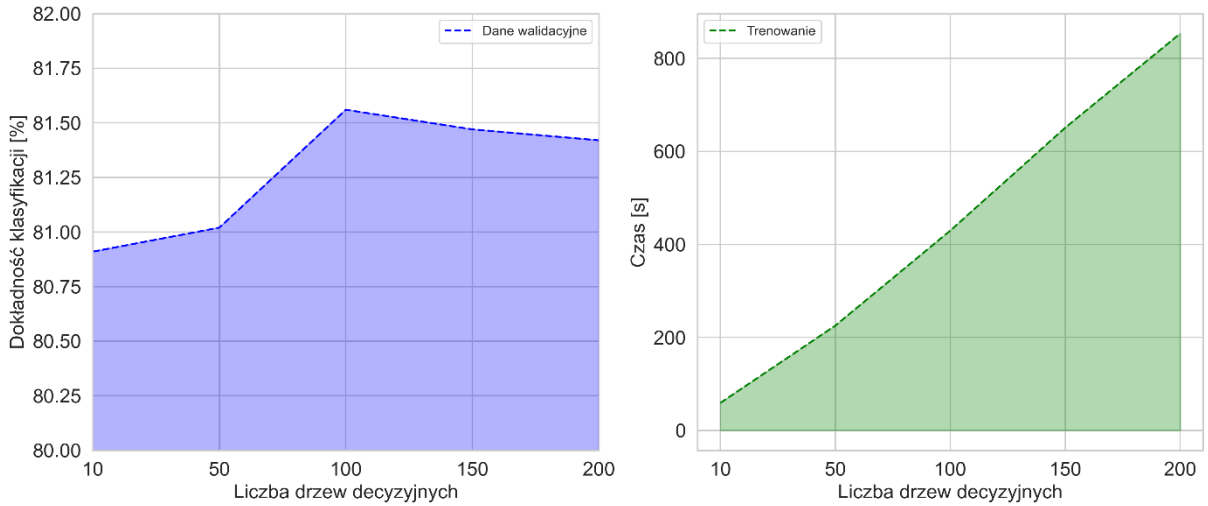
Lp.	Parametr	Testowane wartości
1	Liczba drzew decyzyjnych	10
		50
		100
		150
		200
2	Głębokość pojedynczego drzewa decyzyjnego	2 000
		5 000
		10 000
		50 000
		100 000
		Nielimitowana
3	Maksymalna liczba próbek każdej z klas dla danych uczących	1 000
		5 000
		10 000
		50 000
		100 000
		200 000

W dalszych krokach przedstawiono analizę poszczególnych parametrów. Podczas analizy ważnym aspektem była dokładność klasyfikacji oraz czas uczenia modelu.

Pierwszym testowanym parametrem była liczba drzew decyzyjnych. Określa ona ilość drzew jakie mają zostać wytrenowane podczas działania algorytmu. Analizując Tabela 6 oraz wykresy przedstawione na Rys. 23, można wywnioskować, że zwiększanie liczby drzew powoduje wzrost dokładności klasyfikatora oraz czasu jego trenowania. Należy się jednak skupić na tym, że najlepsze rezultaty osiągnął parametr o wartości 100. Dokładność dla wartości 150 i 200 nie uległa znacznemu polepszeniu, natomiast czas trenowania proporcjonalnie zwiększał się. Biorąc pod uwagę te zależności najlepszym wyborem pod względem dokładności i optymalnemu czasowi trenowania jest parametr o wartości 100.

Tabela 6 Dokładność oraz czas klasyfikacji i uczenia zależnie od ilości drzew decyzyjnych.

Lp.	Liczba drzew decyzyjnych	Głębokość pojedynczego drzewa decyzyjnego	Maksymalna liczba próbek każdej z klas dla danych uczących	Dokładność klasyfikacji dla danych walidacyjnych [%]	Czas uczenia modelu [s]	Czas klasyfikacji danych walidacyjnych [s]
1	10	2 000	5 000	80,91	59,28	3,36
2	50			81,02	225,48	9,13
3	100			81,54	429,82	16,38
4	150			81,53	651,10	23,23
5	200			81,57	853,51	31,44

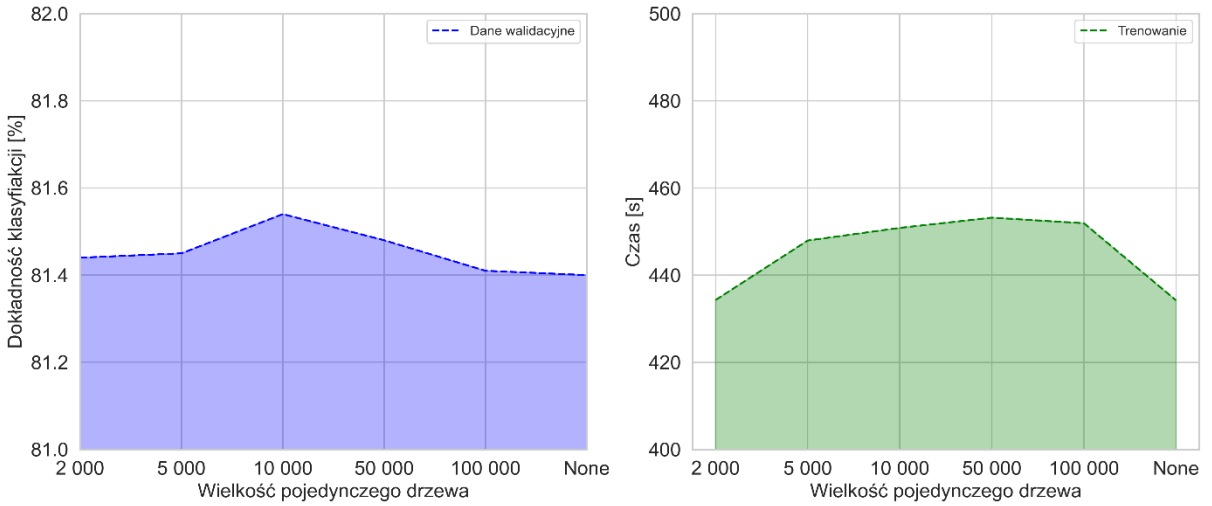


Rys. 23 Po lewej stronie: dokładność danych walidacyjnych zależnie od liczby drzew decyzyjnych; po prawej stronie: czas trenowania zależny od liczby drzew decyzyjnych.

Kolejnym testowanym parametrem jest głębokość pojedynczego drzewa decyzyjnego. Parametr ten odnosi się do maksymalnej liczby poziomów w drzewie. Ograniczając warstwy powoduje on zmianę dokładności oraz czasu trenowania. Tabela 7 oraz wykresy z Rys. 24 przedstawiają zależność dokładności oraz czasu trenowania od głębokości pojedynczego drzewa decyzyjnego. Dokładność klasyfikacji danych walidacyjnych klasuje się pomiędzy 81,40% a 81,54%. Ważny jest mało widoczny wzrost dokładności dla parametru 10 000 z wynikiem równym 81,54%. Biorąc pod uwagę również niewielkie różnice w czasie pomiędzy innymi parametrami, najlepszym parametrem określającym głębokość pojedynczego drzewa jest wartość 10 000.

Tabela 7 Dokładność oraz czas klasyfikacji i uczenia zależnie od głębokości drzewa decyzyjnego.

Lp.	Liczba drzew decyzyjnych	Głębokość pojedynczego drzewa decyzyjnego	Maksymalna liczba próbek każdej z klas dla danych uczących	Dokładność klasyfikacji dla danych walidacyjnych [%]	Czas uczenia modelu [s]	Czas klasyfikacji danych walidacyjnych [s]
1	100	2 000	5 000	81,44	434,28	15,74
2		5 000		81,45	447,93	15,69
3		10 000		81,54	450,84	15,67
4		50 000		81,48	453,18	15,89
5		100 000		81,41	451,91	15,23
6		Nielimitowana		81,40	434,21	15,52



Rys. 24 Po lewej stronie: dokładność danych walidacyjnych zależny od głębokości drzewa decyzyjnego; po prawej stronie: czas trenowania zależny od głębokości drzewa decyzyjnego.

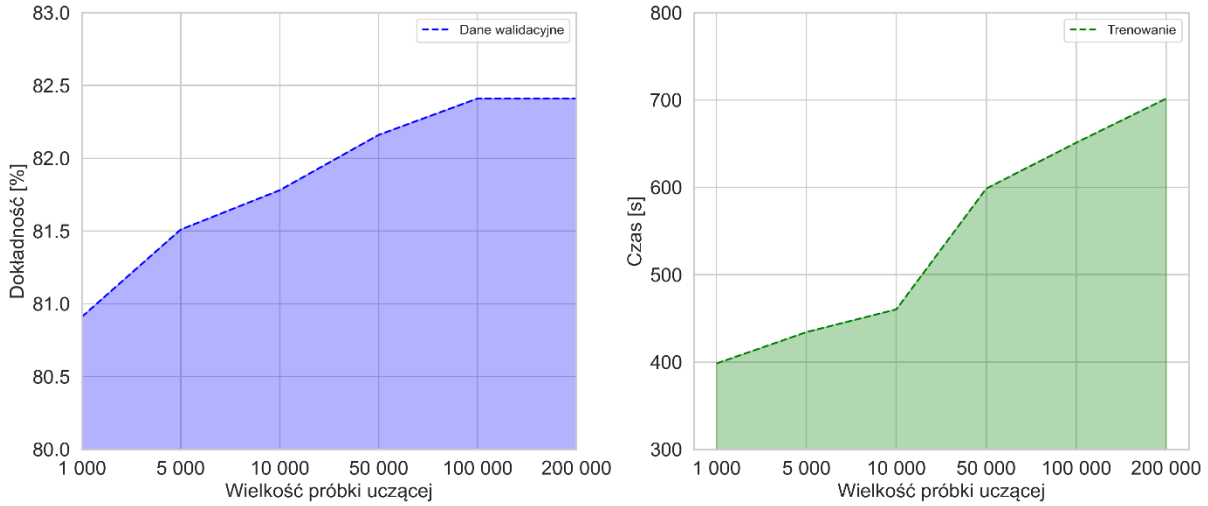
Ostatnim badanym parametrem jest maksymalna liczba próbek dla danych uczących. Parametr pozwala na wybór określonej liczby obserwacji, które zostaną uwzględnione podczas trenowania modelu klasyfikatora. Wybór odpowiedniej wartości tego parametru pozawala na zoptymalizowanie czasu treningu, zrównoważenie zbioru danych oraz poprawienie dokładności klasyfikacji.

Analizując dane zawarte w Tabeli 8 oraz informacje zawarte na wykresie przestawionym na Rys. 25 można zaobserwować, że zwiększanie parametru odpowiedzialnego za maksymalną liczbę próbek dla danych uczących powoduje zwiększanie zarówno dokładności jak i czasu trenowania. Optymalnym parametrem jest maksymalna liczba próbek równa 10 000, ponieważ osiąga zadowalającą dokładność równą 81,78% przy zachowaniu niskiego czasu trenowania. Dalsze parametry o wartościach powyżej 10 000 nie

są korzystne, ponieważ zwiększamy dokładność klasyfikacji dużym kosztem czasu trenowania.

Tabela 8 Dokładność oraz czas klasyfikacji i uczenia zależnie od maksymalnej liczby próbek każdej z klas dla danych uczących.

Lp.	Liczba drzew decyzyjnych	Głębokość pojedynczego drzewa decyzyjnego	Maksymalna liczba próbek każdej z klas dla danych uczących	Dokładność klasyfikacji dla danych walidacyjnych [%]	Czas uczenia modelu [s]	Czas klasyfikacji danych walidacyjnych [s]
1	100	2 000	1 000	80,91	398,30	15,03
2			5 000	81,51	434,19	18,81
3			10 000	81,78	460,27	16,03
4			50 000	82,16	598,58	19,24
5			100 000	82,41	651,09	21,03
6			200 000	82,41	701,56	22,84



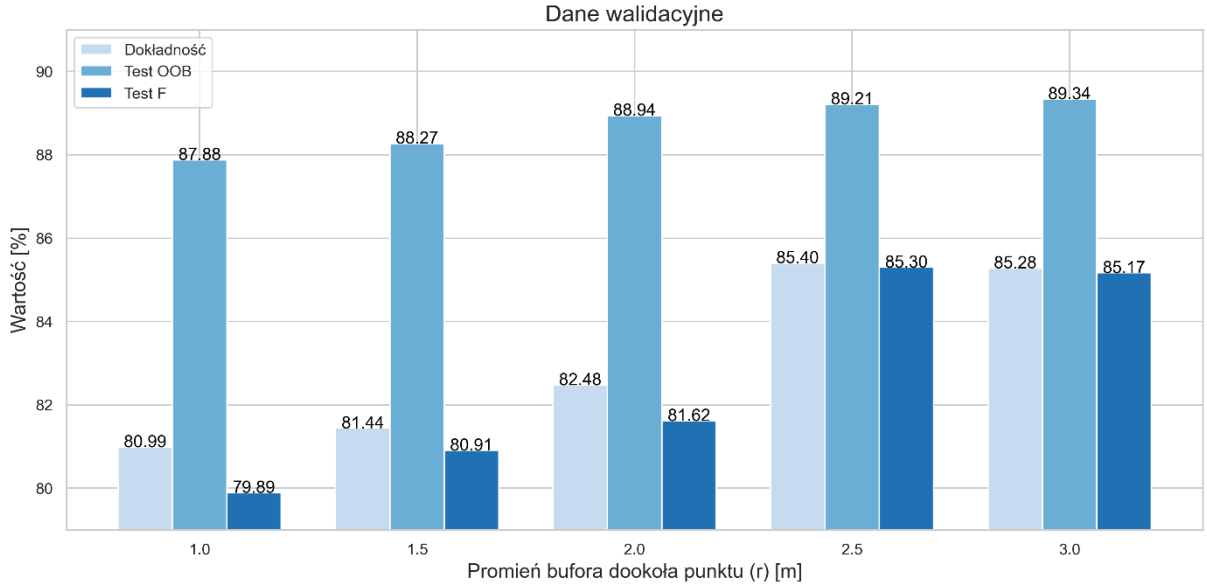
Rys. 25 Po lewej stronie: dokładność danych walidacyjnych zależnie od maksymalnej liczby próbek dla danych uczących; po prawej stronie: czas trenowania zależny od maksymalnej liczby próbek dla danych uczących.

4.3.2 Optymalizacja parametrów elementów wektora cech

Optymalizacja parametrów elementów wektora cech również jak optymalizacja parametrów algorytmu Random Forest, ma na celu poprawienie dokładności klasyfikacji. Testowany był promień bufora dookoła punktów i długość boku sześcianu. Wartości i cechy bazujące na tych parametrach zostały przedstawione w Tabeli 4. Na ich podstawie tworzony był nowy wektor cech w oparciu o różne promienie i boki sześcianów. Następnie

trenowany był klasyfikator dla każdego nowego wektora cech. Klasyfikator stworzony został w oparciu o wybór odpowiednich parametrów algorytmu z poprzedniego etapu optymalizacji.

Łącznie zostało przeprowadzonych 5 testów dla wartości sąsiedztwa od 1,0 m do 3,0 m o skoku co 0,5 m. Dla każdego z testów została obliczona dokładność klasyfikacji zbioru walidacyjnego, test OOB oraz miara F1. Analizując wykres na Rys. 26 można wywnioskować, że największą dokładność równą 85,40% uzyskano dla parametru 2,5 m. Wartości testu OOB oraz testu F1, również uzyskują zadowalające wartości. Parametr o wartości 3,0 m uzyskał podobne wartości testu OOB i F1, ale dokładność spadła o 0,12%. Podsumowując przeprowadzone testy, najlepszym parametrem okazała się być wartość 2,5 m.



Rys. 26 Dokładność, test OOB oraz test F1 dla danych walidacyjnych w zależności od wartości promienia bufora dookoła punktu oraz długości boku sześcianu.

W poniższej Tabeli 9 zostały przedstawione najlepsze, uzyskujące największe dokładności z optymalnym czasem trenowania parametry algorytmu Random Forest i parametry wektora cech. Dodatkowo, do testowanych cech zostały dodane pozostałe cechy nie bazujące na sąsiedztwie.

Tabela 9 Wartości parametrów algorytmu Random Forest, parametrów wektora cech oraz elementy wektora cech.

Wartości parametrów algorytmy Random Forest		
Lp.	**Parametr**	**Wartość**
1	Liczba drzew decyzyjnych	100
2	Głębokość pojedynczego drzewa decyzyjnego	10 000
3	Maksymalna liczba próbek każdej z klas dla danych uczących	10 000
Wartości parametrów wektora cech		
Lp.	**Parametr**	**Wartość [m]**
1	Promień bufora dookoła punktu	2,5
2	Długość boku sześcianu	2,5

Elementy wektora cech			
Lp.	**Nazwa**	**Lp.**	**Nazwa**
1	Intensity	**15**	Surface variation
2	ReturnNumber	**16**	Sphericity
3	NumberOfReturns	**17**	Verticality
4	normal x	**18**	1st eigenvalue
5	normal y	**19**	2nd eigenvalue
6	normal z	**20**	3rd eigenvalue
7	deltaGroundHeight	**21**	deltaLocalHeight
8	Surface density	**22**	imins
9	Sum of eigenvalues	**23**	imeans
10	Omnivariance	**24**	imaxes
11	Eigenentropy	**25**	intensityRange
12	Anisotropy	**26**	VDVI
13	Planarity	**27**	VVI
14	Linearity		

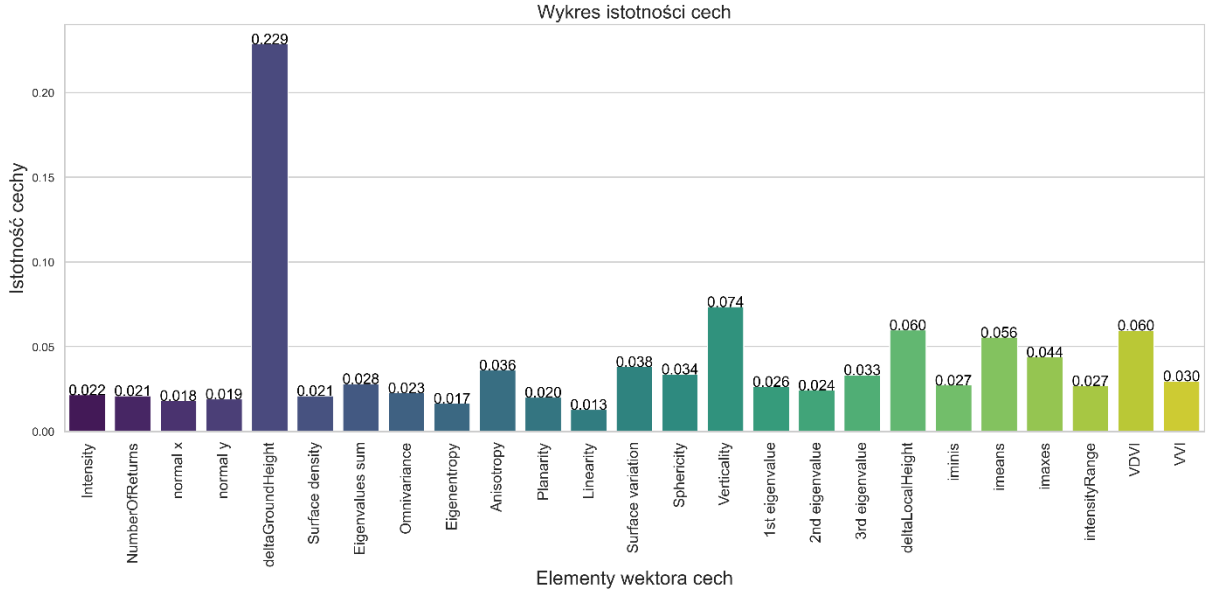
4.3.3 Optymalizacja wektora cech

Ostatnim etapem optymalizacji wzorca klasyfikatora jest odrzucenie najmniej istotnych elementów wektora cech. Wykonując trenowanie oparte na parametrach i cechach z Tabeli 9 otrzymujemy istotności poszczególnych cech. Wykres został przedstawiony wcześniej na Rys. 21. Na jego podstawie można stwierdzić, że najmniej istotnymi cechami są: ReturnNumber oraz normal z.

Cecha ReturnNumber była najmniej istotną, ponieważ prawie 95% punktów miała wartość cechy równą 1. Niska różnorodność cechy spowodowała, że była ona najmniej istotną cechą.

Cecha normal x, wykazała również znikomą istotność. Może być to spowodowane zależnością od innych cech tj.: deltaGroundHeight czy deltaLocalHeight.

Wykres istotności cech po odrzuceniu najmniej istotnych elementów został przedstawiony na Rys. 27. Dokładność klasyfikacji danych walidacyjnych nie zmieniła się i wynosi 96,92%. Czas trenowania również nie uległ zmianie i wynosi 459 sekund. Świadczy to o tym, że usunięte elementy wektora cech były nieistotne i zbędne podczas klasyfikacji.



Rys. 27 Wykres istotności cech po optymalizacji wektora cech.

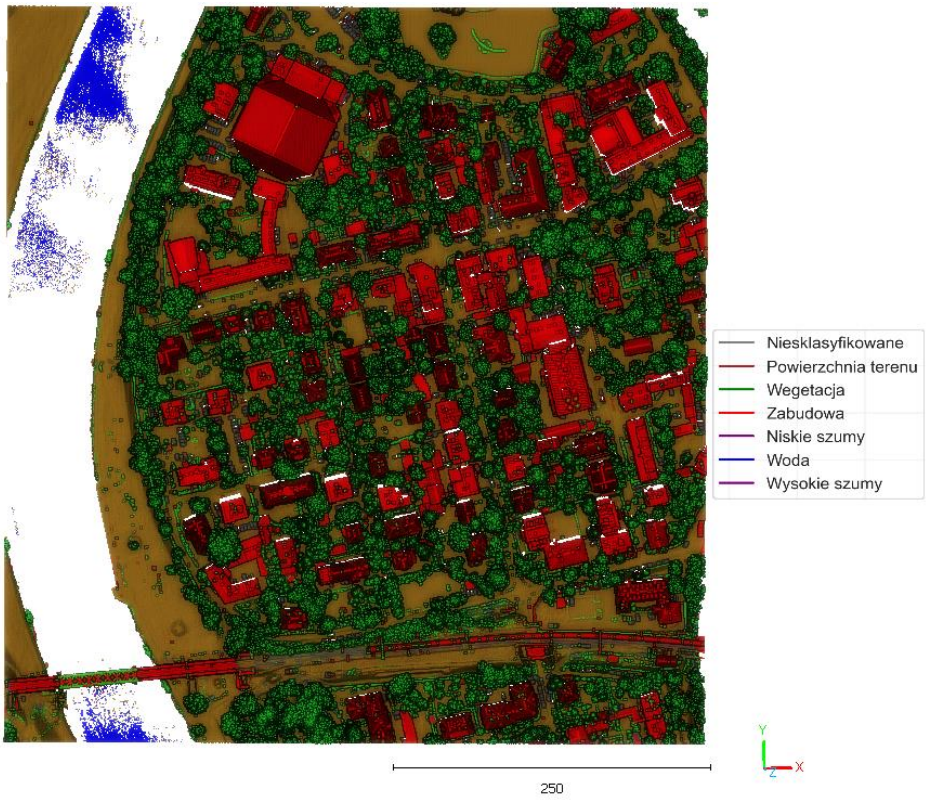
4.4 Analiza dokładności klasyfikacji

Model klasyfikatora został wytrenowany na danych uczących oraz walidacyjnych, dane testowe nie brały udziału podczas trenowania klasyfikatora.

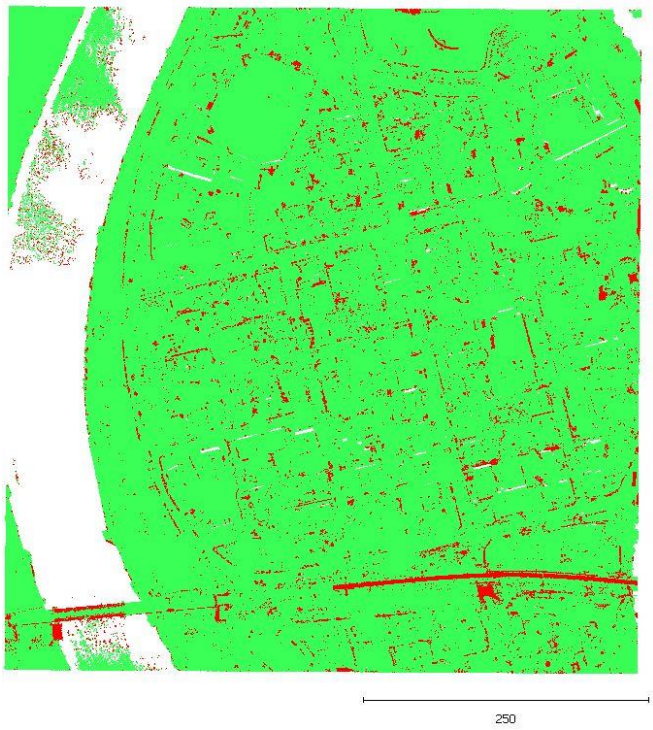
W celu sprawdzenia poprawności klasyfikacji danych testowych wytrenowanym modelem, zostało przeprowadzone porównanie z klasyfikacją manualną. Rys. 28 przedstawia dane testowe które zostały sklasyfikowane za pomocą klasyfikatora, natomiast Rys. 29 przedstawia lokalizację błędnie sklasyfikowanych punktów.

Punkty przedstawione za pomocą zielonego koloru oznaczają, że sklasyfikowano je prawidłowo, natomiast czerwony kolor oznacza lokalizację błędnie sklasyfikowanych punktów. W danych testowych można zauważyć, że wytrenowany model słabo poradził sobie

z pociągami oraz wysokimi elementami mostu. Reszta punktów wygląda na dobrze sklasyfikowane, nie ma grubych błędów a obiekty wyglądają poprawnie geometrycznie.



Rys. 28 Dane testowe, wynik klasyfikacji wytrenowanym modelem.



Rys. 29 Lokalizacja błędnie sklasyfikowanych punktów w danych testowych. Zielony kolor: poprawnie sklasyfikowane; czerwony kolor: błędnie sklasyfikowane.

Tabela 10 Macierz niezgodności danych testowych

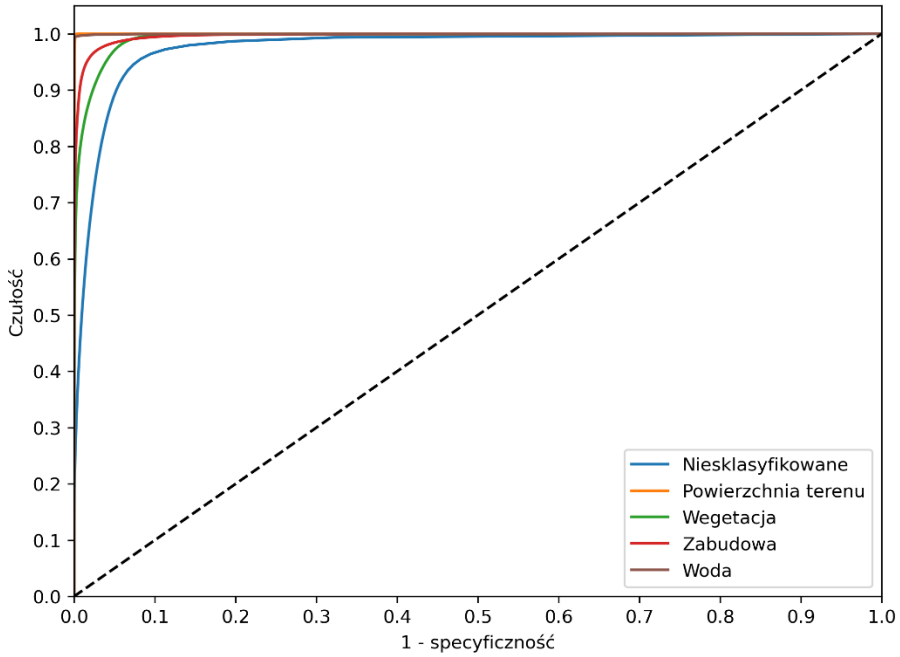
		Dane testowe - klasyfikacja manualna						
		1	**2**	**4**	**6**	**7**	**9**	**18**
Wynik klasyfikacji modelem	**1**	**82 902**	4 500	125 892	26 605	0	0	0
	2	0	**2 833 805**	216	5	0	7	0
	4	12 486	2 868	**1 030 615**	39364	0	0	0
	6	6 998	2 806	39 900	**975 955**	0	0	0
	7	0	1 411	39	23	**0**	0	0
	8	0	2 430	25	203	0	**24 428**	0
	18	0	0	13	3	0	0	**0**

Macierz niezgodności danych testowych została przedstawiona w Tabeli 10. Dokładność klasyfikacji jaka została uzyskana dla danych testowych wynosi 94,90%. Najwyższą dokładność klasyfikacji uzyskano dla punktów należących do powierzchni terenu, wegetacji oraz zabudowy. Klasa powierzchnia terenu została sklasyfikowana na poziomie 99,99%, klasa wegetacja na poziomie 94,93% oraz klasa zabudowa na poziomie 95,15%. Dobre wyniki osiągnęła również klasyfikacja wody, której poziom wynosił 89,65%. Punkty należące do klasy niesklasyfikowane nie osiągnęły zadowalającego poziomu klasyfikacji, ponieważ tylko 34,54% punktów sklasyfikowano poprawnie. Jest to oczekiwany rezultat, ponieważ różnorodność obiektów w klasie 1 nie pozwoliła poprawnie wytrenować modelu. Punkty w klasie 1 stanowią m.in.: samochody, linie kolejowe, pociągi oraz ogrodzenia. Najniższą dokładnością klasyfikacji wykazały się punkty należące do niskich szumów oraz punkty należące do wysokich szumów. Żaden z punktów tych klas nie został poprawnie sklasyfikowany. Mała liczność punktów klasy w danych testowych wpływa na przyznawanie im niskich wag.

Tabela 11 Analiza dokładności z wykorzystaniem współczynnika Kappa i pól pod krzywymi ROC.

Indeksy klas	Dane testowe								
	Icp				Kappa				ROC
	0,9490				0,9171				
	Ec	Eo	Ap	Au	F1	ACC	Precyzja	Czułość	AUC
1	0,1903	0,6544	0,3456	0,8097	0,4844	0,3454	0,8097	0,3456	0,9754
2	0,0049	0,0001	0,9999	0,9951	0,9975	0,9999	0,9951	0,9999	0,9998
4	0,1388	0,0504	0,9496	0,8612	0,9032	0,9493	0,8612	0,9496	0,9929
6	0,0635	0,0485	0,9515	0,9365	0,9439	0,9515	0,9365	0,9515	0,9961
7	0	1	0	1	0	0	0	0	0
9	0,0003	0,0981	0,9019	0,9997	0,9483	0,8925	0,9997	0,9019	0,9996
18	0	1	0	1	0	0	0	0	0

Tabela 11 przedstawia analizę dokładności z wykorzystaniem współczynnika Kappa. Dla danych testowych współczynnik ten wyniósł 0,9171. Wartości współczynnika powyżej 0,8 oznaczają, że klasyfikator jest dobrze zbudowany. Dodatkowo, przedstawione zostały inne parametry analizy dokładności. Błędy wykazują klasy niskich szumów oraz wysokich szumów. Widoczna jest słabość klasyfikatora w klasyfikacji punktów należących do klasy niesklasyfikowane.



Rys. 30 Krzywa ROC dla danych testowych.

Biorąc pod uwagę wykresy krzywych ROC zawarte na Rys. 30 oraz pól pod krzywymi, zawartymi w Tabeli 11, można stwierdzić, że potwierdzają one analizę

z wykorzystaniem współczynnika Kappa. Dla danych testowych klasy powierzchnia terenu, wegetacja, zabudowa i woda posiadają wartości pól pod krzywą powyżej 0.99. Tylko dla punktów należących do klasy niesklasyfikowane wartość ta wynosi 0,9754. Najgorszy wynik otrzymujemy w klasach wysokich szumów oraz niskich szumów, co potwierdza, że klasyfikator nie jest skuteczny dla tych klas.

5 Porównanie osiągnieć z wykorzystaniem Random Forest

(Wang i inni, 2019)

W tym artykule zaprezentowano zastosowanie algorytmu Random Forest w klasyfikacji danych LiDAR w kontekście miejskim. Autorzy zaproponowali wykorzystanie cech porównania pikseli w celu identyfikacji obiektów terenowych. Metoda ta opiera się wyłącznie na informacjach dotyczących wysokości z danych LiDAR, co przekłada się na szybkość obliczeń. Przeprowadzone eksperymenty wykazały, że zaproponowana metoda osiąga dokładność klasyfikacji na poziomie 87,2%, co można uznać za zadowalający wynik, szczególnie biorąc pod uwagę brak informacji spektralnych. Po uwzględnieniu informacji spektralnych, dokładność wzrosła do 90,3%, zbliżając się do wyników osiągniętych przez najnowocześniejsze metody w dziedzinie etykietowania semantycznego. Metoda cechuje się również szybkim obliczaniem cech oraz akceptowalnym czasem trenowania modelu. Autorzy przetestowali różne ustawienia parametrów i stwierdzili, że doprecyzowanie wielu z nich prowadzi do zwiększenia dokładności, ale osiąga się efekt nasycenia. Optymalne ustawienia sugerowane przez testy to około 100 cech , około 2 000 próbek treningowych na klasę i około 300 drzew dla podobnej sceny miejskiej. Wybór liczby cech dla każdego węzła ustawiony na pierwiastek liczby wszystkich cech przyczynia się do lepszej dokładności klasyfikacji. Wnioski z badań sugerują, że zaproponowana metoda oparta na Random Forest jest obiecująca i posiada duży potencjał w klasyfikacji danych LiDAR w środowisku miejskim.

(Bassier, Van Genechten i Vergauwen, 2019)

W artykule przedstawiono metodę automatycznej klasyfikacji danych chmur punktów obiektów strukturalnych w budynkach. Dane zostały podzielone na różne segmenty, które reprezentują odpowiednio piętra, sufity, dachy, belki, ściany oraz szumy. Zastosowano różne metody klasyfikacji, w tym Random Forest, K-nearest neighbors (KNN), Neutral Network, Support Vector Machines (SVM) oraz Boosted Trees. Warto zwrócić uwagę na metodę KNN,

która uzyskała 78,7 % dokładności klasyfikatora przy czasie trenowania 2,1 s. Spośród tych metod Random Forest osiągnął najwyższą dokładność, uzyskując 86% przy czasie trenowania równym 13,7 s. Wyniki pokazują, że klasyfikator Random Forest był w stanie niezawodnie sklasyfikować wcześniej podzielone fragmenty, uwzględniając obserwacje cech. Porównując go do innych metod klasyfikacji, Random Forest osiągnął wyższą skuteczność w klasyfikowaniu różnych klas obiektów budowlanych. Niemniej jednak, nadal występują pewne błędy i ograniczenia, Pewne klasy mają niższą wydajność ze względu na większą wariancję wartości cech wewnątrz klasy. Wyniki można poprawić poprzez podział klas na różne podklasy. Autorzy artykułu zauważyli, również, że większość testowanych metod klasyfikacji była w stanie poprawnie identyfikować klasy obiektów przy odpowiednich cechach i dużej liczbie znanych obserwacji.

(Chehata, Guo i Mallet, 2009)

W badanym artykule przeprowadzono badania mające na celu ocenę możliwości wykorzystania algorytmu Random Forest do klasyfikacji danych związanych z lotniczym skanowaniem laserowym. Dane zostały pomierzone z wykorzystaniem skanera typu full-waveform. Badany obszar obejmował tereny zurbanizowane, a klasy punktów obejmowały roślinność, budynki oraz powierzchnię terenu naturalną oraz sztuczną. Autorzy zaproponowali zastosowanie algorytmu do klasyfikacji pokrycia terenu, reprezentowanego jako raster o rozmiarze piksela 0,5 m, uwzględniając 21 cech obserwacji. Cechy zostały podzielone na pięć kategorii: bazujące na wysokości, bazujące na wpasowanej lokalnej płaszczyźnie, bazujące na odbiciach, bazujące na wartościach własnych oraz parametry wyznaczone z analizy kształtu fali. Na etapie trenowania klasyfikatora, autorzy odrzucili cztery najmniej istotne cechy. Były to cechy tj.: przekrój echa, różnica wysokości pomiędzy pierwszym i ostatnim odbiciem, ogólna liczba odbić i numer odbicia. Otrzymane wyniki wykazały wysoką dokładność klasyfikacji na poziomie 90%. Wynik ten potwierdza skuteczność klasyfikatora Random Forest w kontekście danych związanych z lotniczym skanowaniem laserowym.

6 Podsumowanie

Przedstawiona praca magisterska koncentruje się na wykorzystaniu algorytmu Random Forest do zautomatyzowania klasyfikacji chmur punktów pozyskanych z lotniczego skaningu laserowego. Głównym celem pracy było stworzenie modelu klasyfikatora, który pozwoliłby na precyzyjną klasyfikację danych, przyśpieszenie procesu klasyfikacji oraz wyeliminowania potrzeby doświadczonych operatorów.

Praca zawiera analizę teoretyczną, w której omówiono lotnicze skanowanie laserowe, algorytm płytkiego uczenia maszynowego i wektory cech. W pracy zawarta jest również praktyczna implementacja procesu klasyfikacji. Przeprowadzono ręczną klasyfikację danych uczących, walidacyjnych oraz testowych, a następnie zbudowano klasyfikator oparty na Random Forest. W dalszych etapach przeprowadzono optymalizację wzorca klasyfikatora. Optymalizacja parametrów algorytmu, optymalizacja parametrów wektora cech oraz redukcja wektora cech pozwoliła na uzyskanie modelu klasyfikatora, który zapewnia wysoką dokładność klasyfikacji chmur punktów w wysokości.

Uzyskany model klasyfikatora uzyskał dokładność dla danych walidacyjnych wynoszącą 96,92%, natomiast dla danych testowych uzyskano 94,70% dokładności. Poprawność modelu potwierdzają analizy dokładności z wykorzystaniem współczynnika Kappa i pól pod krzywymi ROC.

Praca ta wnosi wkład w rozwój techniki lotniczego skanowania laserowego poprzez automatyzację klasyfikacji danych i eliminację potrzeby doświadczonego operatora. Wykorzystując stworzony model można dokonać szybszej klasyfikacji podobnych chmur punktów.