

Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH



Instytut Radioelektroniki i Technik Multimedialnych

Praca dyplomowa magisterska

na kierunku Inżynieria Biomedyczna
w specjalności Informatyka Biomedyczna

Porównanie skuteczności algorytmów uczenia maszynowego na podstawie danych pochodzących z badań trakskryptomicznych

inż. Kacper Kubicki
293556

promotor
dr inż. Robert Kurjata

konsultacje
dr inż. Tymon Rubel

WARSZAWA 2024

Streszczenie

Badania przeprowadzone w ramach pracy dyplomowej miały na celu analizę porównawczą metod selekcji najbardziej informacyjnych genów na podstawie danych mikromacierzowych. Badania zostały zrealizowane w oparciu o trzy ogólnodostępne wieloklasowe zbiory danych posiadające od 12000 do ponad 49000 genów, charakteryzujące się dużym niezrównoważeniem pomiędzy występującymi klasami. Analiza selekcji cech obejmowała metody filtrujące, wbudowane oraz metody hybrydowe wykorzystujące algorytmy metod opakowujących. Dla metod filtrujących oraz wbudowanych zbadany został wpływ ilości wyselekcjonowanych genów na wydajność modelu. W przypadku metod hybrydowych sprawdzono wpływ algorytmów opakowujących na wydajność modeli wykorzystując wcześniej wyselekcjonowane za pomocą metod filtrujących i wbudowanych podzbiory. Na podstawie uzyskanych rezultatów stwierdzono, że, metody filtrujące wymagają większej ilości genów niż metody wbudowane, w celu osiągnięcia równie wysokiej wydajności modeli. Wykazano, że zastosowanie algorytmów opakowujących w metodach hybrydowych, bazujących na wcześniej wyselekcjonowanych podzbiorach przy użyciu metod filtrujących, pozwala zwiększyć wydajność modelu, jednocześnie ograniczając wielkość podzbioru. Natomiast metody hybrydowe oparte na podzbiorach wyselekcjonowanych przy użyciu metod wbudowanych powodują znaczną redukcję ilości wyselekcjonowanych genów utrzymując wydajność modeli na tym samym poziomie co pojedyncze metody wbudowane. Dodatkowo w celu określenia wartości optymalnych poddane analizie zostały zarówno parametry metod wbudowanych, jak i algorytmów opakowujących wykorzystanych w metodach hybrydowych. W oparciu o osiągnięte wyniki potwierdzono, że możliwe jest uzyskanie dla konkretnego zbioru optymalnych wartości parametrów, jednak proces ten wymaga sprawdzenia bardzo dużej ilości kombinacji rozwiązań.

Słowa kluczowe: selekcja cech, mikromacierze, ekspresja genów, klasyfikacja wieloklasowa

Abstract

This thesis investigates the comparative analysis of feature selection methods of the most informative genes based on microarray data. The research was carried out using three publicly available multiclass datasets containing from 12,000 to over 49,000 genes, characterized by significant imbalance between classes. Feature selection analysis included filter methods, embedded methods, and hybrid methods using wrapper algorithms. For filtering and embedded methods, the impact of the number of selected genes on model performance was examined. In the case of hybrid methods, the influence of wrapper algorithms on model performance was investigated using subsets previously selected by filtering and embedded methods. Based on the results obtained, it was found that filtering approaches require larger number of genes than embedded approaches to achieve equally high model performance. It was demonstrated that the use of wrapper algorithms in hybrid methods, based on previously selected subsets using filtering methods, allows for an increase in model performance while limiting the subset size. On the other hand, hybrid methods based on subsets selected using embedded methods result in a significant reduction in the number of selected genes while maintaining model performance at the same level as individual embedded methods. Additionally, to determine optimal values, both the parameters of embedded methods and wrapper algorithms used in hybrid methods were analyzed. Based on the achieved results, it was confirmed that obtaining optimal parameter values for a specific dataset is possible, but this process requires checking a very large number of solution combinations.

Key words: feature selection, microarray, gene expression, multiclassification

Spis treści

1	Wprowadzenie	3
2	Wstęp teoretyczny	5
2.1	Oligonukleotydowe mikromacierze ekspresyjne	5
2.2	Metody selekcji cech.....	9
2.2.1	Metody filtrujące	9
2.2.2	Metody opakowujące	13
2.2.3	Metody wbudowane	20
2.2.4	Metody hybrydowe.....	24
2.3	Metody klasyfikacji	25
2.4	Ocena wydajności modelu klasyfikatora.....	28
2.4.1	Makro uśrednianie metryk klasyfikatora	32
2.4.2	Mikro uśrednianie metryk klasyfikatora	33
2.4.3	Ważone uśrednianie metryk klasyfikatora	34
2.5	Podsumowanie przeglądu literaturowego	34
3	Zbiory danych.....	35
4	Metodyka badań	37
4.1	Przetwarzanie wstępne danych	38
4.1.1	Wczytanie danych	38
4.1.2	Eliminacja brakujących wartości	39
4.1.3	Podział zbiorów na grupy treningowe i walidacyjne.....	40
4.1.4	Normalizacja min-max zbiorów.....	41
4.1.5	Wstępna selekcja cech różnicujących.....	41
4.2	Metody selekcji cech.....	42
4.2.1	Metody filtrujące	42
4.2.2	Metody wbudowane	42
4.2.3	Metody hybrydowe.....	43
4.3	Metody klasyfikacji	44
4.4	Ocena wydajności modeli	44
4.5	Środowisko.....	45
5	Wyniki badań i ich interpretacja	46
5.1	Metody filtrujące	46

5.2	Metody wbudowane.....	49
5.2.1	Regresja grzbietowa	50
5.2.2	Las losowy	53
5.2.3	SVM-RFE.....	56
5.3	Metody hybrydowe	60
5.3.1	Metody filtrujące + metody opakowujące	60
5.3.2	Metody wbudowane + metody opakowujące	70
6	Wnioski i plan dalszych badań	81
7	Bibliografia.....	83
8	Spis rysunków	86
9	Spis tabel.....	89
10	Spis załączników	90
11	Załączniki	91
11.1	Załącznik nr 1	91
11.2	Załącznik nr 2	97

1 Wprowadzenie

Według Światowej Organizacji Zdrowia [1] nowotwory są drugą najczęstszą przyczyną zgonów na świecie stanowiąc ponad 10 milionów zgonów rocznie. Oznacza to, że w skali roku co 6 osoba na świecie umiera z powodu nowotworu. Szacuje się, że przy zachowaniu aktualnego współczynnika uleczalności chorób nowotworowych wynoszącego do 50%, w ciągu dwóch najbliższych dekad liczba zgonów spowodowanych nowotworami wzrośnie do 15 milionów przypadków na rok [2].

Wykrycie zmian nowotworowych na wczesnym stadium zaawansowania może mieć kluczowe znacznie dla zwiększenia wskaźników przeżywalności choroby. Wcześniej zdiagnozowany nowotwór otwiera większy zakres możliwości terapeutycznych, w tym stosowanie bardziej ukierunkowanych i skutecznych metod. Dokładne określenie rodzaju nowotworu oraz poznanie mechanizmów z nim związanych są kluczowe przy doborze odpowiedniego leczenia.

Nowotwory są skomplikowanymi chorobami genetycznymi, wynikającymi z niekontrolowanego wzrostu i podziału komórek. Mechanizm powstawania nowotworów jest złożony i związany z mutacjami, które wpływają na aktywność genów i regulację komórek. Znaczący wpływ mają mutacje genów kodujących białka, które uczestniczą w cyklu komórkowym - antyoksydantów i protoonkogenów [3]. W komórkach normalnych antyoksydanty działają hamując procesy rozmnażania, utrzymując stabilność genetyczną, natomiast protoonkogeny pełnią funkcje niezbędne do zachowania integralności tkanek regulując wzrost komórek i hamując ich obumieranie. Gdy dochodzi do mutacji genów regulujących wzrost i podziały komórkowe, może dojść do zaburzenia równowagi pomiędzy proliferacją (podziałem), a apoptozą (obumieraniem) komórek [4]. Są to procesy nie reagujące na naturalne mechanizmy regulacyjne organizmu. Zahamowana aktywność antyoksydantów, przy jednocześnie nadmiernej aktywności protoonkogenów może prowadzić do niekontrolowanego wzrostu komórek w wyniku czego praktycznie każda tkanka organizmu może zostać zaatakowana przez komórki nowotworowe. Wobec tego aktywność genów jest ściśle związana z występowaniem nowotworów, co ma istotne znaczenie w stosowaniu odpowiednich terapii genowych w walce z tą chorobą.

Detekcja korelacji pomiędzy aktywnością genów, a występującymi w organizmach schorzeniami pozwala na opracowanie lepszych strategii diagnozowania i leczenia celowanego choroby. Poznanie dokładnego rodzaju nowotworu umożliwia personalizację terapii dostosowując się do unikalnych cech każdego pacjenta. Dla różnych rodzajów choroby dostępne są specyficzne leki, które celują w konkretne mechanizmy lub geny obecne w komórkach nowotworowych. Celem jest wybranie terapii, która maksymalizuje szanse wyleczenia lub długotrwałą remisję choroby, poprawiając jakość życia pacjenta przy jednoczesnej minimalizacji skutków ubocznych.

Istnieje wiele technik pozwalających na wykrywanie zależności pomiędzy aktywnością genów, a występującymi chorobami, jednak większość z nich posiada znaczne ograniczenia w postaci wysokich

kosztów, trudności w analizie równoczesnej ekspresji wielu genów oraz rozmiarów dochodzących nawet do terabajtów danych. Metodą minimalizującą te ograniczenia są mikromacierze ekspresyjne pozwalające na względnie prostą analizę równoczesnej ekspresji wielu tysięcy genów, będąc jednocześnie bardziej dostępnymi i bardziej dostosowanymi do rutynowych badań ekspresji genów podczas terapii. W związku z tym, mikromacierze ekspresyjne często stanowią preferowaną opcję w badaniach nad ekspresją genów.

Głównym problemem związanym z analizą danych mikromacierzowych jest ich wielowymiarowość przy jednocześnie bardzo małej liczbowości próbek, co w znaczny sposób komplikuje analizę. W związku z tym konieczne jest przeprowadzenie selekcji najbardziej informacyjnych genów w celu identyfikacji podzbioru, który jest najbardziej istotny dla danej choroby, prowadząc do osiągnięcia wydajniejszego i dokładniejszego procesu klasyfikacyjnego.

Przedmiotem badań niniejszej pracy były ogólnodostępne wieloklasowe zbiory mikromacierzowe posiadające od kilkunastu do kilkudziesięciu tysięcy genów. Zbiory charakteryzowały się dużym niezrównoważeniem pomiędzy występującymi klasami, gdzie przynajmniej jedna klasa miała więcej próbek niż pozostałe.

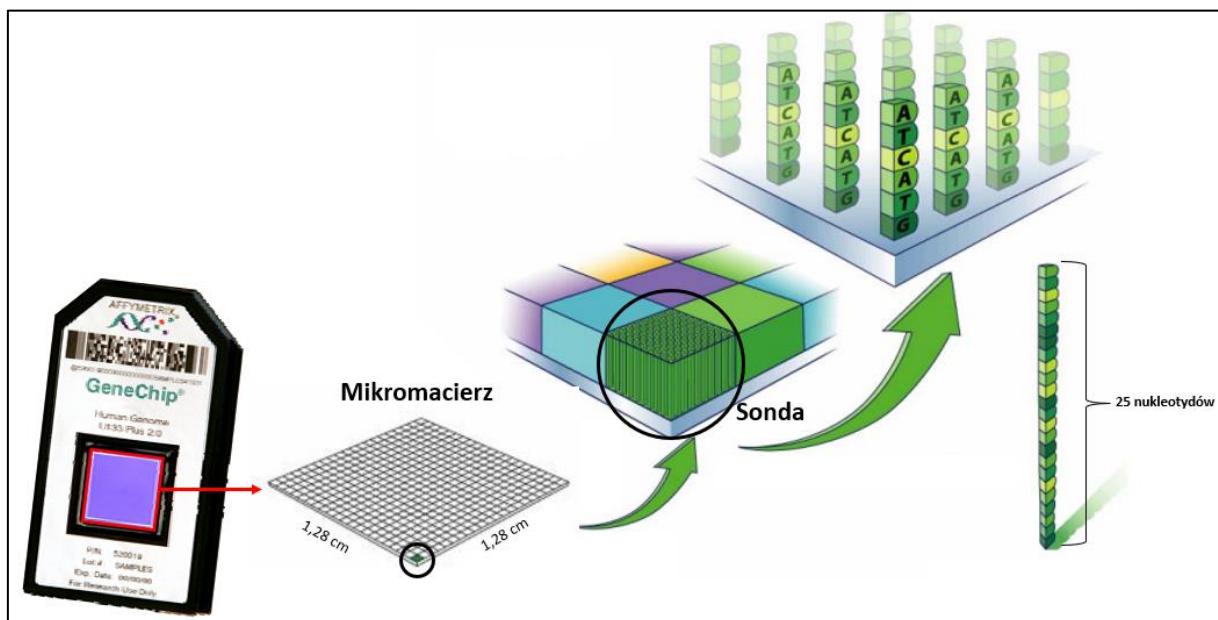
Biorąc pod uwagę złożoność problemu selekcji najbardziej informacyjnych cech dla wieloklasowych zbiorów danych oraz fakt istnienia niewielkiej ilości źródeł, w których autorzy skupiają swoją uwagę na porównaniu metod selekcji dla tak złożonych zbiorów, w niniejszej pracy podjęta została próba analizy porównawczej trzech typów metod selekcji najbardziej informacyjnych genów na podstawie wieloklasowych zbiorów mikromacierzowych. Analiza obejmowała metody filtrujące, wbudowane oraz metody hybrydowe wykorzystujące algorytmy metod opakowujących oraz podzbiory wyselekcjonowane uprzednio za pomocą metod filtrujących i wbudowanych. W przypadku metod filtrujących oraz wbudowanych zbadany został wpływ ilości wyselekcjonowanych genów na wydajność. Dodatkowo dla metod wbudowanych poddany analizie został wpływ optymalnych parametrów zastosowanych algorytmów na wydajność badanych modeli. Ponadto dla modeli hybrydowych sprawdzono wpływ metod opakowujących na wydajność modeli wykorzystując wcześniej wyselekcjonowane podzbiory metod filtrujących i wbudowanych. Dodatkowo poddane analizie zostały parametry algorytmów opakowujących wykorzystanych w metodach hybrydowych selekcji genów.

2 Wstęp teoretyczny

Część teoretyczna niniejszej pracy miała na celu zebranie dotychczasowego stanu wiedzy na temat selekcji najbardziej informacyjnych genów na podstawie mikromacierzowych zbiorów danych. Wobec tego przegląd rozpoczęto od przedstawienia mikromacierzy ekspresyjnych, procesu ich wytwarzania, zasady działania oraz uzyskiwanych na ich podstawie danych. Następnie szczegółowo zestawione zostały różne typy metod selekcji najbardziej informacyjnych cech dla takich zbiorów. Na koniec krótko przedstawiono najczęściej wykorzystywane klasyfikatory oraz przedstawione zostały różne metryki oceny wydajności badanych modeli wykorzystując różne metody uśredniania tych wartości dla problemów wieloklasowych.

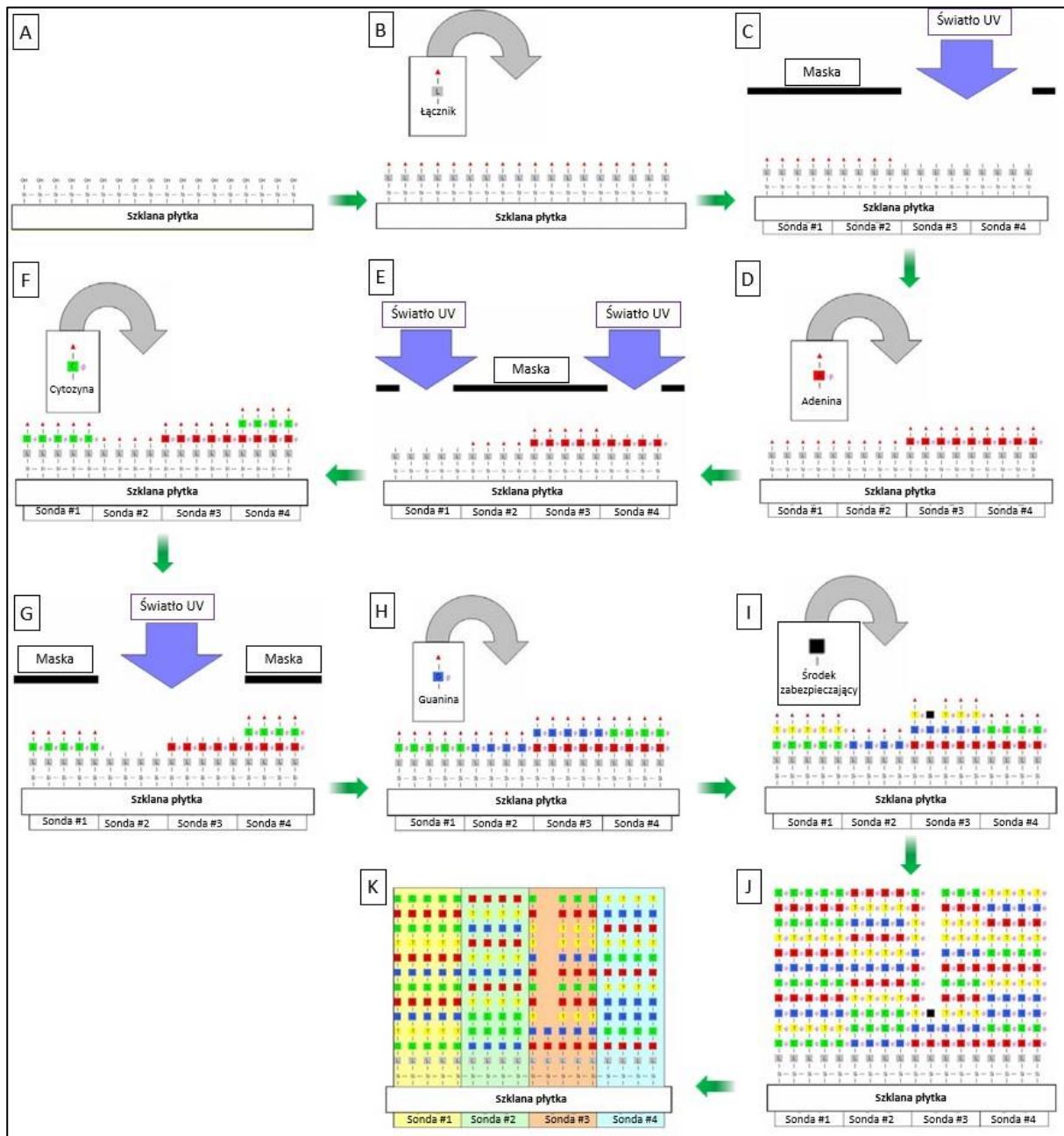
2.1 Oligonukleotydowe mikromacierze ekspresyjne

Oligonukleotydowe mikromacierze ekspresyjne są zaawansowanym narzędziem biotechnologicznym wykorzystywanym do analizy ekspresji genów. Umożliwiają one jednoczesne badanie aktywności wielu tysięcy genów w jednej próbce biologicznej, co pozwala na identyfikację związków między ekspresją genów, a występowaniem chorób w organizmie i zrozumieniem ich molekularnych mechanizmów. Występują one w postaci płyt o wymiarach 1,28 x 1,28 cm, na powierzchni których w uporządkowany sposób rozmieszczone są sondy DNA stanowiące matrycę dla badanych genów. Sondы mają kształt kwadratów o długości boku 5-24 μm w zależności od modelu. Liczba sond na płytce może przekraczać 6,5 miliona, a w każdej z sond znajduje się kilka milionów łańcuchów oligonukleotydowych o długości 25 nukleotydów. Całość zamknięta jest w kasetce ochronnej [7]. Schemat budowy mikromacierzy przedstawiony został na rysunku 1.



Rysunek 1. Schemat budowy mikromacierzy [7]

Proces produkcji chipów przypomina proces stosowany w przemyśle półprzewodnikowym, w którym wykorzystywany jest proces fotolitografii do kontrolowania produkcji wielu warstw materiału. Wytwarzanie rozpoczyna się od szklanej płytki, która kąpana jest w roztworze krzemu w celutworzenia stabilnej warstwy powierzchniowej (rys. 2A). Dochodzi wówczas do połączenia cząsteczek krzemu ze szkłem, zapewniając punkty inicjacji sond DNA. Im bliżej siebie są cząsteczki krzemu, tym gęściej mogą być one umieszczone.

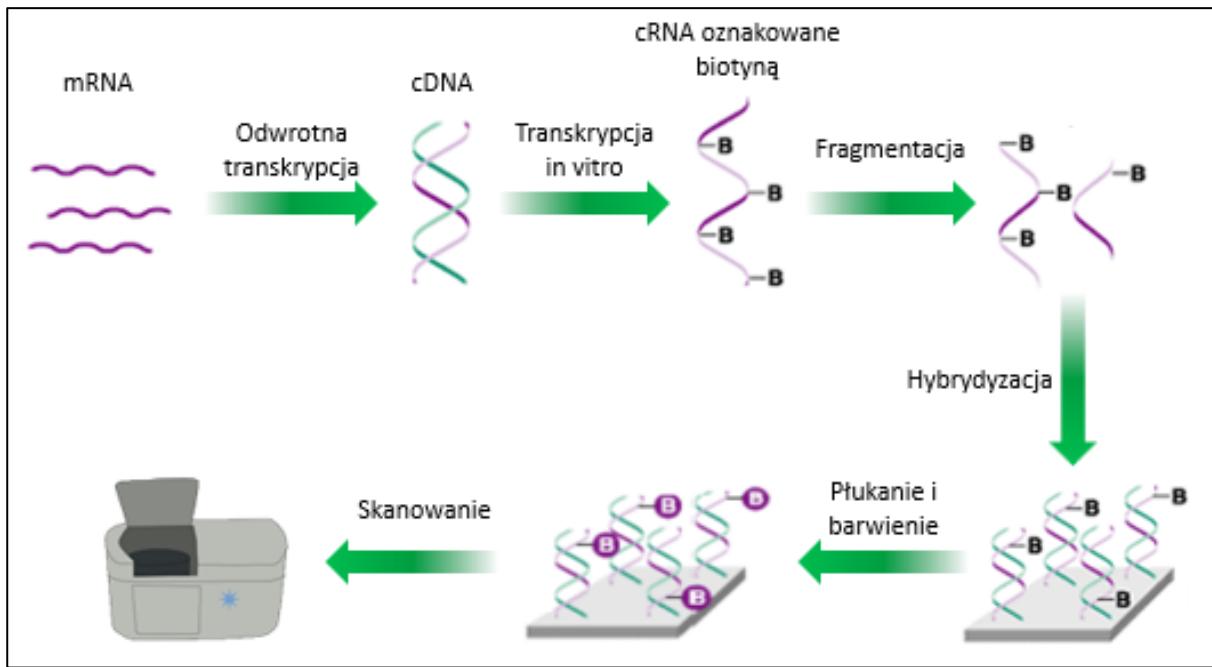


Rysunek 2. Proces wytwarzania mikromacierzy [8]

Następnie do warstwy krzemu przyłączane kowalencyjnie są cząsteczki łącznika z cząsteczkami światłoczułymi, które służą jako mechanizm ochronny przed nieuchcianymi elementami w danej sondzie DNA (rys. 2B). W celu dołączenia do sondy nukleotydów należy wystawić ją na działanie światła

ultrafioletowego, które spowoduje wytrącenie elementu ochronnego z końca wybranej sondy. Podczas procesu aktywacji świetlnej stosowane są maski, które chronią wybrane obszary od ekspozycji świetlnej (rys. 2C). Pozwala to na kontrolowanie budowy oczekiwanej sekwencji nukleotydów na zestawach sond. Kolejnym krokiem jest przemycie płytki roztworem zawierającym pojedyncze, wolne związki wybranego nukleotydu, które połączone są z tą samą światłoczułą cząsteczką oraz boczną grupą ochronną (rys. 2D). Nukleotydy łączą się z niezabezpieczonymi sondami inicjując łańcuch DNA. Etap ten stanowi punkt wyjścia do dodania kolejnego nukleotydu. Nukleotydy są dodawane jeden po drugim w kolejnych iteracjach naświetlania i płukania roztworem, aż do uzyskania określonej liczby nukleotydów na każdej sondzie. Każda z sond zawiera DNA o innej sekwencji. Czasami odblokowany nukleotyd nie wiąże się z następną odpowiednią zasadą. Aby temu zapobiec, przed świetlnym odbezpieczeniem nukleotydu dodaje się czynnik zapobiegający przed dalszym wzrostem nieprawidłowo zbudowanej sondy (rys. 2I), wówczas łańcuch zostaje poświęcony i niekompletny. Niemniej jednak, w każdym elemencie znajduje się mnóstwo identycznych sond, co nie stanowi problemu [8, 101]. Po zakończeniu ostatniego etapu usuwane są boczne grupy oraz środki zabezpieczające (rys. 2K). Ostatecznie każda z sond zawiera DNA o innej sekwencji łańcucha, a zbiór wszystkich sond tworzy mikromacierz.

Zasadą działania mikromacierzy jest komplementarność kwasów nukleinowych. Badanie ekspresji genów obejmuje kilka etapów. Po pierwsze z badanej populacji komórek należy dokonać ekstrakcji mRNA. Następnie wykorzystując proces odwrotnej transkrypcji mRNA materiał jest przekształcany do stabilnej formy komplementarnego DNA (cDNA). Otrzymany cDNA poddany zostaje transkrypcji *in vitro*, polegającej na syntezie cRNA. Kolejnym krokiem jest znakowanie cRNA biotyną, który następnie poddawany jest fragmentacji na krótkie odcinki. Przygotowany w ten sposób materiał zostaje naniesiony na powierzchnię mikromacierzy, gdzie poszczególne rodzaje cząstek mogą ulec hybrydyzacji z sondami zawierającymi komplementarne wobec nich nici. Proces hybrydyzacji materiału odbywa się najczęściej z wykorzystaniem odpowiedniego wyposażenia, gdzie w temperaturze 45°C płytki mikromacierzowe obracają się przez 16 godzin. Jeżeli określony fragment materiału występuje w badanych komórkach, to jego cząsteczki połączą się z odpowiednią sondą. Natomiast jeżeli dany fragment RNA nie występuje w badanych próbkach, to reprezentująca go sonda pozostanie niehybrydyzowana, czyli pozostałe pusta. Niehybrydyzowany materiał zostaje wypłukany z powierzchni płytki. Zhybrydyzowane cząstki barwione są cząsteczkami fluorescencyjnymi, które przylegają do wprowadzonej wcześniej biotyny. Następnie mikromacierz wprowadzona zostaje do skanera, z której obraz fluorescencji sczytywany jest ilościowo za pomocą wiązki lasera [7, 9]. Ogólny schemat przebiegu pomiarów mikromacierzowych przedstawiony został na rysunku 3.



Rysunek 3. Schemat przebiegu pomiarów macierzowych [8]

Bezpośrednim wynikiem pomiaru ze skanera jest 16-bitowy obraz w formacie DAT, w którym wartości fluorescencji przechowywane są jako wartości pikseli. Taki obraz poddaje się przetwarzaniu niskiego poziomu. Jest to zabieg polegający na określeniu wartości liczbowych miar aktywności genów. Intensywność sygnału dla poszczególnych genów jest proporcjonalna do ilości zhybrydowanych fragmentów DNA o danej sekwencji w sondzie. Im większa jest intensywność, tym większa jest miara aktywności danego genu [9]. Przetworzone dane mikromacierzowe mają strukturę $M \times N$, w której każda kolumna reprezentuje inny gen, a każdy wiersz reprezentuje inną macierz. Przykład danych mikromacierzowych przedstawiony został na rysunku 4.

		Identyfikatory zestawów sond (geny)																					
		1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_l_at	1431_at	...	90265_at	90610_at	91617_at	91682_at	91684_g_at	91703_at	91816_f_at	91826_at	91920_at	91952_at	
Próbki	Samples	0	12.198905	8.143128	9.587965	10.658479	7.862327	8.482606	7.937227	5.053111	4.940167	7.811214	...	12.219138	8.824641	9.193525	9.500244	8.894818	7.753351	9.968667	6.163901	10.574310	8.316960
	1	12.209149	8.442529	8.169424	10.576673	7.678072	8.394463	7.540709	4.517276	5.835419	7.062856	...	11.772026	8.781032	8.824004	9.236253	8.439208	8.074677	9.990384	5.044394	10.766529	8.250298	
2	11.483765	8.471675	7.636625	10.808803	6.898450	8.360628	8.448323	4.906891	4.153805	8.077350	...	11.769094	7.643856	8.533719	9.325081	8.711839	8.074141	9.542645	5.672425	10.381002	8.330917		
3	12.385377	8.324477	7.514122	10.876517	7.493455	8.251246	8.346070	5.572890	4.940167	7.542258	...	12.166703	9.142362	8.969243	9.179163	8.176921	7.663696	9.906440	7.367196	10.618844	8.242698		
4	12.225629	8.200653	7.674545	9.982994	7.291861	7.570615	7.044394	4.232661	3.560715	7.7771489	...	11.943650	8.627169	8.780048	9.030667	8.672425	7.909293	9.961160	5.605850	10.259743	7.847997		
...		
175	13.907041	9.347178	8.055282	10.923584	6.711495	9.203348	8.330917	5.593951	4.694880	6.601399	...	9.826548	9.341630	8.662490	9.460456	9.046306	7.859845	10.678336	4.812498	13.252577	9.015136		
176	13.308950	8.706323	8.296916	10.706496	5.325530	7.872521	7.862947	5.177918	6.153805	8.696705	...	9.881267	9.461684	9.143383	9.327777	9.522385	8.085871	10.403119	8.054197	11.740540	7.749534		
177	12.988220	8.410239	7.909893	10.842350	7.285402	8.848993	8.359310	6.580447	4.478972	7.816344	...	11.682468	9.212132	8.647099	9.262800	8.736740	7.852996	10.306062	7.906289	11.973805	8.346514		
178	13.803223	8.615078	8.699052	10.961811	6.721099	9.338736	7.864805	5.602884	4.672425	7.068341	...	10.892346	8.766680	8.923922	9.754721	9.701133	7.557655	10.540031	6.112700	13.558588	9.051481		
179	13.276633	8.712183	7.983564	10.514122	5.963474	8.358431	8.857048	4.240314	6.348728	6.722466	...	10.899508	7.955940	9.076281	9.251246	9.645478	7.358431	10.609456	5.937815	12.128735	8.381975		

180 rows x 49151 columns

Wartości zestawów sond w kolejnych próbkach

Rysunek 4. Przykładowe dane mikromacierzowe

Główny problem związany z mikromacierzami sprowadza się do wielowymiarowości danych przy jednocześnie bardzo małej liczebności próbek, co w znaczny sposób komplikuje analizę danych. W związku z tym konieczne jest przeprowadzenie selekcji genów, w celu niwelacji redundancji i wybrania najbardziej informacyjnych cech. Proces ten umożliwia lepszą i dokładniejszą klasyfikację występującej w organizmie choroby.

2.2 Metody selekcji cech

Głównym celem selekcji cech jest wybór najbardziej informacyjnych i znaczących genów, których wyselekcjonowanie w znaczący sposób wpłynie na zwiększenie dokładności modeli klasyfikacyjnych. Selekcję tę można osiągnąć poprzez usunięcie nieistotnych cech. Zastosowanie selekcji cech posiada wiele zalet, z których najważniejsze to [10]:

- redukcja wymiarowości danych, prowadząca do wydajniejszego i dokładniejszego procesu klasyfikacyjnego,
- lepsza interpretowalność biologiczna, pozwalająca na identyfikację podzbioru genów, które są najbardziej istotne dla danej choroby,
- zmniejszenie wymagań dotyczących pamięci masowej, umożliwiające obniżenie czasochłonności i kosztowności procesu klasyfikacyjnego.

Zalety technik selekcji cech mają swoją cenę, ponieważ poszukiwanie podzbioru optymalnych cech wprowadza dodatkową warstwę złożoności modelu. Zamiast optymalizować parametry modelu klasyfikacyjnego, należy znaleźć optymalne parametry modelu selekcyjnego w celu znalezienia optymalnego podzbioru cech.

Techniki selekcji cech różnią się między sobą sposobem wyszukiwania optymalnych podzbiorów. W kontekście problemu klasyfikacyjnego techniki selekcji cech można podzielić na cztery podstawowe kategorie. Wyróżnia się wśród nich: metody filtrujące, metody opakowujące, metody wbudowane oraz metody hybrydowe [11].

2.2.1 Metody filtrujące

Metody filtrujące selekcji cech używają rankingu jako metryki do oceny wyboru optymalnego podzbioru. Cechy są uszeregowane na podstawie wyników w różnych testach statystycznych pod kątem ich korelacji z przynależnością do danej klasy. Wybierane są te, które uzyskują wynik powyżej określonego progu. Ich główną zaletą jest niska złożoność obliczeniowa, co umożliwia ich łatwe skalowanie do danych o bardzo wielu wymiarach. Metody filtrujące nie opierają się na żadnym algorytmie klasyfikacyjnym, co może prowadzić do generowania modeli o niższej dokładności w porównaniu do innych metod selekcji. Metody te nie uwzględniają interakcji między cechami, w wyniku czego możliwe jest przeoczenie istotnych zależności występujących w danych. Są najczęściej

stosowane jako faza przetwarzania wstępnego, w celu zmniejszenia wymiarowości zbioru [12]. Na rysunku 5 przedstawiony został ogólny schemat metod filtrujących.



Rysunek 5. Schemat metod filtrujących

Do najbardziej popularnych metod filtrujących należą:

A. Informacja wzajemna

Informacja wzajemna to metoda pomiaru nieliniowych relacji pomiędzy dwiema zmiennymi losowymi. Służy do ilościowego określenia zależności między badaną cechą, a przynależnością do danej klasy. Wartość informacji wzajemnej wskazuje w jakim stopniu znajomość wartości jednej zmiennej zmniejsza niepewność drugiej. W przypadku mikromacierzy zmienne losowe odnoszą się do badanych genów, które opisywane są wspólnym wielowymiarowym rozkładem prawdopodobieństwa wobec określonej liczby próbek. Informacja wzajemna między dwiema zmiennymi losowymi X i Y może być zdefiniowana jako:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p_{(X,Y)}(x, y) * \log \left(\frac{p_{(X,Y)}(x, y)}{p_X(x) * p_Y(y)} \right)$$

gdzie $p(x, y)$ to łączne prawdopodobieństwo zaobserwowania wartości x i y dla zmiennych X i Y , a $p(x)$ i $p(y)$ oznaczają brzegowe prawdopodobieństwa w rozkładach zmiennych X i Y . Prawdopodobieństwo łączne odnosi się do prawdopodobieństwa zaobserwowania określonej wartości zarówno dla zmiennej X , jak i dla zmiennej Y w tym samym czasie. Prawdopodobieństwo brzegowe $p(x)$ reprezentuje prawdopodobieństwo zaobserwowania określonej wartości x dla zmiennej X , natomiast prawdopodobieństwo brzegowe $p(y)$ reprezentuje prawdopodobieństwo zaobserwowania określonej wartości y dla zmiennej Y . Gdy stosunek łącznego prawdopodobieństwa do iloczynu prawdopodobieństwa brzegowego jest wysoki, wskazuje to na istnienie znaczącej zależności między dwiema zmiennymi. W przypadku, kiedy wartość ta będzie równa零, dwie zmienne losowe są niezależne [11, 13]. Z całego zestawu wybierane są cechy, które charakteryzują się najwyższymi wartościami zależności pomiędzy cechami.

B. Test niezależności chi-kwadrat (χ^2)

Test niezależności chi-kwadrat to statystyczna metoda nieparametryczna służąca do oceny zależności między badaną cechą, a przynależnością do określonej klasy. Jest to technika, która ocenia każdy gen indywidualnie. Celem tej metody jest identyfikacja genów, których przynależność do danej klasy jest statystycznie istotna. Metoda ta obejmuje kilka kroków. Dla każdej klasy wyznaczona zostaje obserwowana częstość występowania unikalnej wartości ekspresji wszystkich genów. Informacja ta przechowywana jest w dwuwymiarowej tablicy korelacyjnej. Test χ^2 zakłada, że dwie zmienne są niezależne, co oznacza, że zmiana jednej zmiennej nie wpływa na rozkład drugiej. Wobec tego dla każdej komórki w tabeli korelacji wyznaczana zostaje częstość oczekiwana stanowiąca iloczyn sumy wystąpień unikalnych wartości ekspresji dla kolumn (genów) i wierszy (klas) podzielona przez całkowitą wielkość próby. Następnie dla każdego genu obliczana jest statystyka χ^2 , określająca ilościowo poziom w jakim obserwowane częstości różnią się od oczekiwanych. Wartość χ^2 dla danego genu może być zdefiniowana jako:

$$\chi^2 = \sum_{j=1}^r \sum_{s=1}^c \frac{(n_{js} - \mu_{js})^2}{\mu_{js}}$$

gdzie r to liczba genów, c to liczba klas, n_{js} to obserwowana częstość wystąpienia j -ego elementu w klasie s oraz μ_{js} to oczekiwana częstość wystąpienia j -ego elementu w klasie s . Następnie uzyskane wartości statystyki χ^2 porównane zostają do teoretycznego rozkładu χ^2 , który zdefiniowany jest przez stopnie swobody df oraz założony poziom istotności α . Stopnie swobody wyznaczane są następująco:

$$df = (r - 1) * (c - 1)$$

Wynikiem porównania jest wartość p , która stanowi prawdopodobieństwo uzyskania takiej samej lub bardziej skrajnej wartości statystyki χ^2 dla założonego poziomu istotności przeprowadzanego testu. Wartości są szeregowane w rosnącej kolejności wartości p . Geny charakteryzujące się niższymi wartościami p , przekraczającymi ustalony punkt istotności α są uważane za statystycznie istotne. Oznacza to wówczas istnienie zależności pomiędzy badaną cechą, a przynależnością do określonej klasy [12, 14, 15].

C. Analiza wariancji (ANOVA)

Analiza wariancji (ANOVA) jest metodą statystyczną stosowaną do porównywania wartości średnich w kilku porównywanych klasach. W kontekście selekcji cech test ten może być wykorzystany do oceny istotności związku pomiędzy poszczególnymi cechami, a przynależnością do określonej klasy. Metoda analizy wariancji zakłada, że we wszystkich porównywanych cechach

wartości charakteryzują się rozkładem normalnym. Statystyka ANOVA nazywana jest statystyką F , którą wyznaczyć można wykonując następujące kroki:

- 1) Międzygrupowe wyznaczenie różnic, czyli obliczenie sumy kwadratów odchyleń średnich w poszczególnych grupach od średniej ogólnej:

Suma kwadratów odchyleń wyniku od średniej w grupie:

$$SS_B = \sum_{i=1}^k n_k (\bar{X}_k - \bar{X})^2$$

Średni kwadrat międzygrupowy stanowiący miarę rozproszenia średnich w grupach w stosunku do średniej ogólnej:

$$MS_B = SS_B / df_B$$

- 2) Wewnętrzgrupowe wyznaczenie różnic, czyli suma kwadratów odchyleń wszystkich wyników od średnich z odpowiadających im grup:

Suma kwadratów odchyleń średniej w grupie od średniej ogólnej:

$$SS_W = \sum_{i=1}^k (n_k - 1) \sigma_k^2$$

Średni kwadrat wewnętrzgrupowy stanowiący miarę rozproszenia wyników w obrębie grupy, do której należy dany wynik:

$$MS_W = SS_W / df_W$$

- 3) Obliczenie statystyki F :

$$F = \frac{MS_B}{MS_W}$$

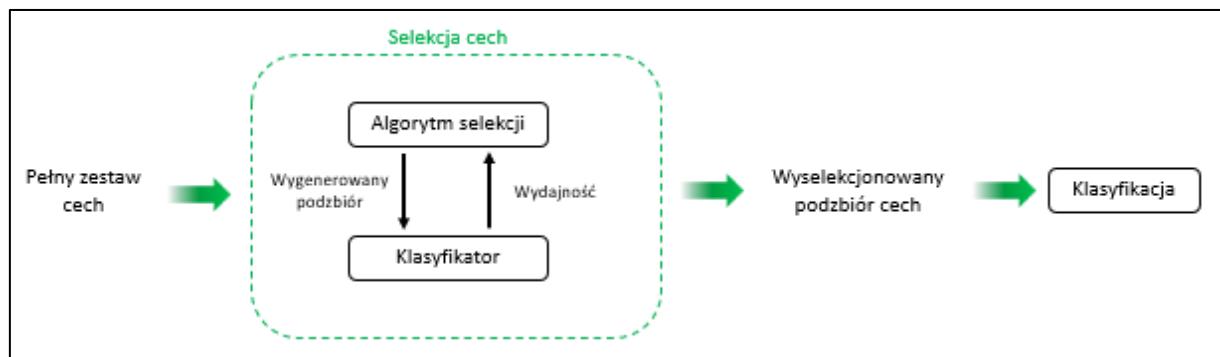
gdzie k to liczba grup, n_k to liczba próbek w grupie k , σ_k to odchylenie standardowe w grupie k , $df_B = (k - 1)$ to międzygrupowa liczba stopni swobody, $df_W = (N - k)$ to wewnętrzgrupowa liczba stopni swobody, N to liczba próbek.

Wartość MS_B reprezentuje zmienność średnich wartości cech względem siebie. Wyznacza o ile średnie wartości grup odbiegają od średniej ogólnej. Im większa jest wartość, tym większa jest różnica pomiędzy średnimi grupami, co wskazuje na to, że cecha może mieć znaczący wpływ na przynależność do danej klasy. Wartość MS_W reprezentuje średnią zmienność każdej cechy. Dokonuje pomiaru, jak rozłożone są poszczególne punkty danych w odpowiednich cechach. Niska

wartość sugeruje, że punkty w każdej grupie są stosunkowo bliskie ze średniej danej grupy. Sugeruje to, że cecha może nie przyczyniać się znacząco do zmienności między grupami. Finalny wynik statystyki F określa ilościowo, o ile większa jest zmienność między średnimi grupowymi w stosunku do zmienności wewnętrz grup. Jeśli średnie międzygrupowe są znacząco różne, a punkty danych w każdej grupie są stosunkowo ciasno skupione to wynik F będzie wysoki. Następnie uzyskane wartości statystyki F porównane zostają do teoretycznego rozkładu F , który zdefiniowany jest przez stopnie swobody df oraz założony poziom istotności α . W przypadku, gdy uzyskana wartość jest wyższa od wartości teoretycznej to wówczas prawdziwe jest stwierdzenie, że zależność pomiędzy badaną cechą, a przynależnością do określonej klasy jest statystycznie istotna [16, 17].

2.2.2 Metody opakowujące

W podejściu opakowującym wybór optymalnego podzbioru odbywa się w interakcji z klasyfikatorem. Metody opakowujące wykorzystują dokładność wybranego algorytmu uczenia maszynowego jako metrykę pomocną w wyborze podzbioru genów, który osiąga najwyższą wydajność predykcyjną dla określonego modelu. Główną zaletą tej metody podczas wyboru optymalnego podzbioru jest branie pod uwagę zależności pomiędzy cechami. Ze względu na wysoką złożoność obliczeniową dla dużych zbiorów są najczęściej stosowane dla wstępnie wyselekcjonowanych danych. Zapobiega to dodatkowo nadmierнемu dopasowaniu modelu prowadzącemu do wyboru cech specyficznych dla danych treningowych, które nie uogólnią dobrze modelu [18, 19]. Na rysunku 6 przedstawiony został ogólny schemat metod opakowujących.



Rysunek 6. Schemat metod opakowujących

Do najbardziej popularnych metod opakowujących zaliczają się:

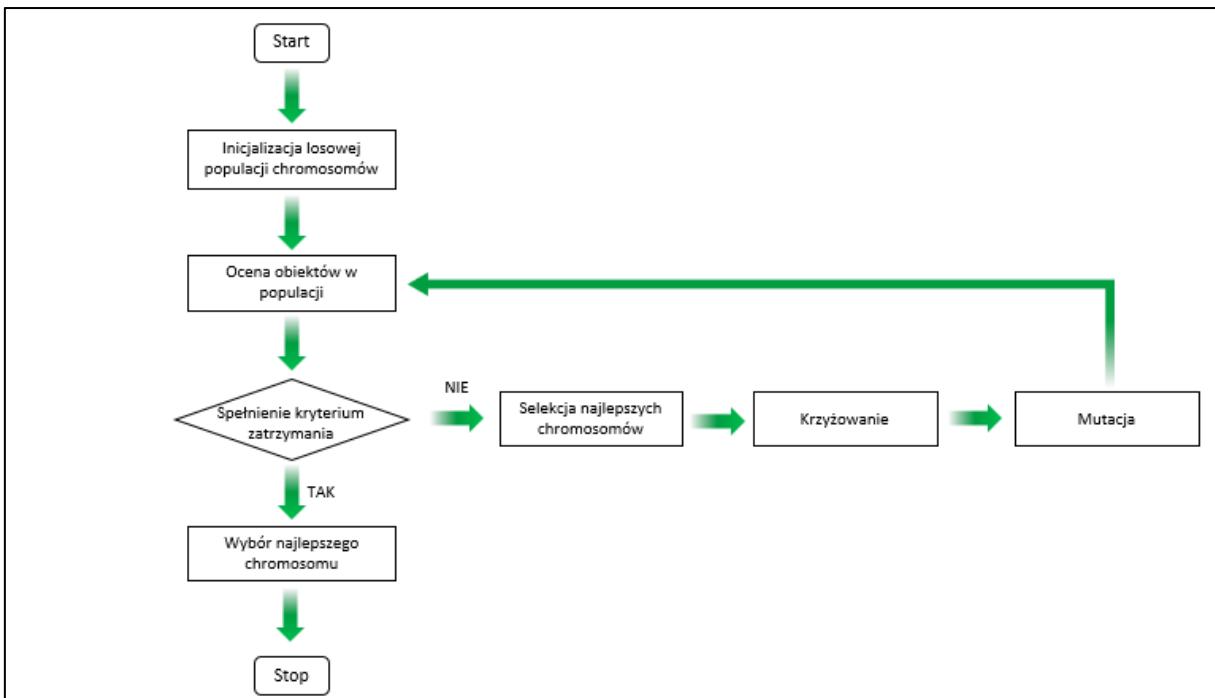
A. Algorytm genetyczny

Algorytm genetyczny (ang. *Genetic Algorithm*) jest heurystyczną techniką optymalizacji i wyszukiwania najbardziej informacyjnych cech, motywowaną procesem naturalnej ewolucji. Algorytm operuje na populacji potencjalnych rozwiązań (chromosomów), stanowiących ciągi symboli binarnych. Operacje w algorytmie są iteracyjnymi procedurami manipulującymi kolejnymi

populacjami chromosomów w celu wytworzenia nowych. Każdy chromosom reprezentuje potencjalne rozwiązanie problemu, którego skuteczność oceniana jest za pomocą funkcji dopasowania. Algorytm genetyczny składa się z wielu komponentów, do których zaliczają się: kodowanie chromosomów, ocena dopasowania, selekcja, krzyżowanie i mutacja.

Proces kodowania polega na określeniu sekwencji znaków binarnych w chromosomach, które stanowią potencjalne rozwiązania problemu. W kontekście selekcji cech długość każdego chromosomu odpowiada liczbie genów w badanym zbiorze danych. Wykorzystując indeksowanie pozycyjne cech każdy bit w chromosomie odpowiada obecności (1) lub nieobecności (0) danego genu ze zbioru definiując w ten sposób kolejne podzbiory. Początkowa populacja chromosomów generowana jest losowo. Jej wielkość powinna być co najmniej równa długości chromosomu, tak aby chromosomy z każdej populacji obejmowały przestrzeń po szukiwań, natomiast nie powinna być ona zbyt duża, ponieważ mogłyby spowodować znaczne spowolnienie algorytmu. Każdy chromosom w populacji jest oceniany za pomocą funkcji dopasowania. Określa ona ilościowo dokładność wybranego algorytmu uczenia maszynowego jako metrykę wydajności predykcyjnej modelu, przypisując każdemu chromosomowi wartość liczbową. W oparciu o uzyskane wyniki wydajności modelu przeprowadzany jest etap selekcji. Ma on na celu wybranie najlepiej przystosowanych genów i odrzucenie tych, które nie nadają się do rozwiązania obecnego problemu. Chromosomy szeregowane są w malejącej kolejności wartości wydajności, te które charakteryzują się najwyższymi wartościami zostają wybrane. Chromosomy wybrane z populacji początkowej są rekombinowane w celu utworzenia nowej populacji. Etap rekombinacji składa się z operacji krzyżowania i mutacji. Operator krzyżowania łączy dwa wybrane chromosomy z populacji początkowej w celu wytworzenia jednego lub dwóch nowych chromosomów w zależności od zaimplementowanej techniki. Jednym z powszechnie stosowanych sposobów jest jednopunktowy operator krzyżowania. Punkt skrzyżowania dwóch chromosomów jest wybierany losowo z jednakowym prawdopodobieństwem. Chromosomy potomne są konstruowane ze znaków pierwszego chromosomu rodzicielskiego występujących przed punktem skrzyżowania i znaków drugiego chromosomu rodzicielskiego występujących po punkcie skrzyżowania. Powstały w ten sposób chromosom poddany zostaje etapowi mutacji. Operator mutacji polega na wprowadzeniu małej, przypadkowej zmiany w chromosomie w celu zachowania różnorodności populacji. Mutacja zachodzi w przypadku, gdy losowo wygenerowana liczba z przedziału [0,1] jest niższa od wcześniej określonego współczynnika mutacji. Wówczas wartość jednego z bitów w chromosomie jest odwracana. Wskaźniki mutacji są zazwyczaj bardzo małe, np. 0,01. Po procesie rekombinacji powstałe chromosomy przekazywane są do następnego pokolenia. Algorytm genetyczny powtarzany jest przez wiele pokoleń, aż do osiągnięcia odpowiedniego kryterium. Kryterium zatrzymania może obejmować maksymalną liczbę pokoleń, zaobserwowanie zbieżności wartości

dopasowania dla kolejnych pokoleń lub osiągnięcie rozwiązania, które spełnia zestaw ograniczeń [20, 21]. Na rysunku 7 przedstawiony został schemat algorytmu genetycznego.



Rysunek 7. Schemat algorytmu genetycznego

B. Algorytm sztucznej kolonii pszczół

Algorytm sztucznej kolonii pszczół (ang. *Artificial Bee Colony Algorithm*) wzorowany jest na zachowaniu pszczół miodnych przy zdobywaniu pożywienia. Należy on do grupy algorytmów rojowych, czyli algorytmów inspirowanych zbiorowym zachowaniem zwierząt, w którym jednostki współpracują ze sobą w celu znalezienia optymalnego rozwiązania problemu. Przestrzeń poszukiwań algorytmu ABC przedstawiona jest jako rój pszczół miodnych, w którym każda pszczoła poszukuje potencjalnego źródła pożywienia zawierającego określoną ilość nektaru. W kontekście selekcji cech problem sprowadza się do wyboru najbardziej informacyjnych cech ze zbioru, przy czym źródło pożywienia reprezentuje potencjalny podzbiór cech, natomiast ilość nektaru nawiązuje do wartości dopasowania, która odzwierciedla dokładność danego podzbioru dla wybranego modelu predykcyjnego. Algorytm ABC zbudowany jest w oparciu o trzy rodzaje pszczół, które stanowią podstawowe elementy algorytmu: pszczoły zatrudnione, pszczoły obserwatorzy i pszczoły zwiadowcy. Model selekcji rozpoczyna się od inicjalizacji losowych rozwiązań, gdzie każda z zatrudnionych pszczół reprezentuje potencjalne źródło pożywienia. Pierwotna populacja wyznaczana jest w następujący sposób:

$$x_{i,j} = x_j^{\min} + \lambda(x_j^{\max} - x_j^{\min})$$

gdzie, $x_{i,j}$ stanowi j -tą cechę dla i -tej pszczoły w przestrzeni poszukiwań $[x_j^{\max}, x_j^{\min}]$, przy czym $i = 1 \dots N, j = 1 \dots D$, gdzie N to liczba pszczołów zatrudnionych, D to liczba cech w zbiorze danych, a λ jest losową liczbą z przedziału $[0,1]$. Algorytm wymaga uprzedniego zdefiniowania wielkości generowanej populacji. Jest to parametr, który należy dostosować w zależności od charakterystyki problemu oraz dostępnych zasobów obliczeniowych. W przypadku przeprowadzania selekcji dla dużych zbiorów danych, większa populacja powinna być korzystniejsza dla skutecznego badania przestrzeni dostępnych rozwiązań, jednak wiąże się to ze znacznie większą złożonością obliczeniową problemu. Z drugiej strony, w przypadku mniejszych zbiorów danych, wprowadzenie większej populacji może prowadzić do zwiększenia złożoności obliczeniowej bez znaczącego wzrostu wydajności. Kolejne kroki algorytmu obejmują aktualizację wygenerowanej populacji źródeł pożywienia. Faza pszczołów zatrudnionych rozpoczyna się od sprawdzenia ilości nektaru $F(x_i)$ dla każdego rozwiązania. Oznacza to, że każdy podzbiór obecny w populacji jest oceniany za pomocą funkcji dopasowania, przy zastosowaniu wybranego algorytmu uczenia maszynowego w celu określenia wydajności predykcyjnej danego rozwiązania. Następnie faza pszczołów zatrudnionych obejmuje opracowanie nowego potencjalnego rozwiązania dla każdej zatrudnionej pszczoły. Proces ten obejmuje następujące kroki:

- 1) Pierwotne rozwiązanie dla i -tej zatrudnianej pszczoły jest kopowane do nowego rozwiązania kandydującego:

$$v_i = x_i$$

- 2) Aktualizowany jest tylko jeden parametr pierwotnego rozwiązania:

$$v_{i,j} = x_{i,j} + \phi(x_{i,j} - x_{k,j})$$

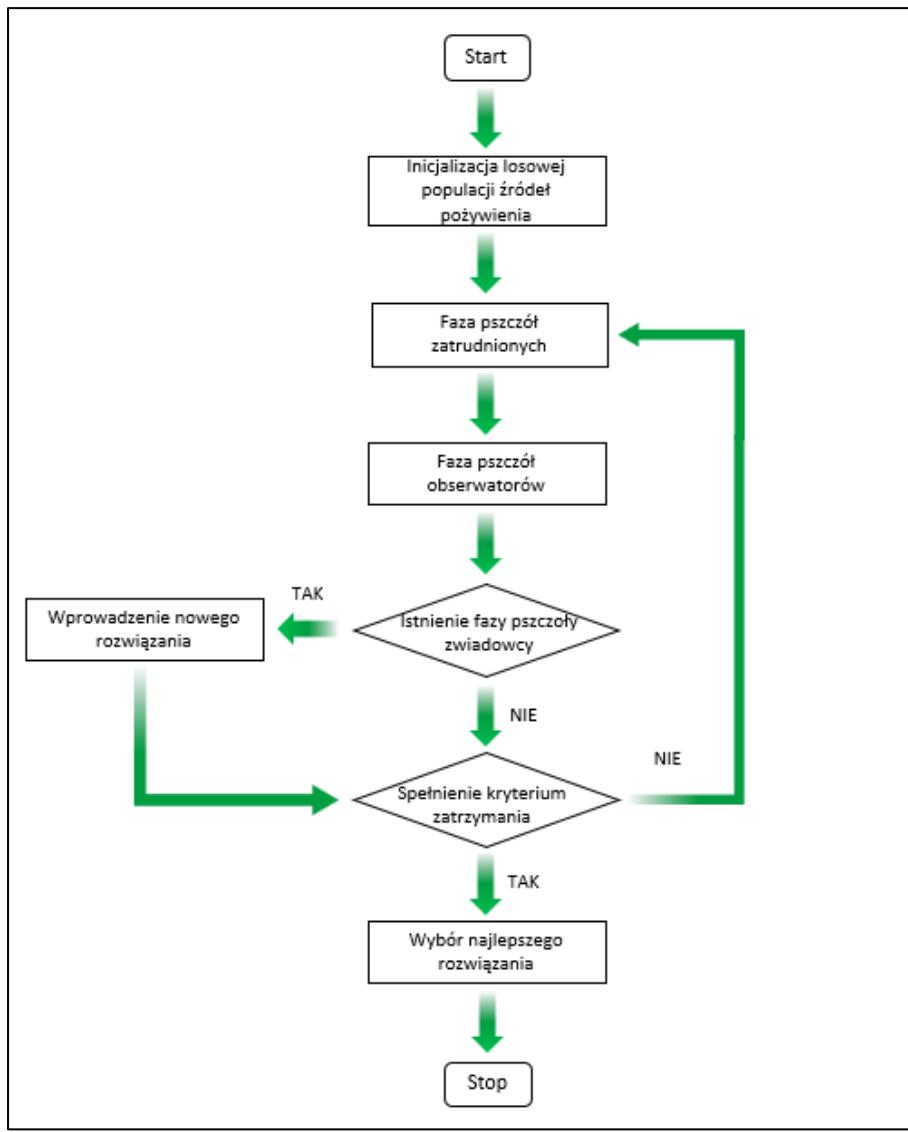
gdzie $v_{i,j}$ to j -ta cecha dla i -tej pszczoły rozwiązania kandydującego, $x_{i,j}$ to j -ta cecha dla i -tej pszczoły rozwiązania pierwotnego, $x_{k,j}$ to j -ta cecha dla k -tej pszczoły rozwiązania pierwotnego, przy czym $i, k \in \{1, 2, \dots, N\}$ oraz $i \neq k, j \in \{1, 2, \dots, D\}$, natomiast ϕ to liczba losowa z przedziału $[-1, 1]$.

Następnie sprawdzana jest wydajność predykcyjna rozwiązania kandydującego. W przypadku, gdy wartość ta jest większa od rozwiązania pierwotnego, to rozwiązanie zostaje zastąpione rozwiązaniem kandydującym. Kolejnym etapem algorytmu jest faza pszczołów obserwatorów. Liczba źródeł pożywienia dla pszczołów obserwatorów jest taka sama jak dla pszczołów zatrudnionych. W tej fazie algorytmu sztucznej kolonii pszczołów, wszystkie pszczoły zatrudnione dzielą się swoimi informacjami o ilości nektaru $F(x_i)$ znajdującego się w źródle pożywienia. Wraz ze wzrostem ilości

nektaru, wzrasta prawdopodobieństwo wyboru tego źródła pożywienia przez pszczoły obserwatorów. Prawdopodobieństwo wyboru danego źródła pożywienia i -tej pszczoły zatrudnionej przez pszczoły obserwatorów wyznaczane jest następująco:

$$p_i = \frac{F(x_i)}{\sum_{i=1}^N F(x_i)}$$

Kierując się znanym prawdopodobieństwem pszczoły obserwatorzy wylatują do źródła pożywienia x_i , zatrzymując się w jego sąsiedztwie. Wówczas sprawdzona zostaje dla niego ilość nektaru. W przypadku, gdy jest ona większa od docelowego źródła pożywienia, to pszczoła obserwator zamienia się w pszczołę zatrudnioną, a w kolejnych iteracjach algorytmu brany pod uwagę jest wyłącznie podzbiór charakteryzujący się wyższą wartością dopasowania. Jeżeli źródło pożywienia charakteryzujące się największą ilością nektaru nie zmienia się przez predefiniowaną liczbę cykli, wówczas zakłada się, że dane źródło pożywienia należy opuścić i rozpocząć fazę pszczoły zwiadowcy. Wobec tego źródło pożywienia zastępowane jest przez dowolne wybrane rozwiązanie wewnętrz przeszukiwanej przestrzeni rozwiązań. Faza ta zapobiega stagnacji pozostałych populacji pszczół. W algorytmie ABC predefiniowana liczba cykli jest istotnym parametrem sterującym, który nazywany jest granicą odrzucenia. W przypadku mniejszej wartości cykli, pszczoły szybciej staną się zwiadowcami, co przyczynia się do większej eksploatacji przestrzeni rozwiązań, ponieważ rozwiązania, które nie ulegają poprawie są zastępowane. Może to jednak prowadzić do ograniczonego przeszukiwania obiecujących obszarów. Z drugiej strony, zastosowanie większej ilości cykli spowodują, że pszczoły pozostaną zatrudnione przez dłuższy czas, nawet gdy ich rozwiązania nie będą ulegały poprawie. Może to prowadzić do przedwczesnej zbieżności w przypadku utknięcia w obszarach o nieoptimalnych rozwiązaniach. Algorytm kontynuuje iterację przez poszczególne fazy, aż do momentu spełnienia kryterium zakończenia, którym może być z góry określona liczba iteracji lub zbieżność rozwiązań [22-25]. Na rysunku 8 przedstawiono schemat algorytmu sztucznej kolonii pszczół.



Rysunek 8. Schemat algorytmu sztucznej kolonii pszczół

C. Algorytm roju cząstek

Algorytm roju cząstek (ang. *Particle Swarm Optimization Algorithm*) wzorowany jest na zachowaniach społecznych stada ptaków, które dzięki współpracy dążą do osiągnięcia wspólnego celu. Zamysłem algorytmu jest znalezienie optymalnego rozwiązania problemu wykorzystując cząstki poruszające się w wielowymiarowej przestrzeni. W kontekście selekcji cech każda z cząstek charakteryzuje się określoną wartością dopasowania danego podzbioru do modelu. Cząstki o najwyższych wartościach dopasowania ustanawiają drogę dla pozostałych cząstek, które podążając za nimi przelatują przez obszar przestrzeni problemowej. Każda cząstka ma swoją pozycję w przestrzeni poszukiwań w danym momencie oraz prędkość z jaką się porusza. Ponadto cząstki rejestrują najlepsze dotychczasowe znalezione rozwiązanie przez każdą z cząstek (rozwiązanie lokalne), a także najlepsze rozwiązanie z całego roju (rozwiązanie globalne). Prędkość ruchu cząstek to wektor, który zależy od położenia najlepszego lokalnego i globalnego rozwiązania

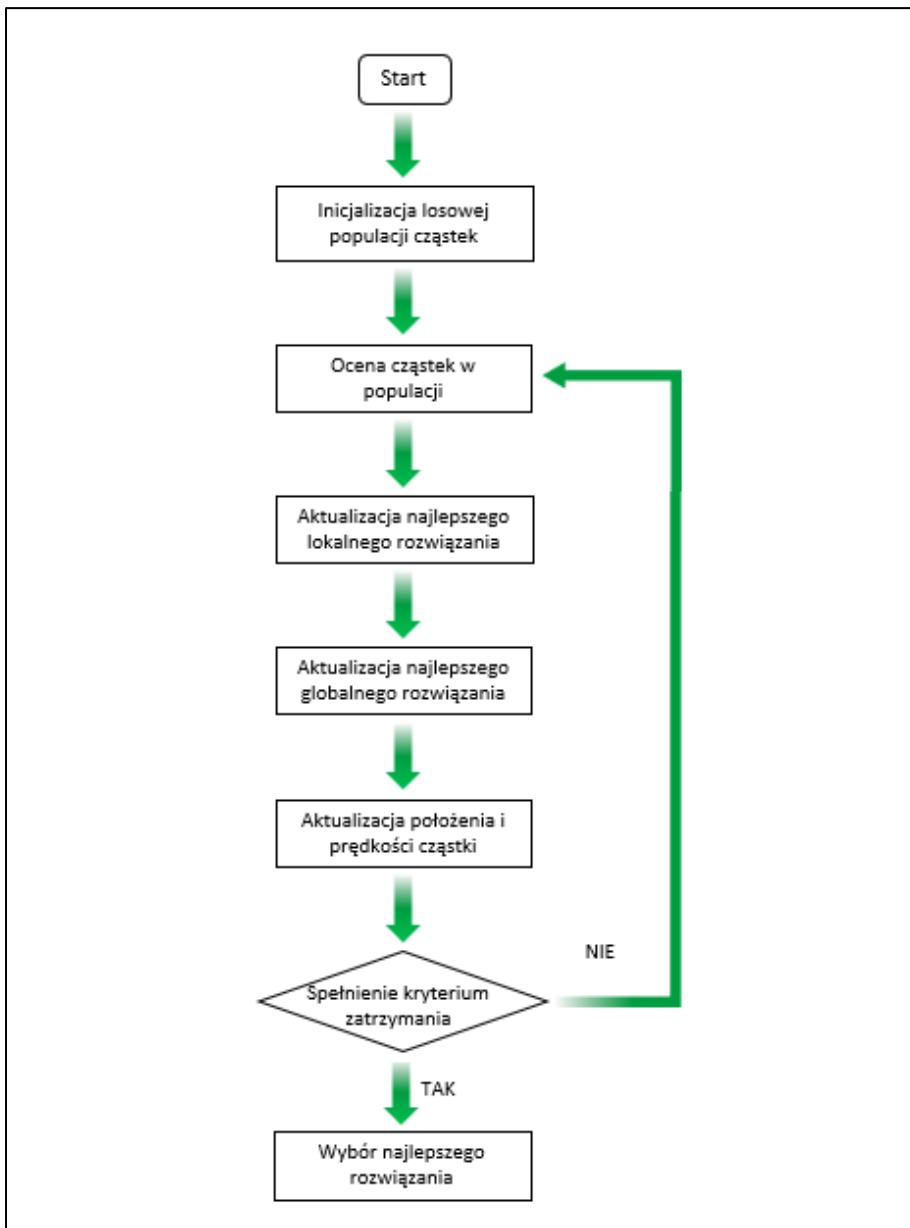
oraz od prędkości cząstek w poprzednich krokach algorytmu. Wzory na obliczenie położenia x_i oraz prędkości v_i danej cząstki i są następujące:

$$x_{ij}^{t+1} = x_{ij}^t + v_{ij}^{t+1}$$

$$v_{ij}^{t+1} = \omega * v_{ij}^t + c_1 * r_1 * (p_{ij} - x_{ij}^t) + c_2 * r_2 * (p_{gj} - x_{ij}^t)$$

gdzie t oznacza t -tą iterację procesu, j oznacza j -tą cechę w przestrzeni poszukiwań, ω to współczynnik bezwładności określający wpływ prędkości w poprzednim kroku algorytmu, c_1 to współczynnik dążenia do najlepszego lokalnego rozwiązania, c_2 to współczynnik dążenia do najlepszego globalnego rozwiązania, p_{ij} to położenie najlepszego lokalnego rozwiązania, p_{gj} to położenie najlepszego globalnego rozwiązania, r_1 oraz r_2 to liczby losowe z przedziału [0,1]. Działaniem algorytmu można sterować poprzez odpowiedni dobór parametrów, ponieważ od ich wartości zależy zachowanie poszczególnych cząstek. Jednak według wielu źródeł [26, 27] optymalną wartością współczynnika bezwładności jest 0,7, natomiast dla obu współczynników dążenia do najlepszych rozwiązań przypisywana jest wartość 2. Wartość współczynnika bezwładności ma wpływ na zdolność cząstek do zachowania poprzedniej prędkości. Im większa jest wartość tego parametru, tym większa jest zdolność cząstek do przeszukiwania nowych rejonów przestrzeni rozwiązań. Wyższe wartości współczynnika dążenia do najlepszego lokalnego rozwiązania powodują większą skłonność cząstki do oscylacji wokół swojej najlepszej pozycji, natomiast wyższe wartości współczynnika dążenia do najlepszego globalnego rozwiązania skutkują zwiększeniem tendencji do grupowania się cząstek wokół najlepszego globalnego rozwiązania.

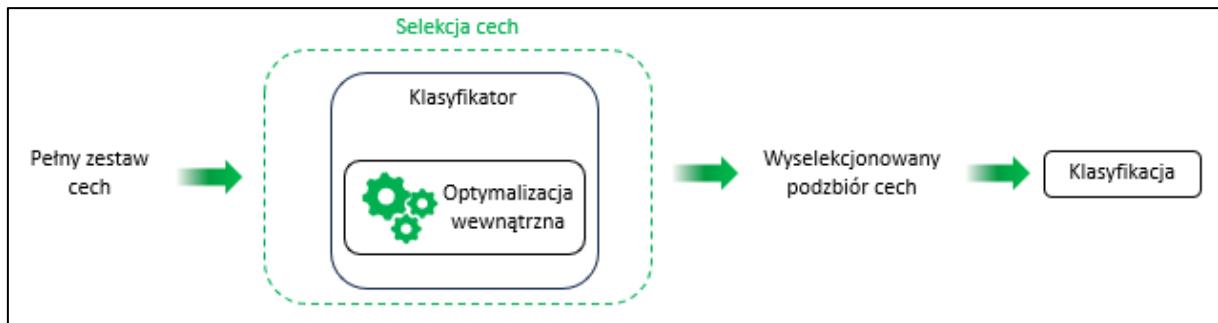
Algorytm roju cząstek rozpoczyna się od losowej inicjalizacji populacji cząstek, gdzie każda ma określoną pozycję oraz prędkość początkową. Następnie w kolejnych iteracjach dla każdej cząsteczki wyznaczona zostaje wartość dopasowania, która odzwierciedla dokładność danego podzbioru dla wybranego modelu predykcyjnego. W przypadku, gdy określone rozwiązanie jest lepsze od najlepszego rozwiązania lokalnego cząsteczki to zostaje ono zaktualizowane. Następnie cząsteczki porównują swoją obecną wartość dopasowania z najlepszym rozwiązaniem globalnym występującym w całym roju. Jeśli dane rozwiązanie jest większe to jest ono aktualizowane zgodnie z pozycją cząstki. Podczas każdej iteracji prędkość i położenie cząstki są aktualizowane na podstawie jej aktualnej pozycji oraz położeniu lokalnego i globalnego rozwiązania według wcześniej zaprezentowanego równania. Algorytm kontynuuje iterację przez poszczególne kroki, aż do momentu spełnienia kryterium zakończenia, którym może być z góry określona liczba iteracji. Wynikiem końcowym algorytmu jest podzbiór cech reprezentowany przez cząsteczkę charakteryzującą się najlepszym globalnym rozwiązaniem [28, 29]. Na rysunku 9 przedstawiony został schemat algorytmu roju cząstek.



Rysunek 9. Schemat algorytmu roju cząstek

2.2.3 Metody wbudowane

W metodach wbudowanych wybór najbardziej informacyjnych cech jest zintegrowany lub wbudowany w algorytm klasyfikatora. Określenie optymalnego podzbioru odbywa się na etapie wykonywania algorytmu klasyfikacji. Wówczas klasyfikator dostosowuje swoje parametry wewnętrzne i określa odpowiednie poziomy ważności dla każdej cechy, w celu uzyskania jak najlepszej dokładności algorytmu. Metody wbudowane stanowią rozwiązanie pośrednie pomiędzy metodami filtrującymi, a metodami opakowującymi łącząc w sobie cechy obu tych metod. Podobnie do metod filtrujących są one stosunkowo nisko kosztowne obliczeniowo. Z drugiej strony, uwzględniają one interakcję pomiędzy cechami, jednak w porównaniu do metod opakowujących są one znacznie mniej podatne na nadmierne dopasowanie [19, 30]. Na rysunku 10 przedstawiony został ogólny schemat metod wbudowanych.



Rysunek 10. Schemat metod wbudowanych

Do najbardziej popularnych metod zaliczają się:

A. Regresja grzbietowa

Regresja grzbietowa (ang. *Ridge regression*) to metoda regresji liniowej, wprowadzająca regularyzację L_2 do estymacji współczynników modelu w celu uniknięcia nadmiernego dopasowania. Regularyzacja L_2 polega na dodaniu do funkcji celu kary proporcjonalnej do kwadratu wartości współczynników regresji. Celem regresji grzbietowej jest minimalizacja funkcji, składającej się z dwóch składników: błędu dopasowania (sumy kwadratów różnic pomiędzy rzeczywistymi wartościami odpowiedzi, a przewidywanymi wartościami modelu) i kary regularizacyjnej L_2 . Ta technika jest szczególnie użyteczna, w przypadku występowania nadmiernej wielowymiarowości danych. W kontekście selekcji cech problem sprowadza się do wyboru najbardziej informacyjnych cech w oparciu o wielkość oszacowanych współczynników. W tym celu regresja grzbietowa opiera się minimalizacji następującego wyrażenia:

$$\sum_{i=1}^N \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

gdzie, N to liczba próbek, p to liczba cech w zbiorze danych, y_i reprezentuje przynależność do danej klasy dla i -tej próbki, x_{ij} reprezentuje wartość dla j -tej cechy i -tej próbki, β_j to współczynnik j -tej cechy, a λ oznacza parametr regularizacji. Im wyższa jest wartość λ , tym większa jest kara i dlatego współczynniki regresji są zmniejszane w kierunku zera. Współczynniki cech o niskim znaczeniu będą miały wartości bliskie zeru, co oznacza, że dana cecha ma niewielki wpływ na model. Poprzez odpowiednie kontrolowanie wartości wyznaczanych współczynników, możliwe jest zidentyfikowanie cech, które są najbardziej istotne dla modelu klasyfikacyjnego [31-33].

B. Las losowy

Las losowy to jeden z najbardziej popularnych algorytmów uczenia maszynowego. Metoda ta charakteryzuje się łatwą interpretacją istotności zmiennych pozwalającą na obliczenie, w jakim stopniu każda cecha przyczynia się do podjęcia decyzji. Lasy losowe składają się z drzew

decyzyjnych. Każde z nich zbudowane jest w oparciu o losowe wygenerowanie podzbiorów na podstawie całego zbioru danych. Proces budowania drzewa decyzyjnego rozpoczyna się od głównego węzła, do którego dodawane są kolejne węzły potomne. Podczas budowania drzewa decyzyjnego nie wszystkie cechy sąbrane pod uwagę przy wyznaczaniu reguły decyzyjnej w węźle. Jest to podejście wykorzystywane w celu ochrony algorytmu przed nadmiernym dopasowaniem modelu do danych. W każdym węźle drzewa algorytm używa miary zanieczyszczenia Gini'ego jako kryterium podziału. Miara ta pomaga w podjęciu decyzji, w jaki sposób należy podzielić dane w każdym węźle drzewa w celu utworzenia możliwie najbardziej jednorodnych węzłów podrzędnych. Współczynnik ten wyznaczany jest jako suma kwadratów prawdopodobieństw przynależności każdej kategorii do danego węzła. Miara Gini'ego mieści się w zakresie od 0 do 1, gdzie im niższa jest wartość w danym węźle, tym dana cecha ma większe znaczenie przy podejmowaniu decyzji. Suma miar dla wszystkich cech na węźle znormalizowana jest do wartości 1. Podczas tworzenia drzewa, algorytm próbuje różnych kombinacji cech i wybiera tę, która minimalizuje wartość. Proces ten jest powtarzany w każdym węźle drzewa, aż do momentu spełnienia kryterium zakończenia, którym może być z góry określona głębokość drzewa. Liczba drzew w lesie losowym jest z góry określona. Jej osiągnięcie świadczy o spełnieniu kryterium zakończenia budowy lasu w algorytmie. Po stworzeniu wszystkich drzew, ostateczna istotność cechy w lesie losowym jest obliczana jako średnia wartość miary zanieczyszczenia Gini'ego spowodowana przez tą cechę we wszystkich drzewach w lesie. Cechy, które przyczyniają się do większego zmniejszenia wartości miary, uważane są za istotniejsze w kontekście selekcji cech [34-36]

C. RFE-SVM

Rekurencyjna eliminacja cech w oparciu o metodę wektorów nośnych (RFE-SVM) to wbudowana metoda uczenia maszynowego, która dokonuje selekcji cech na podstawie wartości wag opisujących granicę decyzyjną liniowego klasyfikatora SVM. Główną zasadą RFE-SVM jest eliminowanie w każdej iteracji cech o najmniejszej wagie. Inicjalizacja algorytmu odbywa się poprzez trenowanie klasyfikatora korzystającego z metody wektorów nośnych na podstawie wszystkich cech obecnych w zbiorze danych. Wektory nośne w metodzie SVM decydują o przebiegu granicy rozdzielającej poszczególne klasy. Stanowią one granice marginesu separacji określającego maksymalną odległość pomiędzy wektorami cech z różnych klas. W oparciu o zbiór wektorów określone zostają wagi dla każdej cechy, za pomocą następującego równania:

$$w_i = \sum_k \alpha_k y_k x_k$$

gdzie w_i reprezentuje wartość wagi dla i-tej cechy ze zbioru, x_k reprezentuje k -ty wektor dla zbioru uczącego, α_k to ustalony podczas procesu trenowania klasyfikatora mnożnik Lagrange'a przypisany

do k -tego wektora, natomiast y_k określa przynależność k -tego wektora do danej klasy. Następnie na podstawie wyliczonych wag tworzony jest ranking cech dla kwadratu ich wartości:

$$c_i = (w_i)^2$$

Cechy o wyższych wagach są uważane za bardziej istotne, zajmując wyższą pozycję w rankingu. Następnie wykorzystując metodę rekurencyjnej eliminacji cech (RFE) każda cecha charakteryzująca się najniższym poziomem istotności w danej iteracji zostaje usunięta ze zbioru. Iteracje algorytmu odbywają się do momentu osiągnięcia predefiniowanej liczby cech w zbiorze, otrzymując najbardziej informacyjny podzbiór cech [37-39].

W tabeli 1 przedstawione zostały zalety oraz ograniczenia wymienionych metod selekcji najbardziej informacyjnych cech. Zgodnie z zestawionymi punktami wszystkie trzy metody charakteryzują się wieloma zaletami oraz ograniczeniami. Metody filtrujące są wydajne obliczeniowo i oferują prostotę interpretacji wyselekcjonowanych cech, ale nie uwzględniają zależności między nimi. Metody opakowujące dostosowują wybór cech do konkretnego modelu uczenia, wychwytując interakcję między cechami, ale są bardzo kosztowne obliczeniowo i podatne na nadmierne dopasowanie. Metody wbudowane biorą pod uwagę zależności między cechami łącząc proces uczenia modelu z wyborem cech oraz zapewniają mniejszą podatność na nadmierne dopasowanie, jednak ich wydajność może być ściśle powiązana z zastosowanym algorymem uczenia.

Biorąc pod uwagę powyższe, nie ma jednego uniwersalnego podejścia, które byłoby odpowiednie dla wszystkich zbiorów danych mikromacierzowych ze względu na ich różnorodność oraz złożoność problemu i wymagań modelu. Fakt ten sygnalizuje istotność stosowania dwuetapowej strategii selekcji cech, którą stanowią metody hybrydowe.

Tabela 1. Zalety i ograniczenia metod selekcji cech [11, 19, 40]

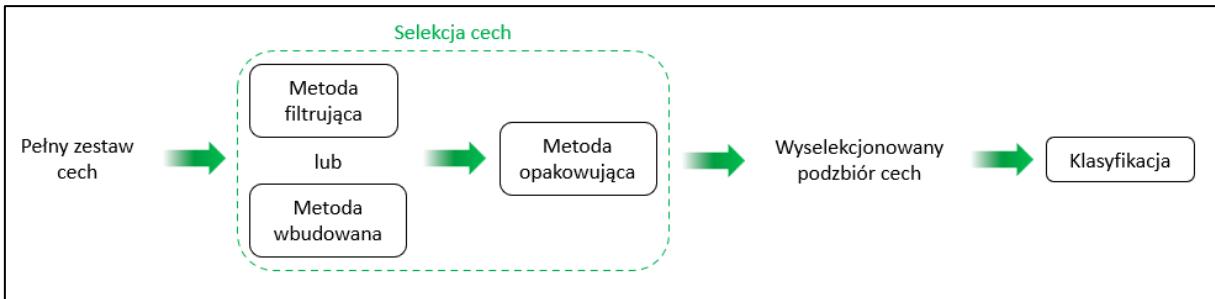
Metoda	Zalety	Ograniczenia
Filtrująca	<ul style="list-style-type: none"> - wysoka wydajność obliczeniowa ze względu na ocenianie każdej cechy niezależnie, - niezależne od modelu uczenia maszynowego, - łatwa interpretacja wyników, - możliwość skalowania do bardzo dużych zbiorów danych 	<ul style="list-style-type: none"> - brak interakcji pomiędzy cechami, - brak optymalizacji pod kątem określonego algorytmu uczenia maszynowego, - możliwość występowania wysokiej redundancji

Metoda	Zalety	Ograniczenia
Opakowująca	<ul style="list-style-type: none"> - zależna od modelu uczenia maszynowego, - wychwytywanie interakcji pomiędzy cechami, - możliwość optymalizacji dla konkretnego algorytmu uczenia 	<ul style="list-style-type: none"> - bardzo wysoki koszt obliczeniowy, ze względu na wielokrotną ocenę modelu dla różnych podzbiorów, - dążenie do nadmiernego dopasowania modelu
Wbudowana	<ul style="list-style-type: none"> - wybór cech jest połączony z procesem uczenia modelu, - wychwytywanie interakcji pomiędzy cechami, - znacznie szybsze od metod opakowujących, - mniejsza podatność na nadmierne dopasowanie w porównaniu z metodami opakowującymi 	<ul style="list-style-type: none"> - wydajność wybranego podzbioru może być ściśle powiązana z wydajnością zastosowanego modelu uczenia maszynowego, - nieznacznie wyższy koszt obliczeniowy od metod filtrujących

2.2.4 Metody hybrydowe

Metody hybrydowe łączą różne techniki selekcji cech w celu optymalizacji wyboru najbardziej informacyjnych cech. Metody te mają na celu uwzględnienie ograniczeń i wykorzystanie mocnych stron poszczególnych metod.

Metody hybrydowe nie ograniczają się do konkretnych technik składowych selekcji. Mogą one zawierać różne algorytmy służące do wyboru cech. Jest to szczególnie istotna cecha w przypadku wielowymiarowych zbiorów danych, dla których kosztowe obliczeniowo metody opakowujące mogą okazać się niepraktyczne bez wcześniejszej redukcji wymiarowości za pomocą technik filtrujących lub wbudowanych. Na rysunku 11 przedstawiony został ogólny schemat metod hybrydowych.



Rysunek 11. Schemat metod hybrydowych

Najczęstsze wykorzystanie metod hybrydowych uwzględnia dwuetapową strategię selekcji cech. Pierwszy krok obejmuje zastosowanie metod filtrujących lub wbudowanych. Ze względu na szybkość i prostotę są one optymalnym wyborem w celu zredukowania początkowego rozmiaru zbioru. Służy to ograniczeniu przestrzeni poszukiwań oraz zredukowania kosztu obliczeniowego dla kolejnego etapu. Drugi krok stanowiący właściwą selekcję cech wykorzystuje jedną z metod opakowujących, ze względu na ich zdolność do wyszukiwania złożonych interakcji między cechami. Zastosowanie dwuetapowej kombinacji metod pozwala na osiągnięcie równowagi między wydajnością obliczeniową modelu, a uchwyceniem interakcji między cechami.

Pomimo swoich zalet, metody hybrydowe nie są pozbawione ograniczeń. Wykorzystanie metod filtrujących na pierwszym etapie selekcji może skutkować usunięciem cech, które mogą okazać się kluczowe w przypadku zbiorów danych, w których interakcje między cechami odgrywają istotną rolę dla dokładności modelu. Z tego powodu potencjalne zastosowanie metod wbudowanych na pierwszym etapie selekcji powinno mieć pozytywny wpływ na jakość budowanego modelu hybrydowego.

Metody hybrydowe zapewniają praktyczne rozwiązanie wyzwań związanych z wielowymiarowymi danymi, jednak niezwykle istotnym czynnikiem jest odpowiedni wybór metod składowych w celu uzyskania optymalnego modelu selekcji najbardziej informacyjnych cech [19].

2.3 Metody klasyfikacji

Klasyfikacja to zadanie polegające na przydzieleniu klasy dla próbki wejściowej, dla której jest ona nieznana. Proces klasyfikacji opiera się na podstawie znanych wartości cech. Przypisania do klasy dokonuje wcześniej wytrenowany model, który uczony jest w oparciu o próbki, dla których klasa oraz wartości cech są znane. W kontekście selekcji cech klasyfikacja odbywa się za pomocą wyselekcjonowanych podzbiorów, które stanowią najbardziej informacyjną część zbioru, a tym samym najlepiej określają przynależność do danej klasy.

Przed przystąpieniem do procesu klasyfikacji należy podzielić dane na grupy treningowe oraz walidacyjne zachowując równowagę próbek ze wszystkich dostępnych klas w obu grupach. Zbiór treningowy to część danych, która używana jest do trenowania klasyfikatora, natomiast zbiór

walidacyjny to część danych, która używana jest do oceny wydajności badanego klasyfikatora po jego wytrenowaniu.

Istnieje wiele algorytmów klasyfikacji. Nie ma jednego uniwersalnego podejścia, które działałaby na wszystkie zbiory danych osiągając zawsze najlepszą skuteczność. Do najczęściej stosowanych algorytmów klasyfikacji należą:

A. Naiwny klasyfikator Bayesa

Naiwny klasyfikator Bayesa (ang. *Naive Bayes classifier*) to algorytm statystyczny oparty na twierdzeniu Bayesa. Twierdzenie Bayesa opisuje relację między prawdopodobieństwem warunkowym dwóch zdarzeń. Mówi, że prawdopodobieństwo wystąpienia jednego zdarzenia, przy założeniu, że inne zdarzenie już zaszło, jest proporcjonalne do prawdopodobieństwa wystąpienia obu tych zdarzeń. W kontekście klasyfikacji, twierdzenie Bayesa pomaga oszacować prawdopodobieństwo przynależności danej próbki do określonej klasy. Naiwny klasyfikator Bayesa zakłada, że cechy są warunkowo niezależne względem klasy. Oznacza to, że wartość każdej cechy jest niezależna od wartości innych cech. Klasyfikator ten charakteryzuje się dużą dokładnością i skalowalnością, nawet dla bardzo dużych zbiorów danych [41, 42].

B. Metoda wektorów nośnych

Metoda maszyny wektorów nośnych (ang. *Support Vector Machine – SVM*) to metoda klasyfikacji wykorzystywana do separacji danych w przestrzeni. Jej celem jest znalezienie hiperpłaszczyzny oddzielającej różne klasy danych, za budowę której odpowiedzialne są wektory nośne. Hiperpłaszczyzny budowane są w taki sposób, aby maksymalizować jej odległość od najbliższych punktów należących do określonych klas. Im większy jest taki margines, tym mniejszy jest błąd klasyfikatora [43].

W przypadku tego klasyfikatora jednym ze sposobów radzenia sobie z nieliniowo rozdzielnymi klasami w złożonych zbiorach danych jest zastosowanie techniki matematycznej, zwanej gaussowską radialną funkcją bazową (ang. *radial basis function – RBF*). Służy ona do przekształcania przestrzeni badanych cech w taki sposób, aby możliwe było utworzenie wielomianowej granicy oddzielającej poszczególne klasy [44]. W celu optymalizacji wydajności modelu tego klasyfikatora istnieją dwa parametry C oraz gamma. Parametr regularizacyjny C kontroluje kompromis jak bardzo model ma dopasować się do danych. Dla dużych wartości C model będzie bardziej skłonny do poprawnej klasyfikacji każdego punktu, nawet jeśli oznacza to skomplikowany model decyzyjny, natomiast dla małych wartości C model będzie bardziej tolerancyjny względem klasyfikacji na danych treningowych, co może prowadzić do uzyskania gładkiego modelu. Parametr gamma kontroluje zakres wpływu każdej próbki w zbiorze. Niskie wartości gamma oznaczają, że wpływ jest szeroki, co oznacza, że punkt wpływa na klasyfikację

innych punktów na dużej odległości, natomiast wysokie wartości gamma oznaczają, że wpływ jest ograniczony do punktów bliższych linii decyzyjnej [45].

Metoda SVM jest klasyfikatorem binarnym, z tego powodu jej użycie dla zbiorów zawierających więcej niż dwie klasy wymaga budowania wielu modeli klasyfikacyjnych. Przykładami dwóch podejść do tego problemu są: *One vs One* oraz *One vs Rest*. Podejście *One vs One* polega na tym, że w czasie treningu tworzonych jest odpowiednia liczba klasyfikatorów binarnych porównujących poszczególne pary klas, natomiast podejście *One vs Rest* polega na trenowaniu klasyfikatorów binarnych budowanych w oparciu o wektory z jednej klasy wobec połączonych wektorów z pozostałych klas [44].

C. Las losowy

Las losowy (ang. *Random Forest*) to model klasyfikacyjny składający się z wielu drzew decyzyjnych. Każde drzewo konstruowane jest osobno na podstawie różnych podzbiorów danych treningowych oraz losowego wyboru cech, co skutkuje poprawą różnorodności modelu. Prowadzi to do zwiększenia zdolności generalizacji i zmniejszeniu ryzyka przeuczenia. Każde drzewo w lesie określa własną prognozę przynależności do określonej klasy dla danego przypadku w oparciu o własne kryteria i zestaw zmiennych. Klasyfikacja z największą liczbą głosów uważana jest za konsensus [46]. Bardziej szczegółowy opis tego modelu przedstawiony został w podpunkcie B podrozdziału 2.3.3.

D. *K*-najbliższych sąsiadów

K-najbliższych sąsiadów (ang. *k-Nearest Neighbors - kNN*) to algorytm klasyfikacji, który opiera się na wyznaczaniu odległości pomiędzy próbami w wielowymiarowej przestrzeni. Klasyfikacja próbki polega na znalezieniu *k* najbliższych sąsiadów danej próbki w przestrzeni cech na podstawie funkcji odległości. Najczęściej stosowaną funkcją odległości jest odległość euklidesowa, która opisana jest następującym równaniem:

$$d(x, y) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$$

gdzie, $d(x, y)$ to odległość euklidesowa między dwiema próbami x i y , N to liczba cech w próbkach, x_i to wartość i -tej cechy w próbce x , y_i to wartość i -tej cechy w próbce y .

Proces klasyfikacji rozpoczyna się od określenia liczby k sąsiadów, którzy zostaną wzięci pod uwagę podczas klasyfikacji. k jest liczbą nieparzystą w celu uniknięcia remisów. Następnie obliczane są odległości pomiędzy badanym obiektem, a pozostałymi próbami w zbiorze

treninguowym za pomocą funkcji odległości. Na ich podstawie wybierane są k próbki o najmniejszych odległościach od badanego obiektu. Finalnie przeprowadzane jest głosowanie większościowe pośród wybranych k sąsiadów w celu określenia przynależności badanej próbki do klasy. Klasa, która pojawia się najczęściej pośród wybranych k sąsiadów, jest przypisywana do badanego obiektu [11, 46].

2.4 Ocena wydajności modelu klasyfikatora

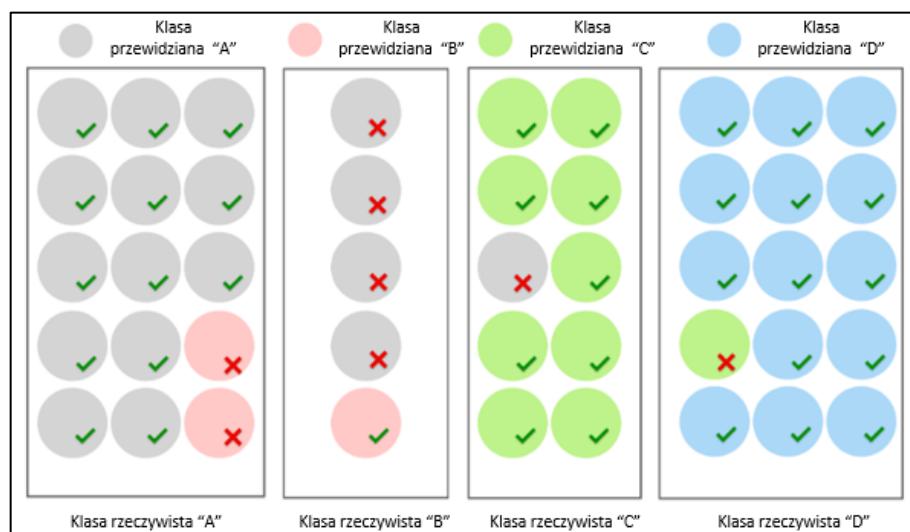
Etap validacji dla zaproponowanego modelu klasyfikacji polega na przeprowadzeniu oceny wydajności i skuteczności klasyfikatora dla badanego zbioru danych. Ocena klasyfikatora pomaga zrozumieć jak dobrze dany model radzi sobie z zadanym zbiorem. Przeprowadzana ona jest w oparciu o dane należące do grupy walidacyjnej.

Fundamentalnym narzędziem przy ocenie wydajności różnych algorytmów klasyfikacji jest macierz pomyłek. Macierz umożliwia zestawienie przewidywanych przez klasyfikator i rzeczywistych instancji klas zbioru. Stanowi ona podstawę definiowania szerokiego zakresu metryk wydajności badanych modeli (dokładność, precyza, czułość, miara F1), czy też technik graficznych obrazujących odsetek błędnych klasyfikacji dla danego punktu odcięcia (ROC). W przypadku klasyfikacji binarnej zastosowanie znajdują wszystkie dostępne metryki wydajności oraz techniki graficzne. Jednak w przypadku klasyfikacji wieloklasowej dostępny jest ograniczony zestaw metryk wydajności. W przypadku analizy ROC może ona znaleźć zastosowanie w problemach klasyfikacji wieloklasowej, jednakże złożoność analizy wzrastająca wraz z liczbą występujących klas w danym zbiorze sprawia, że staje się ona niepraktyczna [47]. Macierz pomyłek dla problemu klasyfikacji binarnej i wieloklasowej przedstawiono na rysunku 12. Każda kolumna macierzy reprezentuje instancje przewidywanej klasy, podczas gdy każdy wiersz reprezentuje instancję rzeczywistej klasy. Element macierzy pomyłek w wierszu i oraz kolumnie j reprezentuje liczbę przypadków, dla których przewidywaną klasą jest j , a rzeczywistą klasą jest i . W takim kontekście macierz pomyłek przedstawia, w jaki sposób model klasyfikacji ulega dezorientacji podczas dokonywania klasyfikacji. Oprócz tego, że macierz może zapewnić wgląd nie tylko w błędy popełniane przez klasyfikator, to informuje również o rodzajach powstających błędów. Dla problemu klasyfikacji binarnej elementy macierzy scharakteryzowane są na podstawie przewidywanej etykiety (pozytywna, negatywna) oraz wyniku porównania przewidywanej z rzeczywistą etykietą klasy (prawda, fałsz): prawdziwie pozytywne (TP), prawdziwie negatywne (TN), fałszywie pozytywne (FP) i fałszywie negatywne (FN).

A	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th colspan="2"></th> <th colspan="2">Klasa przewidziana</th> </tr> <tr> <th colspan="2"></th> <th>Pozytywna</th> <th>Negatywna</th> </tr> <tr> <th rowspan="2" style="text-align: center;">Klasa rzeczywista</th> <th>Pozytywna</th> <td style="text-align: center;">TP</td> <td style="text-align: center;">FN</td> </tr> </thead> <tbody> <tr> <th>Negatywna</th> <td style="text-align: center;">FP</td> <td style="text-align: center;">TN</td> </tr> </tbody> </table>						Klasa przewidziana				Pozytywna	Negatywna	Klasa rzeczywista	Pozytywna	TP	FN	Negatywna	FP	TN																			
		Klasa przewidziana																																				
		Pozytywna	Negatywna																																			
Klasa rzeczywista	Pozytywna	TP	FN																																			
	Negatywna	FP	TN																																			
B	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th colspan="2"></th> <th colspan="4">Klasa przewidziana</th> </tr> <tr> <th colspan="2"></th> <th style="background-color: #cccccc;">K_1</th> <th style="background-color: #cccccc;">K_2</th> <th style="background-color: #cccccc;">...</th> <th style="background-color: #cccccc;">K_N</th> </tr> </thead> <tbody> <tr> <th rowspan="4" style="text-align: center; vertical-align: middle;">Klasa rzeczywista</th> <th>K_1</th> <td style="background-color: #90EE90;">$K_{1,1}$</td> <td style="background-color: #90EE90;">FP</td> <td style="background-color: #90EE90;">...</td> <td style="background-color: #90EE90;">$K_{1,N}$</td> </tr> <tr> <th>K_2</th> <td style="background-color: #90EE90;">FN</td> <td style="background-color: #90EE90;">TP</td> <td style="background-color: #90EE90;">...</td> <td style="background-color: #90EE90;">FN</td> </tr> <tr> <th>...</th> <td style="background-color: #cccccc;">...</td> <td style="background-color: #cccccc;">...</td> <td style="background-color: #cccccc;">...</td> <td style="background-color: #cccccc;">...</td> </tr> <tr> <th>K_N</th> <td style="background-color: #90EE90;">$K_{N,1}$</td> <td style="background-color: #90EE90;">FP</td> <td style="background-color: #90EE90;">...</td> <td style="background-color: #90EE90;">$K_{N,N}$</td> </tr> </tbody> </table>							Klasa przewidziana						K_1	K_2	...	K_N	Klasa rzeczywista	K_1	$K_{1,1}$	FP	...	$K_{1,N}$	K_2	FN	TP	...	FN	K_N	$K_{N,1}$	FP	...	$K_{N,N}$
		Klasa przewidziana																																				
		K_1	K_2	...	K_N																																	
Klasa rzeczywista	K_1	$K_{1,1}$	FP	...	$K_{1,N}$																																	
	K_2	FN	TP	...	FN																																	
																																	
	K_N	$K_{N,1}$	FP	...	$K_{N,N}$																																	

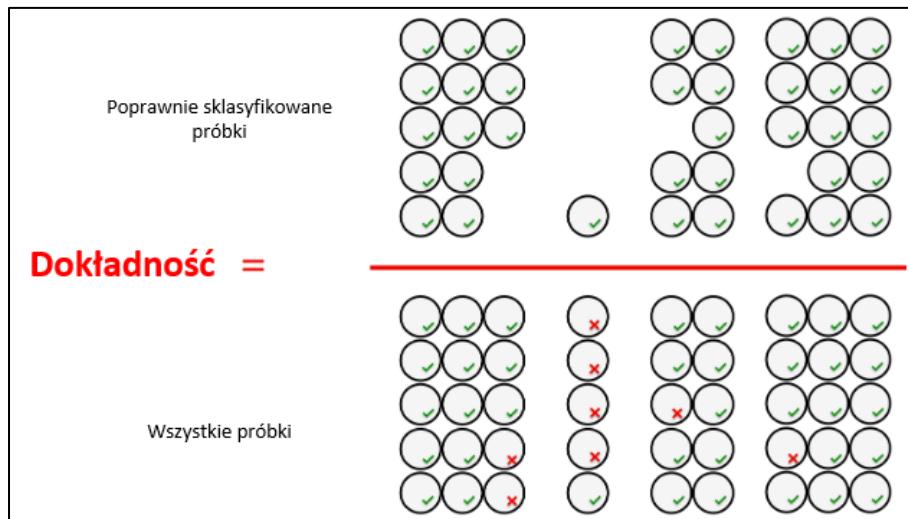
Rysunek 12. Macierz pomyłek dla problemu klasyfikacji binarnej (A) oraz klasyfikacji wieloklasowej (B) [47]

W przypadku wieloklasowych zbiorów najczęściej stosowanymi metrykami oceny są: dokładność, precyzaja, czułość oraz miara F1. W celu zilustrowania omawianych metod wydajności dla problemów wieloklasowych modeli klasyfikacyjnych na rysunku 13 przedstawiony został przykładowy podział wieloklasowego zbioru walidacyjnego. Rzeczywista przynależność danych próbek do określonych klas ograniczona została granicami prostokąta, natomiast kolorowymi kołami zaznaczono klasy przewidziane, stanowiący rezultat klasyfikacji zastosowanego na zbiorze modelu [48, 49].



Rysunek 13. Przykładowy podział czteroklasowego zbioru danych przedstawiający potencjalną przydział klas przewidzianych do klas rzeczywistych [48]

Dokładność klasyfikatora dla wieloklasowych zbiorów danych (rys. 14) to miara przedstawiająca stosunek wszystkich poprawnie sklasyfikowanych próbek do wszystkich próbek w zbiorze. Jest to miara bardzo prosta do zinterpretowania, stosowana do ogólnej oceny klasyfikatora. Jednak może być ona myląca w przypadku nierównoważnych zbiorów danych, w których liczba próbek jednej klasy dominuje nad innymi [49].

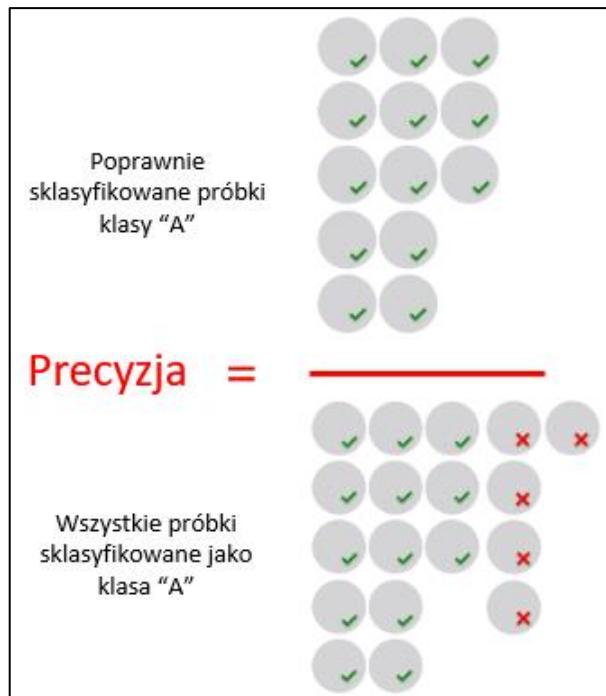


Rysunek 14. Dokładność klasyfikatora dla przykładowego, wieloklasowego zbioru [48]

Precyza klasyfikatora dla danej klasy (rys. 15) w wieloklasowym zbiorze danych to stosunek poprawnie sklasyfikowanych próbek określonej klasy do wszystkich próbek przewidzianych do tej klasy. Innymi słowy, precyza określa zdolność modelu do prawidłowego identyfikowania próbki określonej klasy [49]. Im mniej fałszywych trafień daje klasyfikator, tym większa jest jego precyza. Określana jest następującym równaniem:

$$Precyza = \frac{TP}{TP + FP}$$

gdzie, TP określa liczbę próbek poprawnie sklasyfikowanych do danej klasy, natomiast FP określa próbki błędnie sklasyfikowane jako należące do danej klasy, podczas gdy w rzeczywistości do niej nie należą.

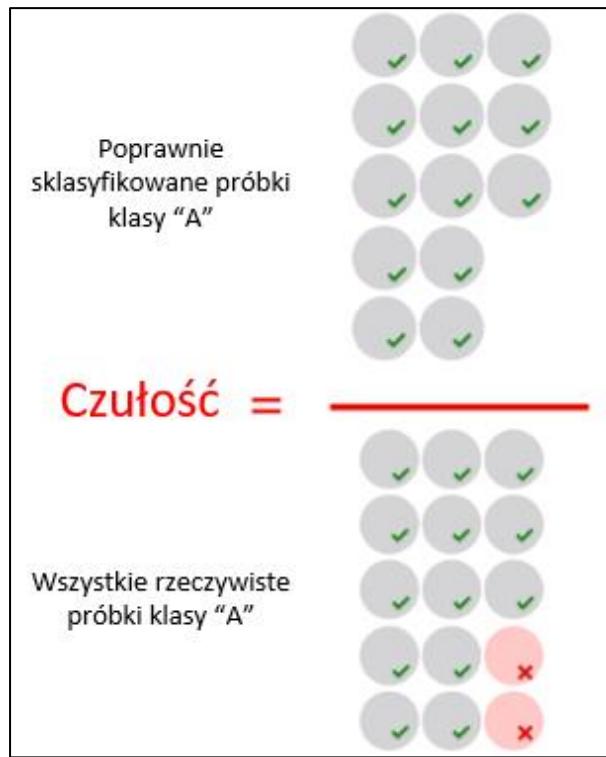


Rysunek 15. Precyza klasyfikatora dla danej klasy dla przykładowego, wieloklasowego zbioru [48]

Czułość klasyfikatora dla danej klasy (rys. 16) w wieloklasowym zbiorze danych określany jest przez stosunek poprawnie sklasyfikowanych próbek określonej klasy do wszystkich próbek obecnych w danej klasie, a więc czułość określa zdolność modelu do identyfikowania wszystkich próbek dla określonej klasy [49]. Wartość czułości jest określana w następujący sposób:

$$\text{Czułość} = \frac{TP}{TP + FN}$$

gdzie, FN określa próbki, które zostały błędnie sklasyfikowane jako nie należące do danej klasy, podczas gdy w rzeczywistości do niej należą. Im mniej jest wyników fałszywie negatywnych, tym większa jest czułość klasyfikatora.



Rysunek 16. Czułość klasyfikatora dla danej klasy dla przykładowego, wieloklasowego zbioru [48]

Zatem im wyższa jest precyzyja i czułość klasyfikatora, tym lepsze jest jego działanie, ponieważ wykrywa on większość próbek pozytywnych (wysoka czułość) i nie wykrywa wielu próbek, które nie powinny zostać wykryte (wysoka precyzyja). W celu ilościowego określenia tych miar, stosowana jest metryka nazywana miarą F1.

Miara F1 klasyfikatora określana jest jako średnia harmoniczna precyzyji i czułości klasyfikatora, przy czym wynik miary F1 osiąga najlepszą wartość dla 1, a najgorszy wynik dla 0. Im wyższa jest wartość precyzyji i czułości klasyfikatora, tym wyższa jest miara F1. Względny udział precyzyji oraz czułości w wyznaczaniu miary F1 jest równy [49]. Wzór na wyznaczanie wartości miary F1 jest następujący:

$$F1 = \frac{2 * (\text{Precyzyja} * \text{Czułość})}{\text{Precyzyja} + \text{Czułość}}$$

W celu oceny wydajności wieloklasowych zbiorów danych wartość precyzyji, czułości oraz miary F1 wyznaczane są w oparciu o uśrednienie otrzymanych wartości poszczególnych klas. Spośród różnych metod uśredniania wyróżnić można: makro i mikro uśrednianie oraz uśrednianie ważone.

2.4.1 Makro uśrednianie metryk klasyfikatora

Makro uśrednianie polega na obliczeniu precyzyji i czułości dla każdej klasy osobno, a następnie uśrednieniu tych wyników uwzględniając liczbę klas. Podejście to nadaje równą wagę każdej klasie obecnej w zbiorze, niezależnie od liczby należącej do niej próbek, co stanowi istotną właściwość w

przypadku zrównoważonych zbiorów danych, w których żadna z klas nie dominuje nad innymi pod względem liczebności [49, 50]. Wzory na makro uśrednianie precyzji, czułości oraz miary F1 dla wieloklasowego zbioru danych o N klasach są następujące:

$$Precyzja_M = \frac{1}{N} \sum_{i=1}^N Precyzja_i$$

$$Czułość_M = \frac{1}{N} \sum_{i=1}^N Czułość_i$$

$$F1_M = \frac{1}{N} \sum_{i=1}^N F1_i$$

2.4.2 Mikro uśrednianie metryk klasyfikatora

Mikro uśrednianie polega na obliczeniu ogólnej liczby TP, FP i FN oraz obliczeniu precyzji i czułości na podstawie tych wartości. Metoda skupia się na całkowitej poprawności klasyfikacji próbek we wszystkich klasach, ignorując różnice w rozmiarach klas. Mikro uśrednianie jest często stosowane w przypadku niezrównoważonych zbiorów danych, w których jedna lub kilka klas może mieć znacznie więcej próbek niż pozostałe klasy. Wzory na mikro uśrednianie precyzji, czułości oraz miary F1 są następujące:

$$Precyzja_\mu = \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N TP_i + \sum_{i=1}^N FP_i}$$

$$Czułość_\mu = \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N TP_i + \sum_{i=1}^N FN_i}$$

$$F1_\mu = \frac{2 * (Precyzja_\mu * Czułość_\mu)}{Precyzja_\mu + Czułość_\mu}$$

przypadku stosowania mikro uśredniania wszystkie przedstawione powyżej metryki będą takie same, ponieważ za każdym razem, gdy wystąpi w klasyfikacji wynik fałszywie dodatni to zawsze będzie również wynik fałszywie ujemny i odwrotnie. Jeśli przewidywana jest klasa „A”, a rzeczywistą klasą danej próbki jest „B” wówczas istnieje wynik FP dla klasy „A” oraz wynik FN dla klasy „B”. Jeśli przewidywanie dla danej próbki jest poprawne to istnieje tylko wynik TP, w związku z czym nie ma żadnego wyniku FP ani FN. Nie ma więc możliwości, żeby w przypadku zastosowania mikro uśredniania zwiększała się tylko liczba wyników FP lub FN, ale nie oba [51].

2.4.3 Ważone uśrednianie metryk klasyfikatora

Uśrednianie ważone polega na uwzględnianiu wag dla występujących w zbiorze danych klas. Waga dla każdej klasy jest wyznaczana na podstawie ilości jej wystąpienia w danym zbiorze. Metryki służące do oceny wydajności modelu wyznaczane są jako średnie ważone z poszczególnych klas. Innymi słowy każdej z występujących klas przypisywana jest rzetelna wartość W_N stanowiąca proporcję występujących próbek w klasie względem wszystkich próbek w zbiorze. Podejście to jest przydatne dla niezrównoważonych zbiorów danych, dla których podczas wyznaczania metryk przypisuje się większe znaczenie klasom o większej liczebności. Wzory na ważne uśrednianie precyzji, czułości oraz miary F1 dla wieloklasowego zbioru danych o N klasach są następujące:

$$Precyzja_W = \sum_{i=1}^N Precyzja_i * W_i$$

$$Czułość_W = \sum_{i=1}^N Czułość_i * W_i$$

$$F1_W = \sum_{i=1}^N F1_i * W_i$$

2.5 Podsumowanie przeglądu literaturowego

Przegląd literaturowy miał ma celu analizę dotychczasowego stanu wiedzy na temat selekcji najbardziej informacyjnych genów na podstawie mikromacierzowych zbiorów danych. W przeglądzie przedstawiona została istotność analizy ekspresji genów ludzkich w celu klasyfikacji występującej w organizmie choroby. Wobec tego przedstawione zostały mikromacierze ekspresyjne, proces ich wytwarzania, zasada działania oraz uzyskiwane na ich podstawie dane. Podkreślony został główny problem związany z ich wykorzystaniem, sprowadzający się do wielowymiarowości uzyskiwanych danych przy jednocześnie bardzo małej liczbie próbek. W związku z tym stwierdzono, że w celu usunięcia redundancji danych i polepszenia wydajności modeli klasystycznych, konieczne jest przeprowadzenie selekcji najbardziej informacyjnych genów pośród wszystkich dostępnych w zbiorze.

3 Zbiory danych

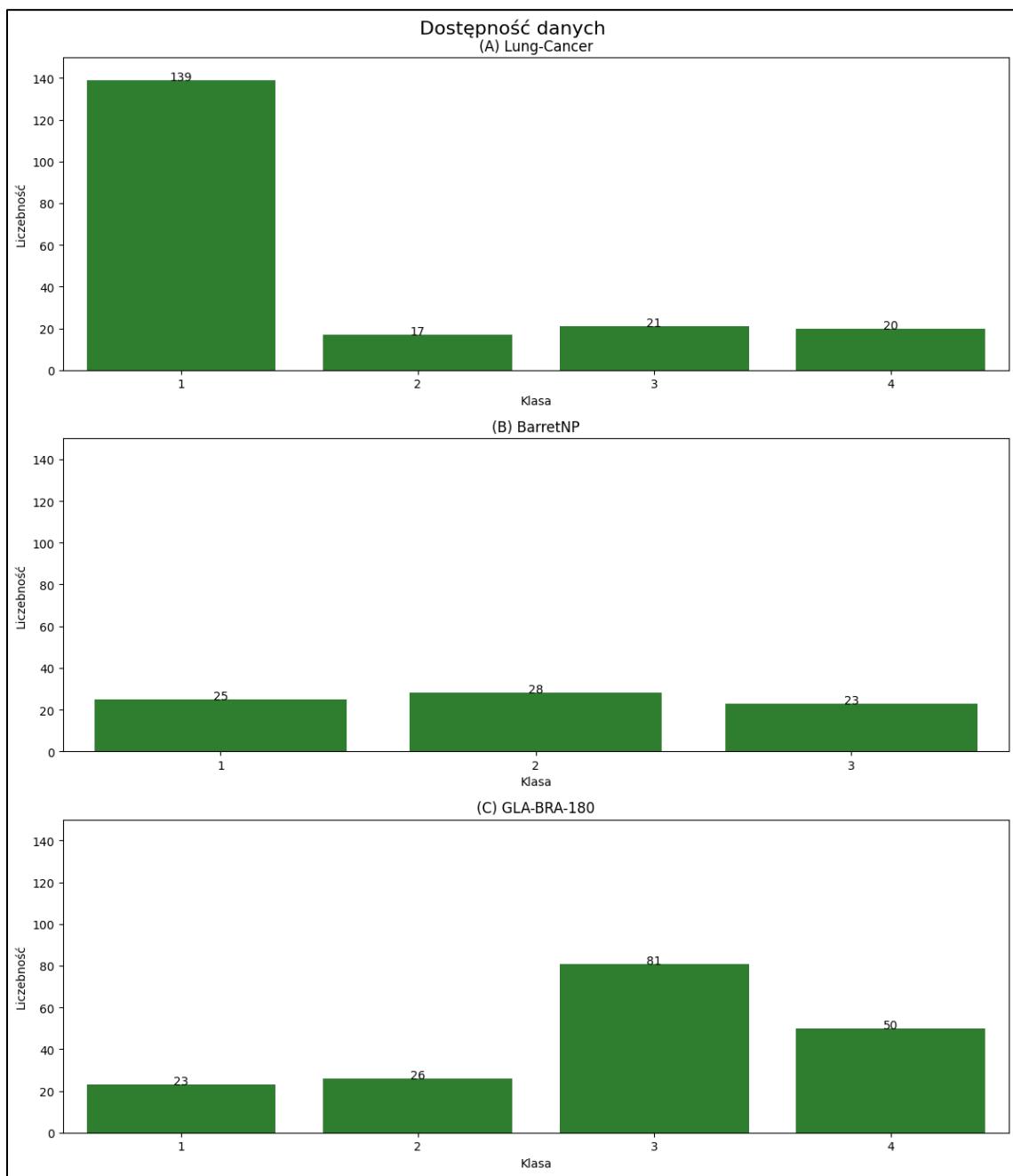
Badania przeprowadzone zostały na trzech ogólnodostępnych mikromacierzowych zbiorach danych. Obiekt zainteresowań sprowadził się wyłącznie do wieloklasowych zbiorów, ponieważ różnice między klasami w zbiorach binarnych byłyby zbyt jednoznaczne i nie stanowiłyby to problemu dla modelów klasyfikacyjnych.

W tabeli 2 przedstawione zostały szczegóły dotyczące wykorzystanych zbiorów danych. Dla każdego z nich uwzględnione zostały: nazwa zbioru, liczba próbek, liczba cech, liczba klas oraz dodatkowy opis zbioru.

Tabela 2. Zestawienie wykorzystanych zbiorów danych

Nazwa zbioru	Liczba próbek	Liczba cech	Liczba klas	Opis
<i>Lung-Cancer</i>	197	12600	4	Zbior zawierający pięć podtypów raka płuc
<i>BarretNP</i>	76	22277	3	Zbior zawierający trzy warianty choroby refluksowej przełyku
<i>GLA-BRA-180</i>	180	49151	4	Zbior zawierający cztery typy agniogenezy glejaków

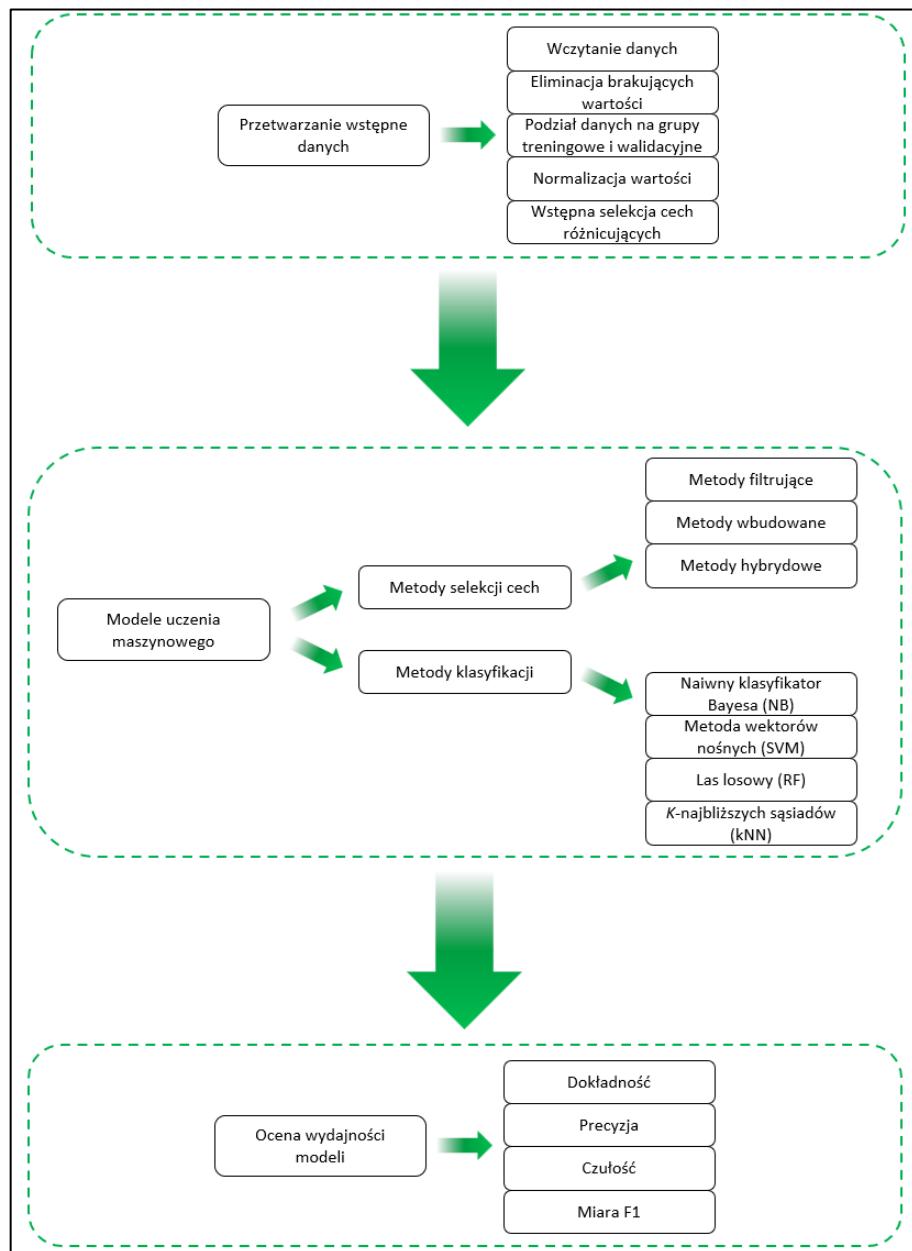
Na rysunku 17 przedstawiona została dostępność punktów dla poszczególnych klas w badanych zbiorach danych. Zbiory danych różnią się zarówno pod względem liczby próbek, jak i ilością klas. Zbior *Lung-Cancer* (rys. 16A) wyróżnia się największą liczbą próbek jak również największą liczbą klas. Najmniejszym zbiorem, mającym poniżej 100 próbek jest *BarretNP* (rys. 17B). W 2 z 3 wykorzystanych zbiorów widoczna jest znaczna dysproporcja pomiędzy występującymi klasami, co sprawia, że zbiory są zróżnicowane. W przypadku zbioru *Lung-Cancer* próbki należące do klasy 1 stanowią blisko 70% wszystkich próbek ze zbioru. Brak równowagi pomiędzy klasami dotyczy również zbioru *GLA-BRA-180* (rys. 17C), w którym dwie najbardziej popularne klasy stanowią ponad 70% wszystkich próbek. Jedynym zrównoważonym zbiorem jest *BarretNP*, w którym liczba próbek przypadających na poszczególne klasy zawiera od 23 do 28 przypadków.



Rysunek 17. Dostępność danych dla poszczególnych klas w badanych zbiorach danych

4 Metodyka badań

Na rysunku 18 przedstawiony został schemat metodyki badań. Zastosowany schemat można podzielić na 3 główne etapy. Pierwszy etap sprowadzony został do wstępnego przetwarzania danych. Drugi etap polegał na selekcji najbardziej informacyjnych cech ze zbiorów, a następnie klasyfikacji danych. Ostatnim etapem badań było przeprowadzenie oceny wydajności modelu wykorzystując metryki takie, jak: dokładność, precyzja, czułość oraz miara F1 modelu.



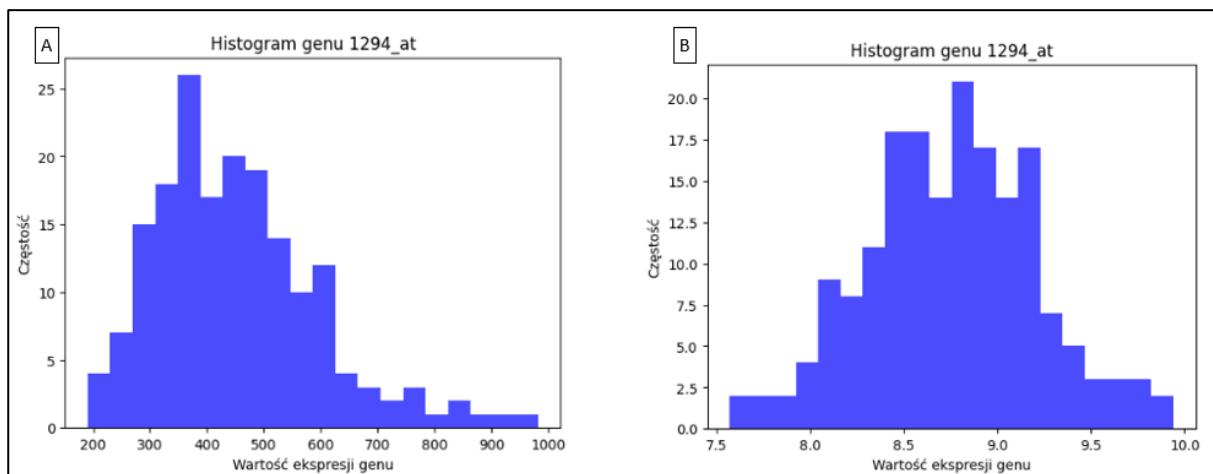
Rysunek 18. Schemat metodyki badań

4.1 Przetwarzanie wstępne danych

Dane mikromacierzowe uzyskane bezpośrednio z pomiarów oprócz tego, że mogą posiadać błędy techniczne w postaci brakujących wartości, to dodatkowo dane te mogą być przedstawiane w różnych skalach, co może powodować problemy w porównywaniu wyników pomiędzy różnymi zbiorami. Z tego powodu niezbędnym elementem analizy jest poddanie danych mikromacierzowych ich wstępнемu przetworzeniu. Etap ten składał się z następujących kroków: wczytania danych, eliminacji brakujących wartości w zbiorach, podzielenia zbiorów na grupy treningowe i walidacyjne oraz normalizacji wartości ekspresji genów. Dodatkowo zbiory poddane zostały wstępnej selekcji poprzez usunięcie cech nieistotnych z punktu widzenia analizy.

4.1.1 Wczytanie danych

Pierwszym krokiem badań było wczytanie zbiorów, na podstawie których zbudowane zostały modele. W przypadku niektórych zbiorów, których wartości ekspresji genów charakteryzowały się rozkładem odbiegającym od rozkładu normalnego, należało przeprowadzić transformację logarytmiczną [52]. Etap ten jest niezwykle istotnym elementem umożliwiającym późniejsze porównywanie różnych macierzy. Wobec tego wartości z badanych zbiorów przekształcone zostały za pomocą logarytmu binarnego. Na rysunku 19 zaprezentowany został histogram wybranego genu ze zbioru *GLA-BRA-180* przed przeprowadzeniem transformacji logarytmicznej (rys. 19A) i po transformacji (rys. 19B).



Rysunek 19. Histogram wybranego genu ze zbioru *GLA-BRA-180* przed przeprowadzeniem transformacji logarytmicznej (A) i po transformacji (B).

4.1.2 Eliminacja brakujących wartości

Kolejnym etapem była eliminacja brakujących wartości w zbiorach. Jest to kolejny, istotny element przetwarzania wstępnego danych, ponieważ brakujące wartości mogą znacząco wpływać na zależności pomiędzy badanymi genami. Problem ten można rozwiązać na wiele sposobów [53]:

A. Uzupełnianie średnią

Brakujące wartości uzupełniane są wartościami średnimi ekspresji genów dla określonego genu w oparciu o wartości ekspresji dla pozostałych próbek. Jest to prosta metoda, stosowana przy założeniu, że brakujące wartości są w przybliżeniu podobne do obserwowanych.

B. Uzupełnianie metodą „Hot Deck”

Metoda ta polega na identyfikacji k najbliższych sąsiadów próbki z brakującymi danymi na podstawie profilów ekspresji genów, korzystając z miary odległości euklidesowej. Brakujące wartości uzupełniane są poprzez obliczenie średniej wartości k najbliższych sąsiadów dla danego genu. Metoda ta zakłada, że próbki o podobnych profilach ekspresji mają podobne wartości ekspresji genów.

C. Uzupełnianie oparte na modelu

Metoda ta polega na zbudowaniu statystycznego modelu do oszacowania brakujących wartości, który tworzony jest na podstawie dostępnych wartości ekspresji genów. Wówczas wartości ekspresji innych genów są stosowane jako predyktory do zbudowania modelu regresji i przewidzenia brakujących wartości dla określonego genu.

D. Wielokrotne uzupełnianie

Metoda polegająca na wygenerowaniu kilku zestawów danych z uzupełnionymi danymi. Wówczas w każdym z tych zbiorów, każdą brakującą wartość zastępuje się różnymi uzupełnionymi wartościami na podstawie różnych procesów stochastycznych. Następnie każdy z tych zestawów jest analizowany oddziennie, a wyniki są łączone przy uwzględnieniu zmienności pomiędzy zestawami danych z uzupełnieniem.

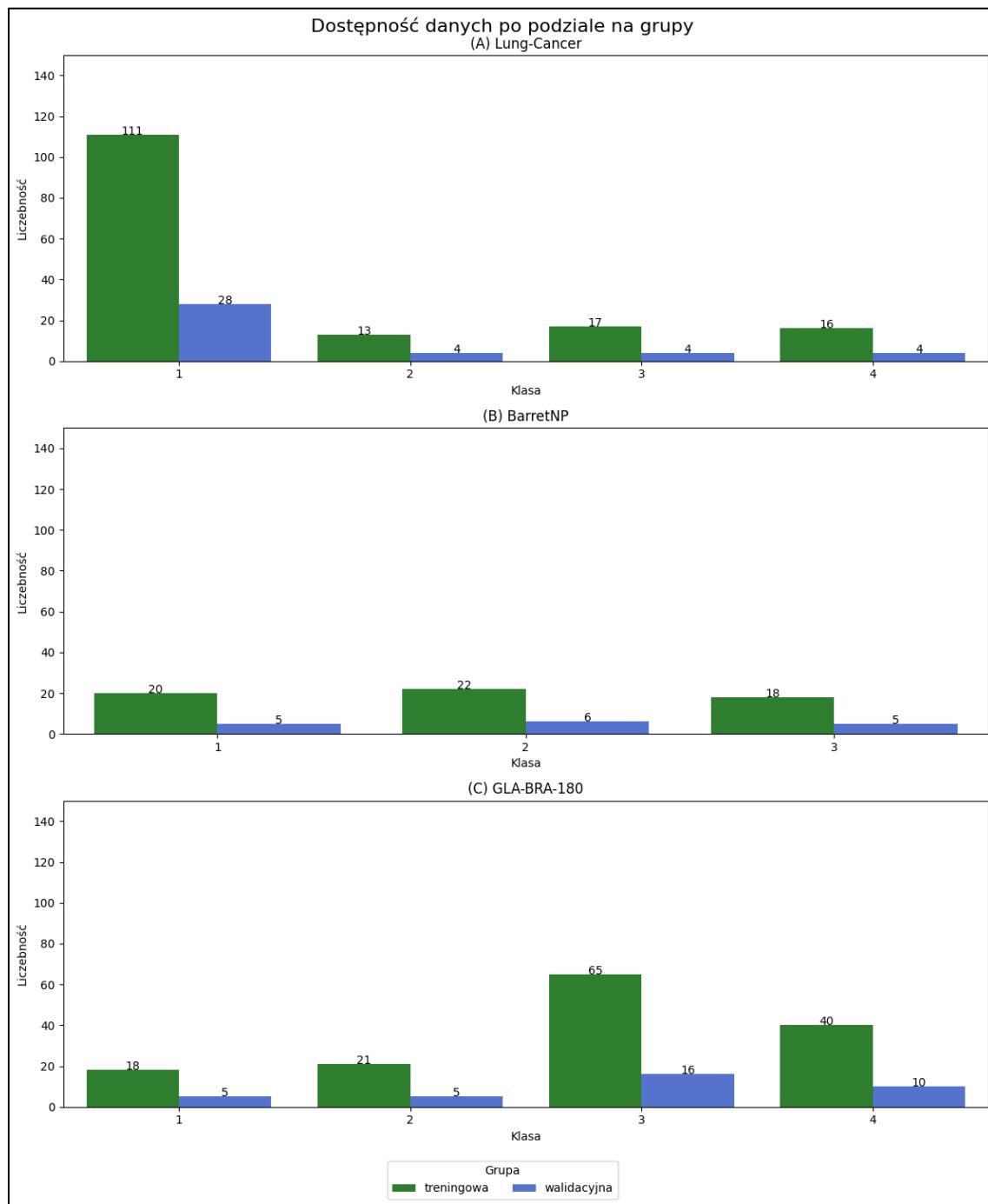
E. Uzupełnianie metodą „Cold Deck”

Jest to metoda, która polega na wykorzystywaniu zewnętrznych źródeł informacji, takich jak dane z innych podobnych badań lub źródła danych zewnętrznych, do oszacowania brakujących wartości w badanym zestawie danych.

W przypadku poniższej pracy brakujące wartości zastąpione zostały wartościami średnimi przypadającymi na daną cechę X_i .

4.1.3 Podział zbiorów na grupy treningowe i walidacyjne

Następnie dane zostały podzielone na zbiory treningowe oraz walidacyjne (rys. 20), w których dane treningowe stanowiły 80% zbioru. Dane treningowe wykorzystane zostały do trenowania modeli, natomiast dane walidacyjne służyły do oceny ich wydajności. Istotnym czynnikiem podczas podziału danych na grupy było zachowanie równowagi próbek ze wszystkich dostępnych klas w obu grupach.



Rysunek 20. Dostępność danych dla poszczególnych klas w badanych zbiorach danych po podziale na grupy treningowe i walidacyjne

4.1.4 Normalizacja min-max zbiorów

Kolejnym krokiem przetwarzania wstępniego danych było zastosowanie normalizacji min-max dla operowanych danych. Normalizacja min-max polega na sprowadzeniu wartości danej cechy X_i do przedziału [0; 1]. Dla wszystkich wartości cechy X_i wyznaczana jest nowa wartość w oparciu o następujące równanie:

$$X_i(j)' = \frac{X_i(j) - X_i^{\min}}{X_i^{\max} - X_i^{\min}}$$

gdzie X_i^{\min} reprezentuje minimalną wartość cechy X_i , a X_i^{\max} wartość maksymalną. Istotnym czynnikiem jest, aby do procesu normalizacji wartości minimalne i maksymalne określonej cechy wyznaczyć jedynie na podstawie danych treningowych. Następnie te same wartości X_i^{\min} oraz X_i^{\max} stosowane są do normalizacji zarówno danych treningowych, jak i danych walidacyjnych.

4.1.5 Wstępna selekcja cech różnicujących

Istotnym elementem przetwarzania wstępniego było usunięcie ze zbiorów cech, które były nieistotne z punktu widzenia analizy. W poniżej pracy wstępna selekcja polegała na identyfikacji i usunięciu cech ze zbiorów, które charakteryzowały się minimalną wariancją pośród wszystkich próbek. Cechy o niskiej wariancji są uważane za nieistotne, ponieważ zmiany w ich ekspresji są nieznaczne w porównaniu do innych genów. Wobec tego wszystkie geny, których wariancja znalazła się poniżej progu o wartości 0,01 zostały wykluczone ze zbioru i uznane za bezwartościowe. W tabeli 3 przedstawiona została liczba dostępnych cech badanych zbiorów danych przed i po wstępnej selekcji cech różnicujących.

Tabela 3. Zestawienie badanych zbiorów danych po wstępnej selekcji cech różnicujących

Nazwa zbioru	Liczba cech przed wstępnią selekcją	Liczba cech po wstępnej selekcji
<i>Lung-Cancer</i>	12600	12600
<i>BarretNP</i>	22277	14000
<i>GLA-BRA-180</i>	49151	49151

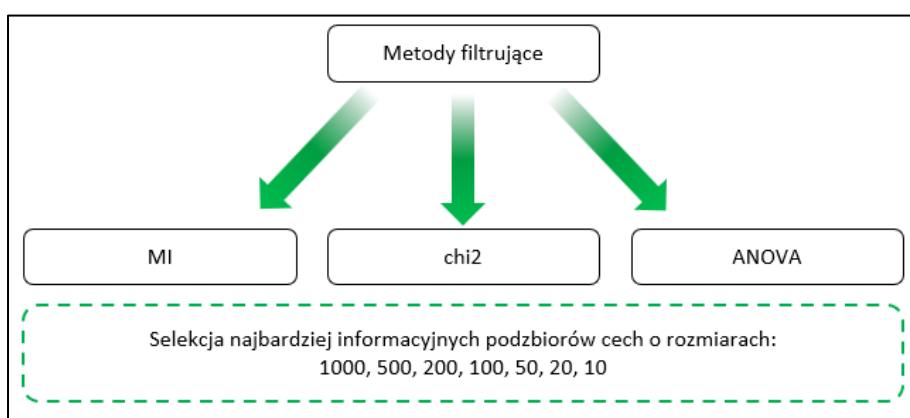
4.2 Metody selekcji cech

Problem przedstawiony w pracy sprowadza się do analizy porównawczej różnych metod selekcji najbardziej informacyjnych cech. Z tego powodu poddane analizie zostały następujące metody selekcji:

- metody filtrujące,
- metody wbudowane,
- metody hybrydowe, stanowiące połączenie metod filtrujących i wbudowanych z metodami opakowującymi.

4.2.1 Metody filtrujące

Zaimplementowane zostały 3 metody filtrujące: informacja wzajemna (MI), test niezależności chi-kwadrat (chi2) oraz analiza wariancji (ANOVA). Dla każdej z wymienionych metod filtrujących, dla wszystkich badanych zbiorów danych iteracyjnie wybrano kolejne podzbiory najbardziej informacyjnych cech o rozmiarach: 1000, 500, 200, 100, 50, 20, 10. Każdy kolejny podzbiór cech wybrany został na podstawie poprzedniego zbioru, w celu zminimalizowania redundancji danych. Wszystkie wyselekcjonowane podzbiory cech zapisane zostały do pliku CSV, w celu umożliwienia ich dalszego wykorzystania dla metod hybrydowych. Struktura takiego pliku CSV zawierała nazwy cech, odpowiadające im wartości oraz przynależność próbek do określonych klas. Na rysunku 21 przedstawiony został schemat metodyki badań dla metod filtrujących.

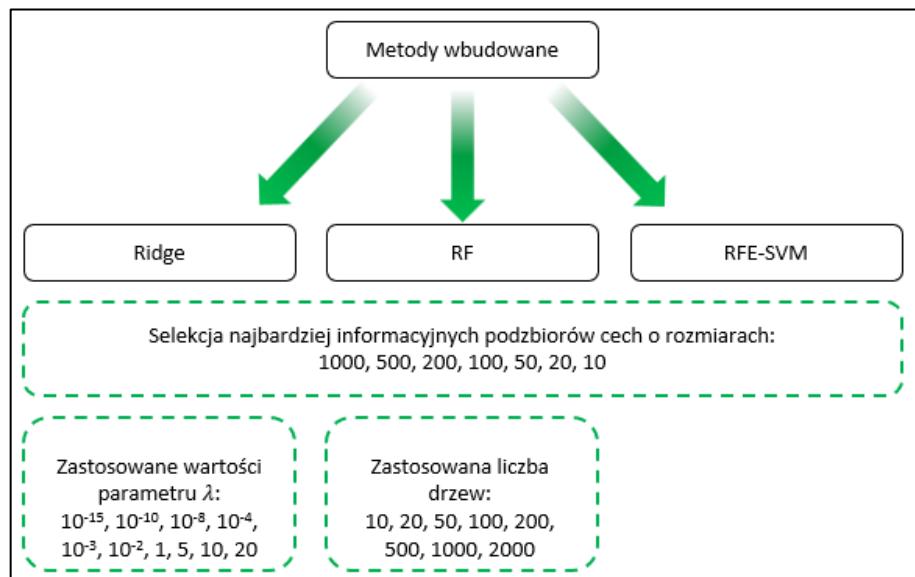


Rysunek 21. Schemat metodyki badań metod filtrujących

4.2.2 Metody wbudowane

W celu porównania metod wbudowanych selekcji cech zaimplementowano 3 wybrane metody: regresja grzbietowa (Ridge), las losowy (RF) oraz rekurencyjna eliminacja cech w oparciu o metodę wektorów nośnych (SVM-RFE). Dla każdej z wymienionych metod wybrano podzbiory o rozmiarze: 1000, 500, 200, 100, 50, 20, 10 najbardziej informacyjnych cech ze wszystkich badanych zbiorów danych. Każdy kolejny podzbiór cech wyselekcjonowany został na podstawie poprzedniego zbioru, w celu zminimalizowania redundancji danych. Identycznie jak w przypadku metod filtrujących wszystkie

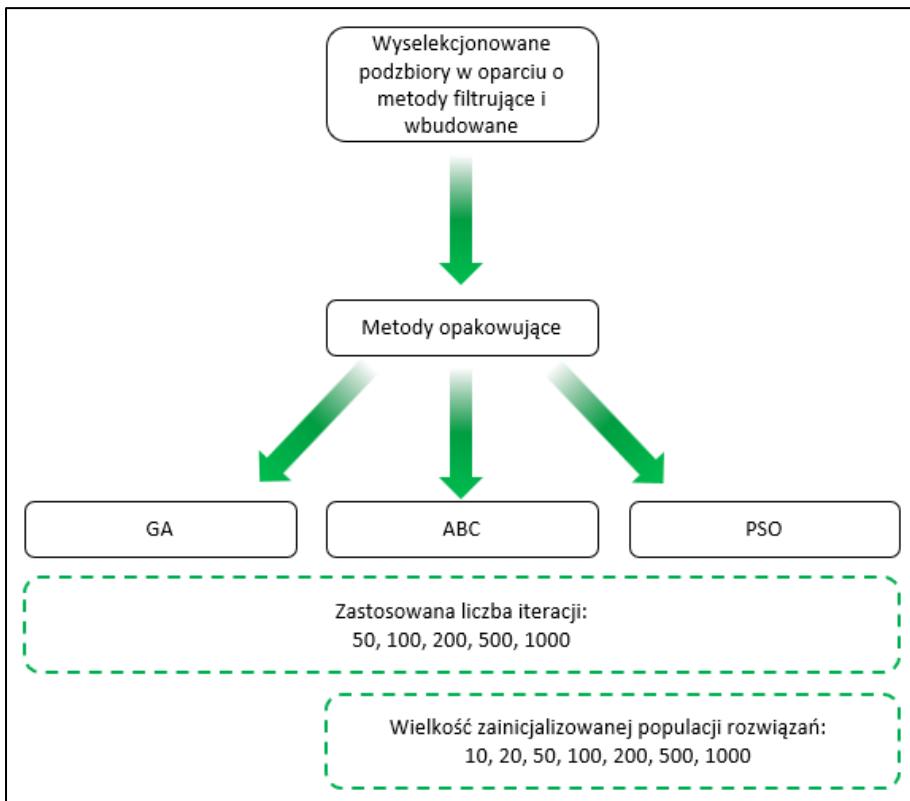
wyselekcjonowane podzbiory cech zapisane zostały do pliku CSV, w celu umożliwienia ich dalszego wykorzystania dla metod hybrydowych. Dodatkowo w przypadku metody Ridge zbadano wpływ wartości parametru regularyzacji λ (lambda), stosując następujące wartości parametru: $10^{-15}, 10^{-10}, 10^{-8}, 10^{-4}, 10^{-3}, 10^{-2}, 1, 5, 10, 20$. Natomiast w metodzie lasu losowego poddany analizie został wpływ liczby drzew w algorytmie, stosując wartości: $10, 20, 50, 100, 200, 500, 1000, 2000$. Schemat metodyki badań dla metod wbudowanych przedstawiony został na rysunku 22.



Rysunek 22. Schemat metodyki badań metod wbudowanych

4.2.3 Metody hybrydowe

Badania nad metodami hybrydowymi rozpoczęto od zestawienia wyselekcjonowanych wcześniej przy wykorzystaniu metod filtrujących i wbudowanych podzbiorów, które uzyskały najwyższe wartości wydajności dla każdego z badanych zbiorów danych. Wobec tego zestawiono po 3 podzbiory uzyskane za pomocą obu tych metod na pojedynczy zbiór danych, uzyskując łącznie 18 podzbiorów. Następnie na tych wyselekcjonowanych podzbiorach dokonano analizy porównawczej trzech metod opakowujących, takich jak: algorytm genetyczny (GA), algorytm sztucznej kolonii pszczół (ABC) oraz algorytm roju częstek (PSO). We wszystkich tych algorytmach funkcję dopasowania stanowił klasyfikator k -najbliższych sąsiadów (kNN). Dla każdej metody przeprowadzone zostały badania dla następującej liczby iteracji: 50, 100, 200, 500, 1000. Osiągnięcie predefiniowanej liczby iteracji stanowiło kryterium spełniające zatrzymanie algorytmu. Dodatkowo dla algorytmów ABC oraz PSO przeprowadzono badania względem optymalnej wielkości inicjalizowanej populacji rozwiązań. Przeprowadzone one zostały dla następujących wielkości: 10, 20, 50, 100, 200, 500, 1000. Schemat metodyki badań dla zastosowanych algorytmów opakowujących wykorzystanych w metodach hybrydowych przedstawiony został na rysunku 23.



Rysunek 23. Schemat metodyki badań metod hybrydowych

4.3 Metody klasyfikacji

Podzbiory najbardziej informacyjnych cech uzyskane za pomocą wszystkich wymienionych metod selekcji, poddane zostały klasyfikacji. Zastosowane zostały 4 metody klasyfikacji zbiorów: naiwny klasyfikator Bayesa (NB), maszyna wektorów nośnych (SVM), las losowy (RF) oraz k -najbliższych sąsiadów (kNN). Każda z zastosowanych metod była trenowana na zbiorze treningowym, a walidacja przebiegała na zbiorze walidacyjnym stanowiącym 20% próbek.

Ponieważ głównym celem tej pracy nie było optymalizowanie parametrów klasyfikatorów, lecz przeprowadzenie szczegółowej analizy metod selekcji najbardziej informacyjnych genów, wartości parametrów dla zastosowanych klasyfikatorów zostały starannie wybrane na podstawie przeglądu literatury. Dlatego dla klasyfikatora SVM, wartości parametrów gamma i C zostały ustalone na odpowiednio: 0,0001 oraz 1000 [54, 55]. Dodatkowo w przypadku tego klasyfikatora zastosowano podejście budowania klasyfikatora *One vs Rest*. W klasyfikatorze RF przypisano algorytmowi 128 drzew [56], natomiast dla klasyfikatora kNN zdefiniowano 5 sąsiadów [57], co w obu przypadkach powinno stanowić dobry kompromis pomiędzy wysoką wydajnością, a efektywnością obliczeniową.

4.4 Ocena wydajności modeli

Ze względu na wielowymiarowość danych oraz nierównowagę klas w badanych zbiorach zastosowana została metoda uśredniania ważonego metryk. Analiza wyników przeprowadzona została

w oparciu o metryki takie jak dokładność, precyza, czułość oraz miarę F1, zastosowaną ze względu na duże zróżnicowanie klas w badanych zbiorach danych. Jest to wskaźnik uwzględniający zarówno precyzję, czyli zdolność modelu do poprawnego identyfikowania pozytywnych przypadków, jak i czułość, czyli zdolność do wykrywania wszystkich istotnych przypadków. Dzięki temu miara F1 oferuje kompleksową ocenę wydajności modelu w sytuacjach, gdzie występuje istotna nierównowaga między klasami.

4.5 Środowisko

Badania przeprowadzone zostały w języku programowania *Python*. Wybór tego języka był podyktowany kilkoma istotnymi czynnikami, które znaczowo wpłynęły na efektywność oraz jakość opracowywanego rozwiązania.

Python to interpretowany język programowania wysokiego poziomu, który zdobył dużą popularność w dziedzinie informatyki i analizy danych. Jego główną ideą jest czytelność i klarowność kodu źródłowego, co ułatwia rozumienie oraz utrzymywanie projektów programistycznych. Składnia języka jest prosta i przejrzysta, co sprzyja sprawnemu pisaniu kodu. Python oferuje rozbudowany pakiet bibliotek, co sprawia, że jest doskonałym narzędziem do implementacji systemów z zakresu sztucznej inteligencji, a w szczególności algorytmów uczenia maszynowego.

W ramach tej pracy wykorzystana została popularna biblioteka *scikit-learn*, która dostarcza gotowe implementacje algorytmów związanych z przetwarzaniem danych oraz uczenia maszynowego. Do prezentacji uzyskanych rezultatów wykorzystane zostały biblioteki *matplotlib* oraz *seaborn*. Są to narzędzia umożliwiające tworzenie różnorodnych wykresów, co pozwoliło na efektywne zaprezentowanie wyników oraz ich interpretację.

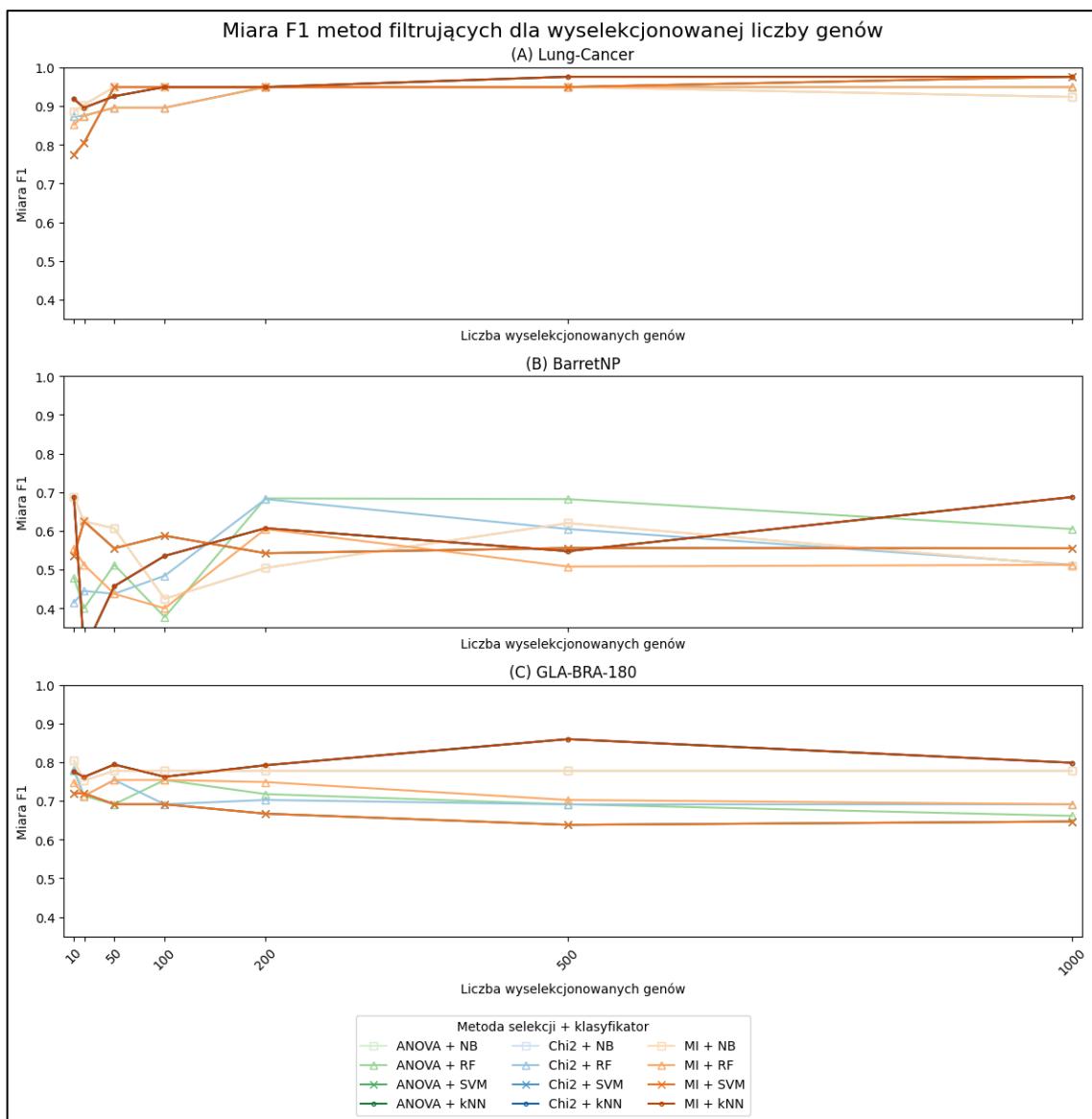
5 Wyniki badań i ich interpretacja

W niniejszym rozdziale pracy skoncentrowano się na prezentacji uzyskanych rezultatów i ich interpretacji. Podzielono go na podrozdziały, omawiające kolejno wyniki metod filtrujących, wbudowanych i hybrydowych.

5.1 Metody filtrujące

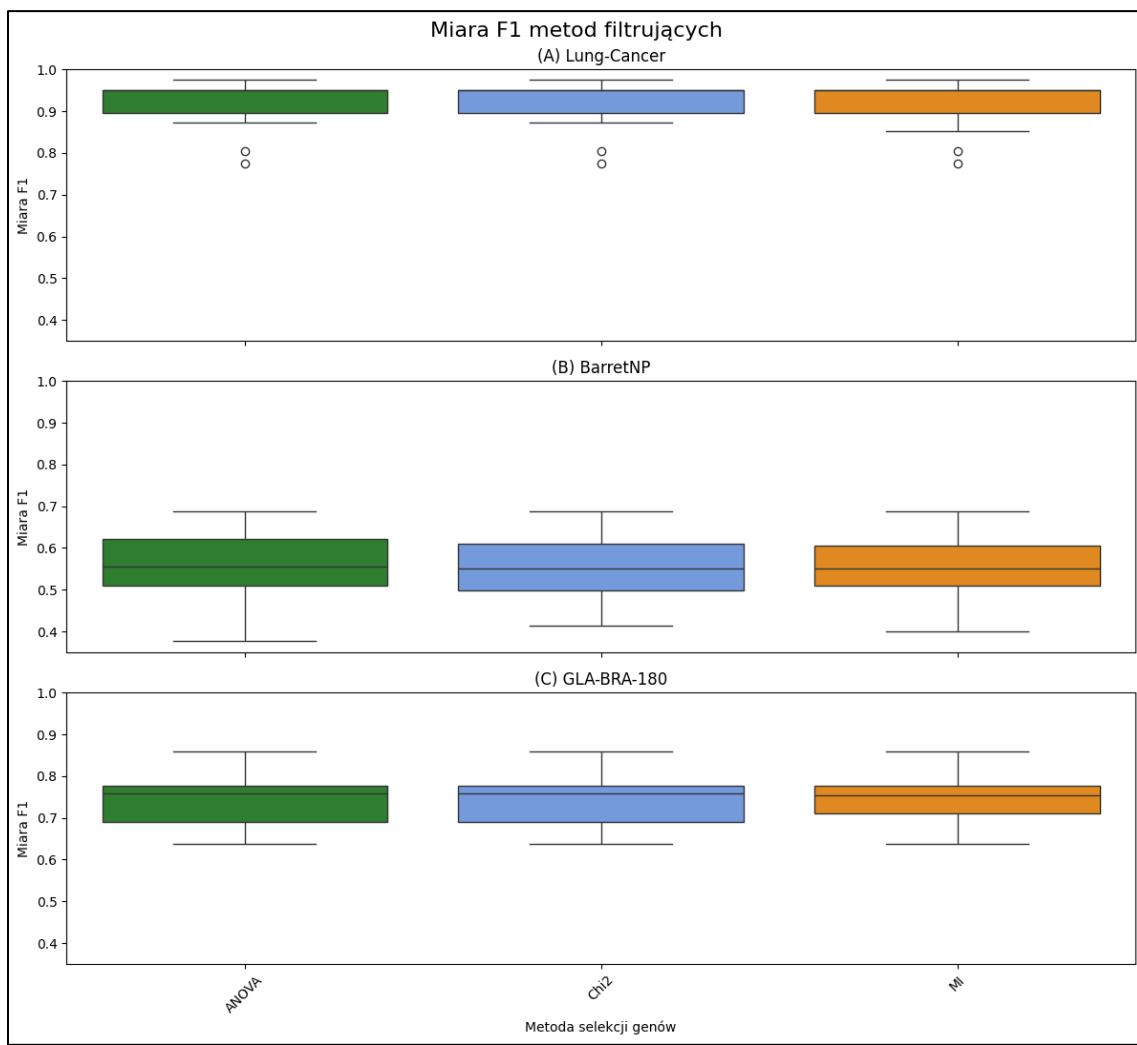
W celu szczegółowego omówienia wyników uzyskanych dla metod filtrujących na podstawie badanych zbiorów danych charakteryzujących się stosunkowo dużym zróżnicowaniem klas, wykorzystana została miara F1. Rezultaty miary F1 metod filtrujących względem określonej liczby wyselekcjonowanych genów zaprezentowane zostały na rysunku 24. Uzyskane wyniki pozostałych metryk: dokładności, precyzji oraz czułości dla metod filtrujących zamieszczone zostały w załączniku nr 1.

Uzyskane wartości miary F1 dla zbioru *Lung-Cancer* (rys. 24A) wskazują, że wszystkie zastosowane modele osiągają nieco niższe wartości dla najmniejszych podzbiorów danych, a wraz ze wzrostem wielkości tych podzbiorów, ich wydajność wzrasta. Wartości miary F1 dla wszystkich badanych modeli ulegają stabilizacji dla podzbioru składającego się z 200 genów. Zauważać można, że modele wykorzystujące klasyfikator kNN, już dla najmniejszego podzbioru składającego się z 10 genów osiągają ponad 90%. Dodatkowo najwyższa wartość dla tego zbioru uzyskana została dla wszystkich modeli wykorzystujących ten sam klasyfikator kNN osiągając 98% dla grupy 500 genów o najwyższej informacyjności. W przypadku zbioru *BarretNP* (rys. 24B) rezultaty miary F1 są znacznie niższe od wartości uzyskanych dla pozostałych zbiorów. Dokładnie taka sama wartość, stanowiąca wartość maksymalną wynoszącą blisko 70%, została uzyskana dla tego zbioru trzykrotnie: 2 razy dla modeli wykorzystujących klasyfikator kNN (10 i 1000 genów) oraz raz dla modelu wykorzystującego klasyfikator NB (10 genów). Wyniki miary F1 dla zbioru *GLA-BRA-180* (rys. 24C) dla wszystkich zastosowanych modeli osiągają stosunkowo wysokie wartości nawet dla bardzo małych podzbiorów. Maksymalną wartość osiągającą modele wykorzystujące klasyfikator kNN, dla podzbiorów składających się z 500 najbardziej informacyjnych genów osiągając niemal 90% wydajności. Rezultaty dla pozostałych modeli dla tego zbioru zawierają się w przedziale od 80% do 65%.



Rysunek 24. Miara F1 metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora

Uzyskane rezultaty miary F1 metod filtrujących bez rozróżnienia na rodzaj zastosowanego klasyfikatora oraz liczbę wyselekcjonowanych cech przedstawione zostały na rysunku 25. Zgodnie z zaprezentowanymi rezultatami miary F1 wyniki wydajności pomiędzy zastosowanymi metodami filtrującymi dla wszystkich badanych zbiorów danych są identyczne. Oznacza to, że w przypadku analizowanego wieloklasowego problemu klasyfikacji zastosowane metody filtrujące nie różnią się pod względem wydajności, co potwierdzają niemalże takie same rezultaty. Uzyskane wyniki pozostałych metryk: dokładności, precyzyji oraz czułości dla metod filtrujących zamieszczone zostały w załączniku nr 1.



Rysunek 25. Miara F1 badanych metod filtrujących

Dodatkowo na podstawie przeprowadzonej analizy metod filtrujących zauważać można, że zbiór *Lung-Cancer* (rys. 25A) niezależnie od zastosowanego modelu charakteryzuje się bardzo wysokimi wynikami wydajności nawet dla najmniejszych podzbiorów najbardziej informacyjnych genów, co wskazuje na to, że jest to zbiór łatwy do klasyfikacji. Z drugiej strony, najniższymi wynikami spośród wszystkich badanych zbiorów charakteryzuje się zbiór *BarretNP* (rys. 25B), co może sugerować dużą złożoność tego zbioru.

Celem analizy wyników dla metod filtrujących było znalezienie podzbioru dla każdego z badanych zbiorów danych, który spośród zastosowanych modeli osiągnął najwyższą wydajność. Wobec tego w tabeli 4 przedstawiono parametry, dla których uzyskano najwyższe wyniki wydajności dla każdej zastosowanej metody filtrującej i zestawu danych, uwzględniając wielkość wyselekcjonowanego podzbioru oraz wykorzystany klasyfikator. Warto zaznaczyć, że dla zbioru *BarretNP* zdecydowano się wskazać na podzbiory charakteryzujące się maksymalną wartością wydajności o największej ilości cech, ponieważ uznano te zbiorы za lepsze do klasyfikacji dla metod hybrydowych. Zestawione w tabeli podzbiory wykorzystane zostały do przeprowadzenia badań dla metod hybrydowych.

Tabela 4. Wyselekcjonowane podzbiory za pomocą metod filtrujących charakteryzujące się najwyższą wydajnością

Metoda	Zbiór danych	Liczba wyselekcjonowanych genów	Klasyfikator uzyskujący najlepsze wyniki	Uzyskana wartość miary F1
ANOVA	<i>Lung-Cancer</i>	500	kNN	0,98
ANOVA	<i>BarretNP</i>	1000	kNN	0,69
ANOVA	<i>GLA-BRA-180</i>	500	kNN	0,86
Chi2	<i>Lung-Cancer</i>	500	kNN	0,98
Chi2	<i>BarretNP</i>	1000	kNN	0,69
Chi2	<i>GLA-BRA-180</i>	500	kNN	0,86
MI	<i>Lung-Cancer</i>	500	kNN	0,98
MI	<i>BarretNP</i>	1000	kNN	0,69
MI	<i>GLA-BRA-180</i>	500	kNN	0,86

5.2 Metody wbudowane

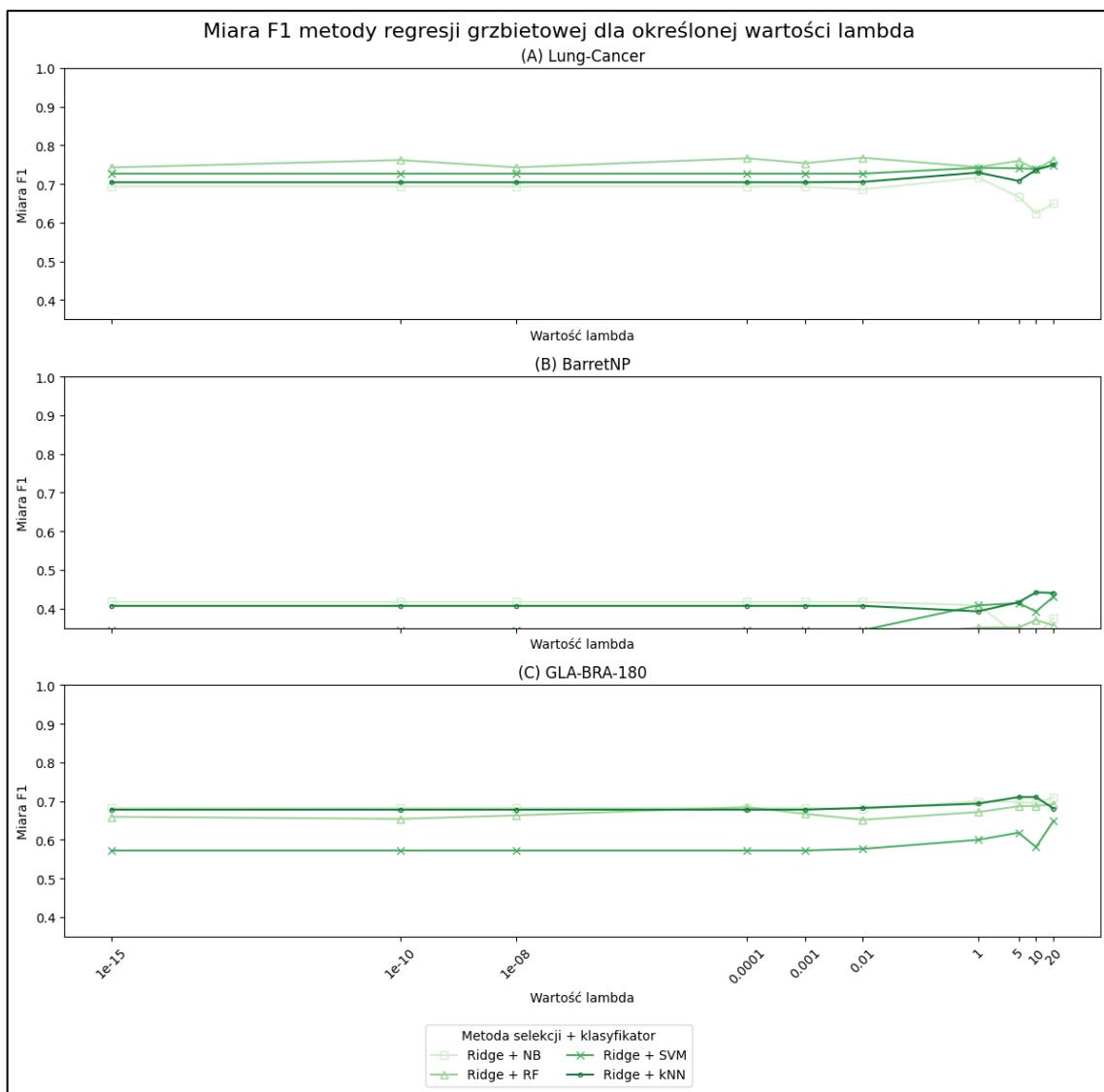
Kolejnym krokiem analizy było zbadanie wydajności metod selekcji cech dla metod wbudowanych. Model wykorzystujący metodę Ridge sprawdzony został pod kątem zależności wartości lambda na wydajność modeli. Wartość metryki dla danej wartości lambda stanowiła uśredniony wynik wszystkich badanych podzbiorów przypadających dla określonej wartości. W przypadku analizy metody selekcji RF zbadana została zależność liczby drzew na wydajność modelu, również stanowiąca uśredniony wynik wszystkich podzbiorów dla określonej liczby. Następnie dla wszystkich badanych metod, włącznie z metodą SVM-RFE, poddany analizie został wpływ wielkości wyselekcjonowanych podzbiorów na wydajność modelu. W celu znalezienia podzbioru dla każdego z badanych zbiorów danych, który spośród zastosowanych modeli osiąga najwyższą wydajność, zestawiono parametry, dla których uzyskano najwyższe wyniki wydajności dla każdej zastosowanej metody i zestawu danych.

Identycznie, jak w przypadku metod filtrujących, w celu szczegółowego omówienia wyników uzyskanych dla metod wbudowanych na podstawie badanych zbiorów danych charakteryzujących się stosunkowo dużym zróżnicowaniem klas, wykorzystana została miara F1. Graficznie przedstawione rezultaty pozostałych metryk dla metod wbudowanych zamieszczone zostały w załączniku nr 2.

5.2.1 Regresja grzbietowa

Modele wykorzystujące metodę Ridge do określenia najbardziej informacyjnych genów ze zbiorów zostały sprawdzone dla następujących wartości lambda: 10^{-15} , 10^{-10} , 10^{-8} , 10^{-4} , 10^{-3} , 10^{-2} , 1, 5, 10, 20. Uśrednione rezultaty miary F1 metod filtrujących względem tych wartości zaprezentowane zostały na rysunku 26.

Na podstawie uzyskanych rezultatów można stwierdzić, że uzyskane wykresy zależności wartości lambda od wydajności modeli charakteryzują się wypłaszczonym przebiegiem. To sugeruje, że zastosowane wartości lambda dla badanych zbiorów danych nie miały istotnego wpływu na osiągane rezultaty. W przypadku zbioru *Lung-Cancer* (rys. 26A) rezultaty różnych modeli są ze sobą bardzo zbliżone. Warto zauważyć, że dla tego zbioru rodzaj zastosowanego klasyfikatora ma istotniejszy wpływ na wydajność niż konkretne wartości lambda. W przypadku zbiorów *BarretNP* (rys. 26B) oraz *GLA-BRA-180* (rys. 26C) można dostrzec niewielki wzrost uśrednionych wartości wydajności modeli dla wyższych wartości lambda. Zgodnie z teorią regresji grzbietowej, złożoność modelu rośnie wraz ze wzrostem wartości lambda, co prowadzi do ograniczania nadmiernego dopasowania modelu [32]. Uzyskane rezultaty mogą sugerować, że te dwa zbiory są złożone.

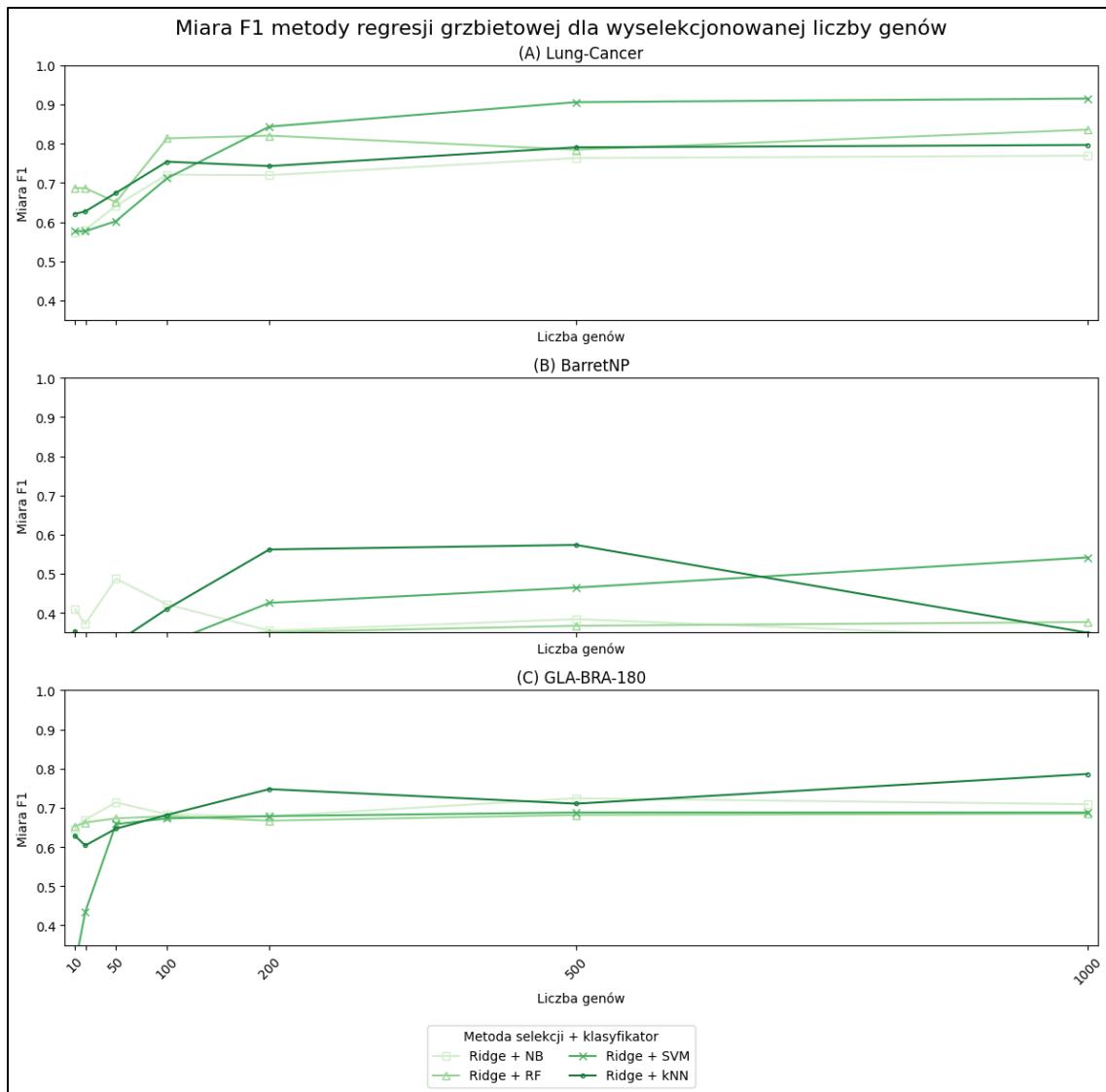


Rysunek 26. Średnia miara F1 metody Ridge dla określonej wartości lambda oraz zastosowanego klasyfikatora

Następnie poddane analizie zostały uśrednione rezultaty wydajności metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów wszystkich badanych zbiorów danych, które zaprezentowane zostały na rysunku 27. Wyniki wskazują na istotną korelację między liczbą wyselekcjonowanych genów, a wydajnością modelu dla wszystkich badanych zbiorów danych.

Dla zbioru *Lung-Cancer* (rys. 27A), zauważono, że dla metody regresji grzbietowej wzrost liczby wyselekcjonowanych genów prowadzi do znacznego wzrostu wydajności. Osiągnięcie ponad 90% miary F1 dla 500 najbardziej informacyjnych genów sugeruje, że dla tego zbioru większa liczba genów przyczynia się do lepszej klasyfikacji. W przypadku zbioru *BarretNP* (rys. 27B) maksymalna uśredniona wydajność przypada na 500 genów. Natomiast dla zbioru *GLA-BRA-180* (rys. 27C)

maksymalna uśredniona wartość miary F1 przypada na 1000 genów. Dla tych dwóch trudnych zbiorów danych największa uśredniona wartość wydajności została osiągnięta dla klasyfikatora kNN.



Rysunek 27. Średnia miara F1 metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora

W tabeli 5 przedstawiono parametry, dla których uzyskano najwyższe wyniki wydajności regresji grzbietowej dla każdego zestawu danych, uwzględniając wielkość wyselekcjonowanego podzbioru, wartość lambda oraz wykorzystany klasyfikator. Wszystkie zestawione w tabeli podzbiory wykorzystane zostały do przeprowadzenia badań dla metod hybrydowych.

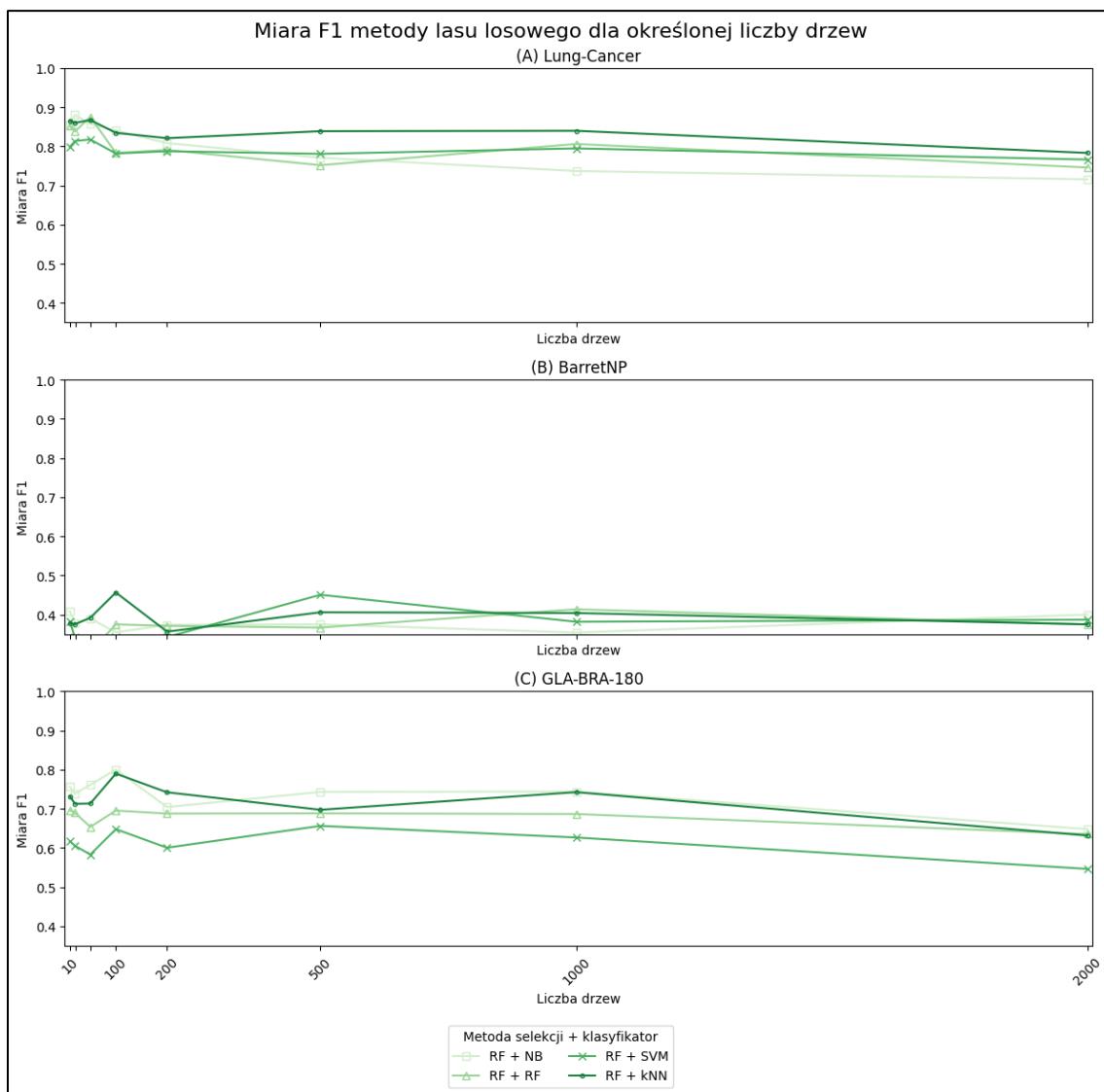
Tabela 5. Wyselekcjonowane podzbiory za pomocą metody Ridge charakteryzujące się najwyższą wydajnością

Metoda	Zbiór danych	Liczba wyselekcjonowanych genów	Wartość lambda	Klasyfikator uzyskujący najlepsze wyniki	Uzyskana wartość miary F1
Ridge	<i>Lung-Cancer</i>	500	10^{-2}	SVM	0,95
Ridge	<i>BarretNP</i>	500	10	kNN	0,64
Ridge	<i>GLA-BRA-180</i>	1000	5	kNN	0,79

5.2.2 Las losowy

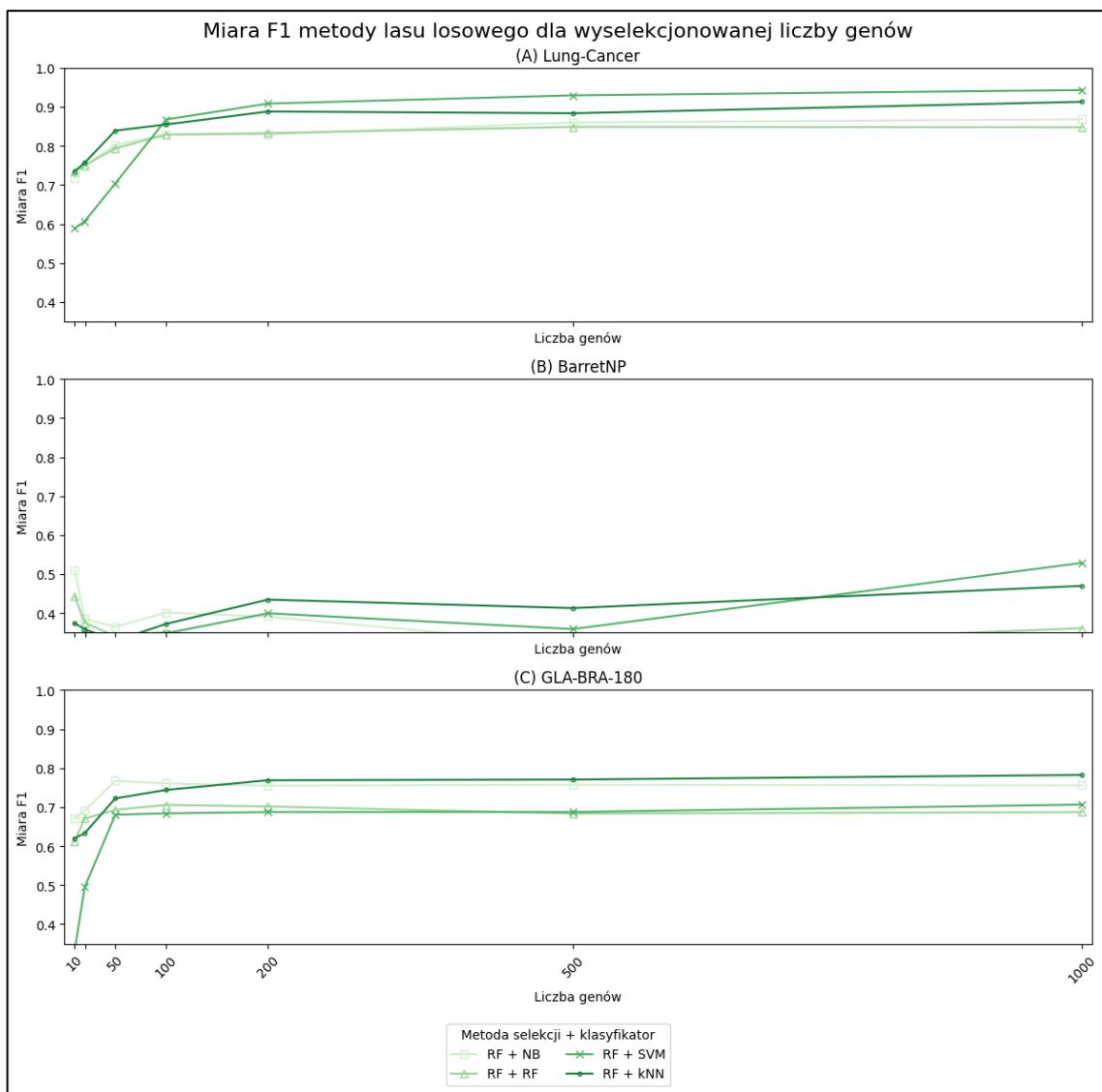
Wyniki średnich wartości miary F1 dla modeli wykorzystujących metodę selekcji RF do określenia najbardziej informacyjnych genów ze zbiorów zostały sprawdzone dla następujących populacji drzew: 10, 20, 50, 100, 200, 500, 1000, 2000. Uzyskane wartości przedstawione zostały na rysunku 28.

Dla zbioru *Lung-Cancer* (rys. 28A) zauważać można, że modele osiągają nieco wyższą wydajność dla mniejszej liczby drzew, przy czym najniższa średnia wydajność jest uzyskiwana dla największej analizowanej liczby drzew. W przypadku zbioru *BarretNP* (rys. 28B) nie można stwierdzić zależności wpływu ilości drzew na wydajność modeli, jednak warto zaznaczyć, że dla 100 oraz 500 drzew uśrednione wyniki wydajności są nieco wyższe od pozostałych. Dla zbioru *GLA-BRA-180* (rys. 28C) istnieje duża rozbieżność w wydajności pomiędzy różnymi modelami. W tym przypadku istotniejszy na rezultat modeli jest wpływ zastosowanego klasyfikatora niż ilość drzew w algorytmie.



Rysunek 28. Średnia miara F1 metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora

Tak jak w przypadku poprzedniej metody, kolejnym krokiem było sprawdzenie średnich wyników wydajności metody lasu losowego dla określonej liczby wyselekcjonowanych genów badanych zbiorów danych. Uśrednione rezultaty miary F1 przedstawione zostały na rysunku 29. Wyniki wskazują, że wraz ze wzrostem liczby wyselekcjonowanych genów dla wszystkich zbiorów, wydajność modelu rośnie, osiągając dla metody lasu losowego maksymalne wartości dla 1000 genów.



Rysunek 29. Średnia miara F1 metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora

Warto zauważyć, że model wykorzystujący las losowy zarówno jako metodę selekcji cech, jak i klasyfikator, nie osiąga najwyższych wyników. Może to sugerować, że inne kombinacje metod mogą być bardziej efektywne dla tych konkretnych zbiorów danych.

W tabeli 6 zaprezentowano parametry, dla których uzyskano najwyższe wyniki wydajności lasu losowego dla każdego zestawu danych, uwzględniając wielkość wyselekcjonowanego podzbioru, zastosowaną populację drzew oraz wykorzystany klasyfikator. Wszystkie zestawione w tabeli podzbiory wykorzystane zostały do przeprowadzenia badań dla metod hybrydowych.

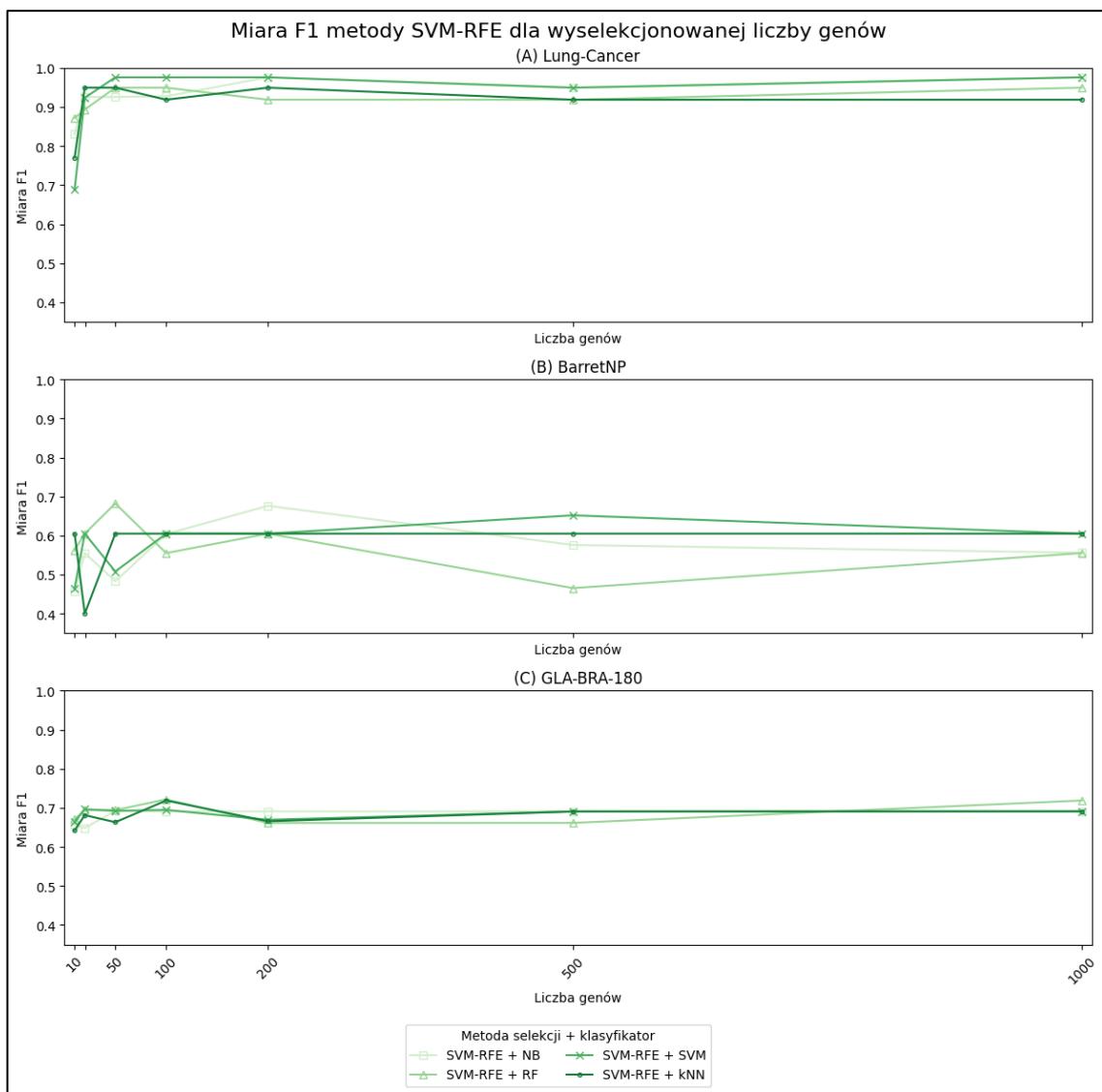
Tabela 6. Wyselekcjonowane podzbiory za pomocą metody lasu losowego charakteryzujące się najwyższą wydajnością

Metoda	Zbiór danych	Liczba wyselekcjonowanych genów	Liczba drzew	Klasyfikator uzyskujący najlepsze wyniki	Uzyskana wartość miary F1
RF	<i>Lung-Cancer</i>	500	1000	SVM	0,95
RF	<i>BarretNP</i>	200	500	SVM	0,70
RF	<i>GLA-BRA-180</i>	500	50	kNN	0,86

5.2.3 SVM-RFE

Wyniki miary F1 metody SVM-RFE dla określonej liczby wyselekcjonowanych genów wszystkich badanych zbiorów danych zaprezentowane zostały na rysunku 30. Uzyskane rezultaty dla modeli wykorzystujących metodę selekcji SVM-RFE charakteryzują się stosunkowo wysokimi wartościami dla wszystkich badanych zbiorów danych. W porównaniu do pozostałych wbudowanych metod selekcji wyniki wydajności są znacznie wyższe dla najmniejszych podzbiorów.

Dla zbioru *Lung-Cancer* (rys. 30A), maksymalna wartość wydajności już dla 50 najbardziej informacyjnych genów osiąga niemal 98% skuteczności przy wykorzystaniu klasyfikatora SVM. Sugeruje to, że dla tego stosunkowo prostego do klasyfikacji zbioru, metoda SVM-RFE jest w stanie znacznie zredukować liczbę genów, jednocześnie utrzymując wysoką skuteczność klasyfikacji. W przypadku zbioru *BarretNP* (rys. 30B), maksymalna wartość wydajności, wynosząca niemal 70%, osiągnięta zostaje również dla 50 genów przy wykorzystaniu klasyfikatora RF. Sugeruje to, że dla tego trudnego zbioru możliwe jest uzyskanie efektywnej selekcji cech przy zastosowaniu tej metody. Dla zbioru *GLA-BRA-180* (rys. 30C) maksymalna wartość wydajności osiągnięta zostaje dla 100 genów, również dla klasyfikatora RF. Dodatkowo zauważono, że wykorzystanie klasyfikatora kNN również generuje bardzo wysokie rezultaty, bardzo zbliżone do tych uzyskanych dla maksymalnych wartości.



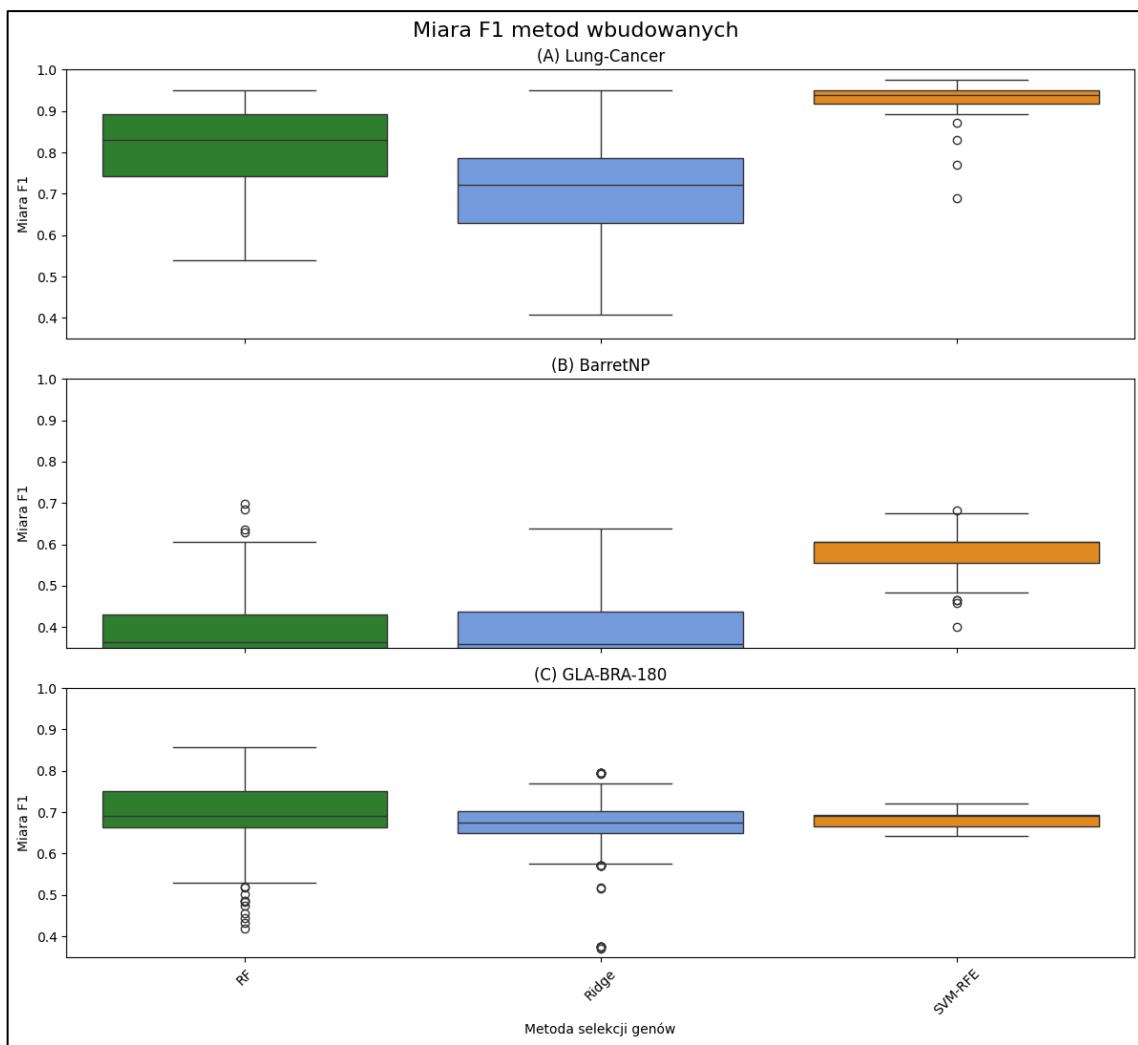
Rysunek 30. Miara F1 metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora

W tabeli 7 przedstawiono parametry, dla których uzyskano najwyższe wyniki wydajności dla każdego zestawu danych, uwzględniając wielkość wyselekcjonowanego podzbioru oraz wykorzystany klasyfikator. Wszystkie zestawione w tabeli podzbiory wykorzystane zostały do przeprowadzenia badań dla metod hybrydowych.

Tabela 7. Wyselekcjonowane podzbiory za pomocą metody SVM-RFE charakteryzujące się najwyższą wydajnością

Metoda	Zbiór danych	Liczba wyselekcjonowanych genów	Klasyfikator uzyskujący najlepsze wyniki	Uzyskana wartość miary F1
SVM-RFE	<i>Lung-Cancer</i>	50	SVM	0,98
SVM-RFE	<i>BarretNP</i>	50	RF	0,68
SVM-RFE	<i>GLA-BRA-180</i>	100	RF	0,72

Analiza porównawcza metod wbudowanych dla badanych zbiorów danych została przedstawiona na rysunku 31. Porównania dokonano na podstawie wszystkich uzyskanych wyników. Na podstawie uzyskanych rezultatów można zauważyć, że metoda SVM-RFE osiągnęła najlepsze wyniki dla zbiorów *Lung-Cancer* (rys. 31A) i *BarretNP* (rys. 31B). Warto zauważyć imponujące przewyższenie wyników dla trudnego zbioru *BarretNP*. W obu przypadkach metoda SVM-RFE osiągnęła najwyższą wydajność przy bardzo małej liczbie wyselekcjonowanych genów, co sugeruje jej efektywność w selekcji cech. Dla zbioru *GLA-BRA-180* (rys. 31C) charakteryzującego się największą liczbą genów, najskuteczniejszą metodą okazał się być RF.



Rysunek 31. Miara F1 dla wszystkich zastosowanych metod wbudowanych

W porównaniu wyników metod wbudowanych do rezultatów uzyskanych na podstawie metod filtrujących (rys. 25) zauważalne są istotne różnice pomiędzy badanymi zbiorami danych. Dla zbioru *Lung-Cancer* wszystkie metody filtrujące uzyskały identyczne maksymalne wyniki, wynoszące 98%. W przypadku metod wbudowanych, tylko metoda SVM-RFE osiągnęła taką wartość, charakteryzując się jednocześnie najmniejszym podziobrem, składającym się tylko z 50 wyselekcjonowanych genów. Przy analizie zbioru *BarretNP*, wszystkie metody filtrujące uzyskały identyczną maksymalną wartość wynoszącą 69%. W przypadku metod wbudowanych wyłącznie metoda RF przekroczyła tę wartość, osiągając 70% skuteczności, dla 200 najbardziej informacyjnych genów tego zbioru. Pozostałe metody wbudowane uzyskały nieco niższe wyniki. Podobnie jak w przypadku zbioru *Lung-Cancer*, metoda SVM-RFE charakteryzuje się najmniejszą różnicą maksymalnych i minimalnych rezultatów. Dla zbioru *GLA-BRA-180* wszystkie metody filtrujące osiągnęły wartość wydajności wynoszącą 86%, natomiast jedynie metoda RF pośród metod wbudowanych uzyskała identyczny rezultat. Pozostałe metody wbudowane uzyskały nieco niższe wyniki, przy czym metoda SVM-RFE wypadła najgorzej.

5.3 Metody hybrydowe

W badaniach nad metodami hybrydowymi, eksplorowano skuteczność połączeń podejścia opakowującego ze wcześniejszą selekcją cech przy użyciu metod filtrujących i wbudowanych. W ramach tych badań, zastosowano analizę na podzbiorach, które dla badanych metod filtrujących oraz wbudowanych wykazały najlepszą wydajność dla każdego zbioru danych. Podzbiory te zestawione zostały w tabelach 4-7.

Model hybrydowy wykorzystujący metodę opakowującą algorytmu genetycznego (GA) sprawdzony został wyłącznie pod kątem zależności liczby iteracji algorytmu na wydajność badanych modeli. W przypadku modeli hybrydowych wykorzystujących metody opakowujące w postaci algorytmu sztucznej kolonii pszczół (ABC) oraz algorytmu roju częstek (PSO) zbadana została zależność liczby iteracji algorytmów oraz wielkości zainicjalizowanych populacji rozwiązań na ich wydajność. Wartości metryk dla określonej liczby iteracji czy wielkości zainicjalizowanych populacji rozwiązań stanowiły uśrednione wyniki wszystkich badanych rozwiązań przypadających dla określonych wartości. Skuteczność algorytmów sprawdzona została dla następujących wartości iteracji: 50, 100, 200, 500, 1000. Natomiast wielkość zainicjalizowanych populacji rozwiązań sprawdzona została dla wartości wynoszących: 10, 20, 50, 100, 200, 500, 1000.

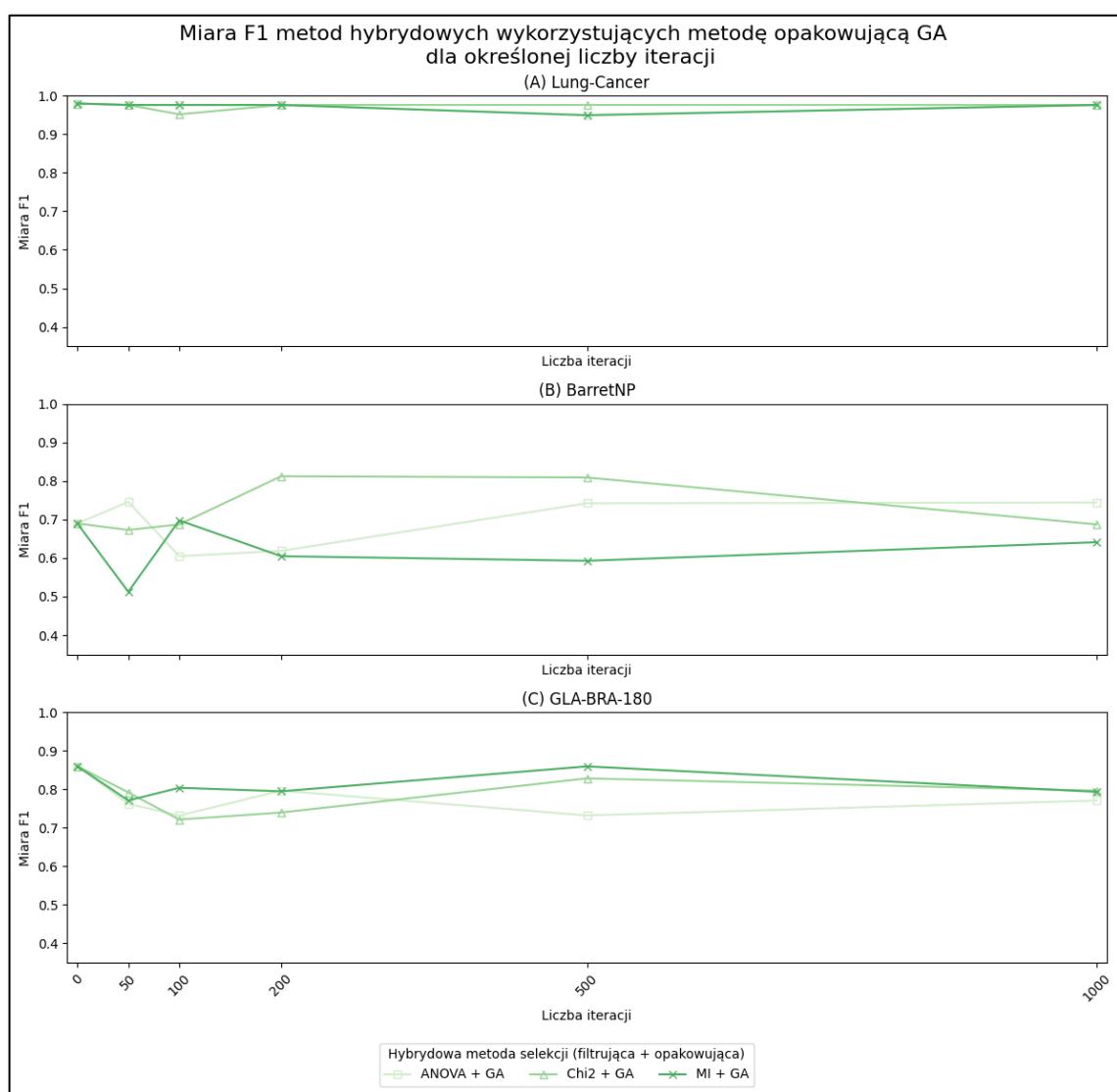
Tak samo, jak w przypadku analizy wyników metod filtrujących i wbudowanych, wyniki zostały przedstawione przy użyciu miary F1. Co więcej, ocena wydajności metod hybrydowych została przeprowadzona wyłącznie dla tego klasyfikatora, w przypadku którego wybrane podzbiory uzyskane na podstawie metod filtrujących i wbudowanych osiągnęły maksymalne wartości wydajności (tab. 4-7).

W metodologii przeprowadzonych badań uwzględniono, że dla każdego zaimplementowanego algorytmu opakowującego w ramach podejścia hybrydowego, jako metodę dopasowania wykorzystano klasyfikator kNN (rozdział 5.2.3). Wybór ten został podjęty po analizie metod filtrujących i wbudowanych, które jednoznacznie wykazały, że zastosowanie tego konkretnego klasyfikatora przyniosło maksymalne wartości wydajności dla 12 z 18 uzyskanych podzbiorów, co zostało szczegółowo przedstawione w tabelach 4-7.

5.3.1 Metody filtrujące + metody opakowujące

Wyniki miary F1 metody hybrydowej wykorzystującej metodę opakowującą GA dla określonej liczby iteracji wszystkich wyselekcjonowanych przez metody filtrujące podzbiorów danych zaprezentowane zostały na rysunku 32. Dla zbioru *Lung-Cancer* (rys. 32A) można zauważyć, że ilość iteracji nie wpływa istotnie na wyniki, ponieważ rezultaty utrzymują się na bardzo zbliżonym poziomie do wyników wejściowych (0 iteracji). W przypadku zbioru *BarretNP* (rys. 32B) można zauważyć, że dla podzbioru wyselekcjonowanego za pomocą metody chi2, metoda GA osiąga najwyższe wyniki dla 200 i 500 iteracji algorytmu, co przekłada się na zwiększenie wydajności modelu w porównaniu do

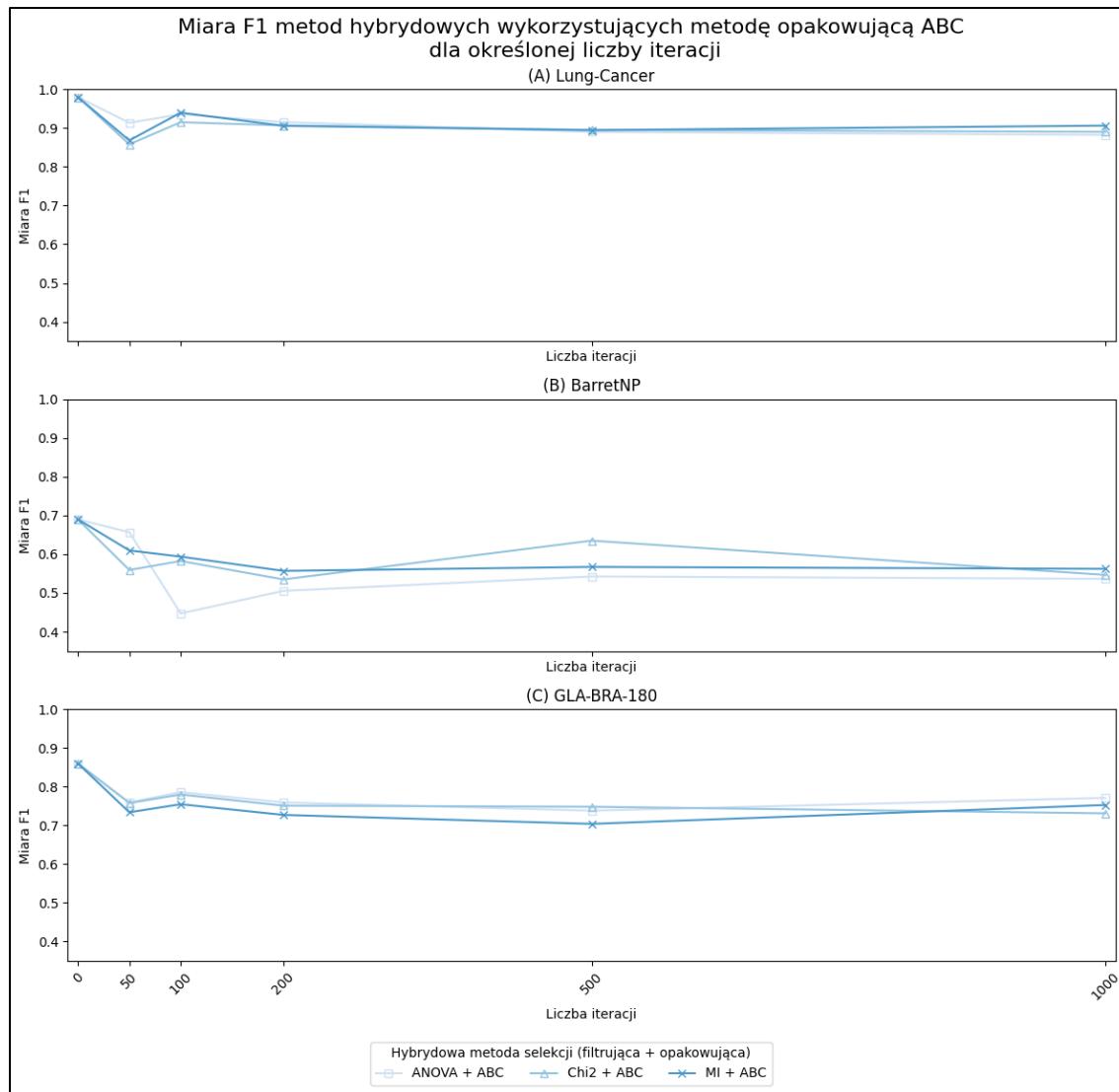
wcześniej wyselekcjonowanego podzbioru za pomocą metody chi2, osiągając w obu przypadkach ponad 80%, w porównaniu do wejściowych 69%. Dodatkowo zauważać można, że dla tego zbioru metoda GA zastosowana na podzbiorze wstępnie wyselekcjonowanym metodą ANOVA przy niewielkiej ilości iteracji powoduje spadek wydajności, jednakże wraz z zwiększeniem liczby iteracji wydajność rośnie, osiągając dla 1000 iteracji wynik zbliżony do wyniku wejściowego. W przypadku zbioru *GLA-BRA-180* (rys. 32C) wszystkie rezultaty osiągają bardzo zbliżone wyniki dla kolejnych wartości iteracji algorytmu, uzyskując dla podzbiorów wcześniej wyselekcjonowanych przez metody filtrujące chi2 oraz MI wartości zbliżone do wejściowych dla 500 iteracji.



Rysunek 32. Miara F1 metody hybrydowej wykorzystującej algorytm genetyczny (GA) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów

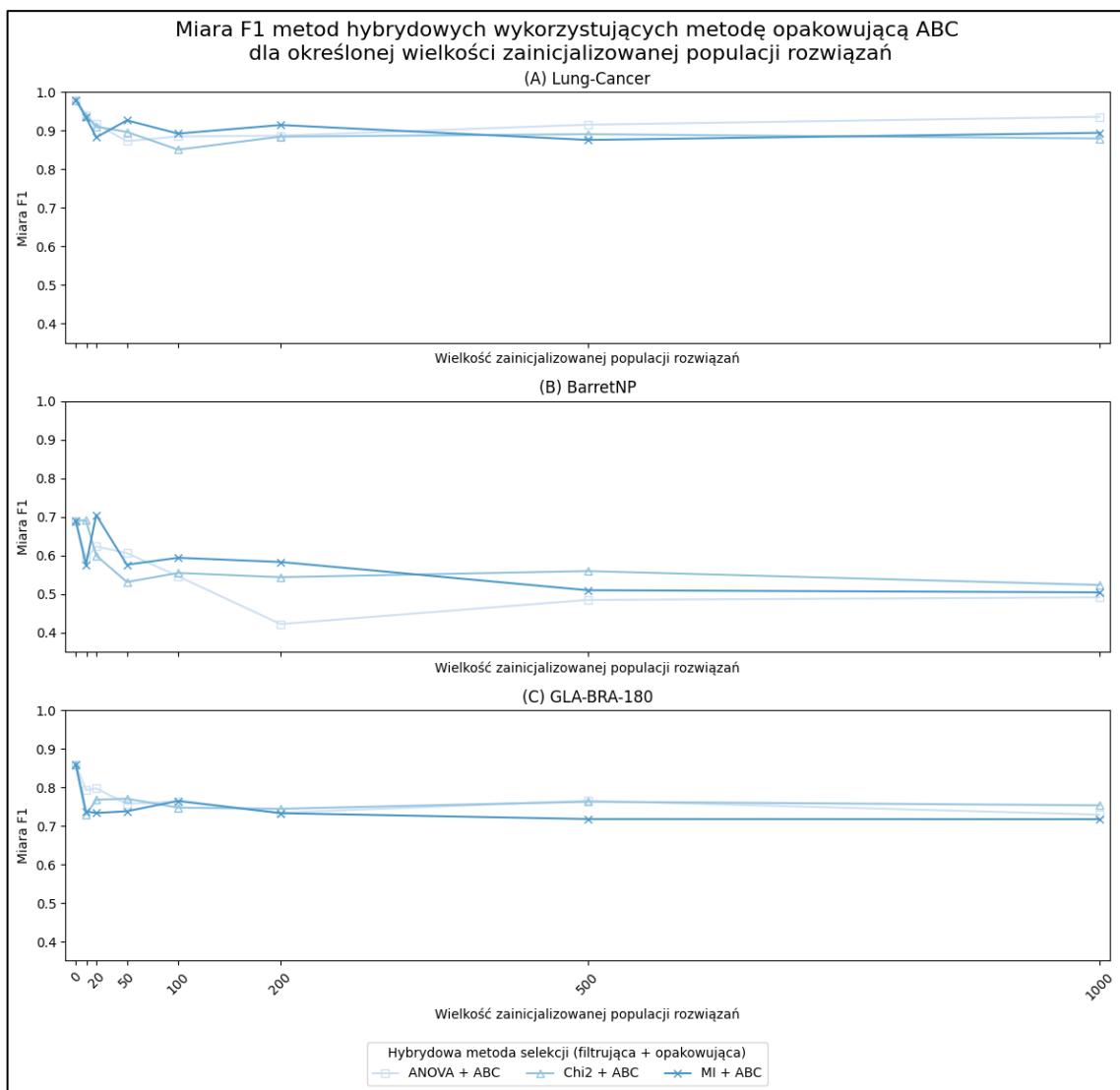
Na rysunku 33 zaprezentowano uśrednione rezultaty miary F1 dla metod hybrydowych, wykorzystujących algorytm ABC, w zależności od liczby iteracji. Dla każdej wartości iteracji, uzyskane wyniki są bardzo zbliżone dla wszystkich wcześniej wyselekcjonowanych podzbiorów metodami

filtrującymi, osiągając nieco niższe wartości niż wynik wejściowy. Dodatkowo model ten utrzymuje stabilność rezultatów dla kolejnych iteracji algorytmu.



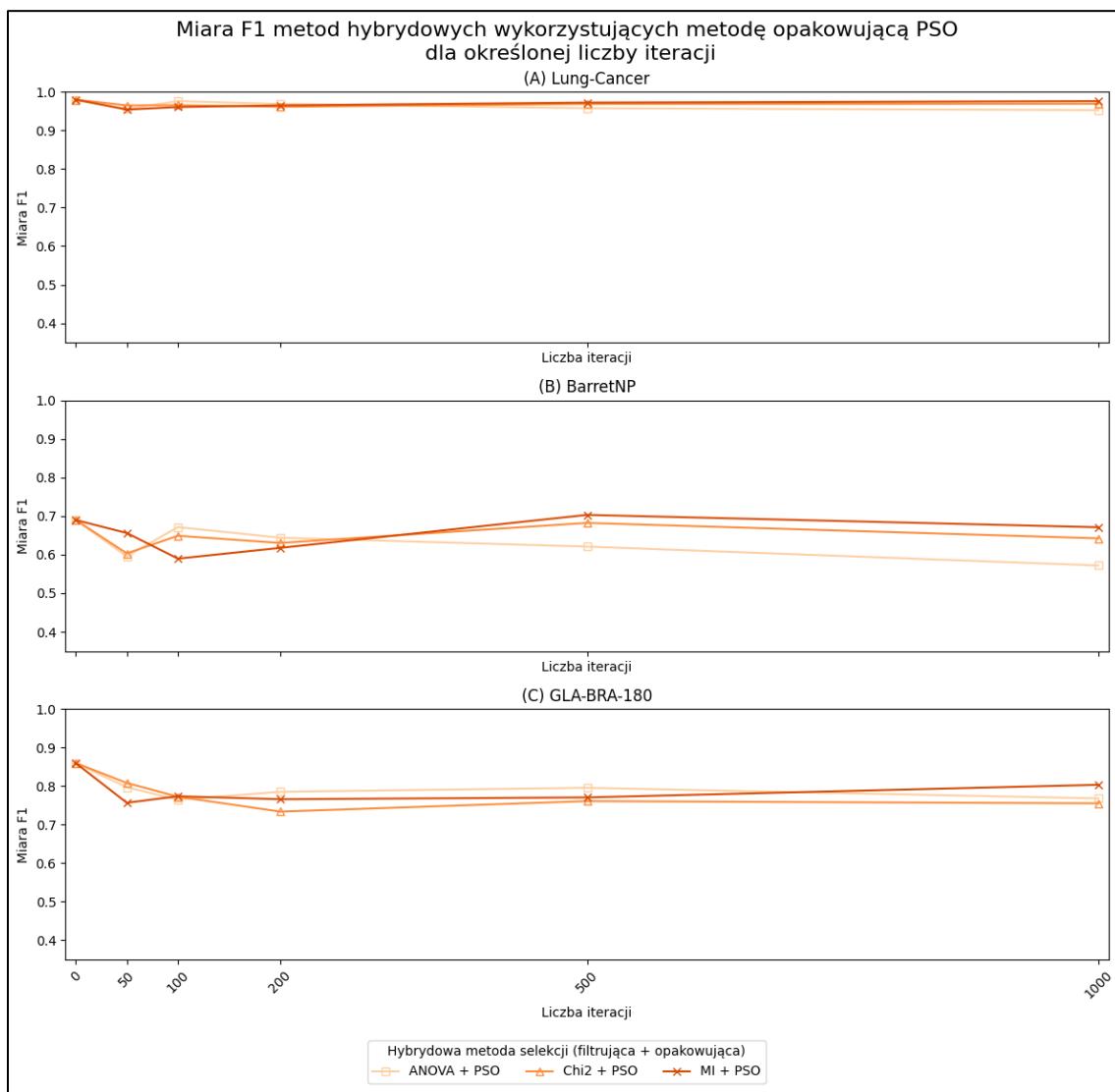
Rysunek 33. Średnia miara F1 metody hybrydowej wykorzystującej algorytm sztucznej kolonii pszczół (ABC) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów

Na rysunku 34 przedstawione zostały uśrednione wyniki miary F1 dla metod hybrydowych, wykorzystujących algorytm ABC, w zależności od wielkości zainicjalizowanej populacji rozwiązań algorytmu. Podobnie jak w przypadku analizy wpływu ilości iteracji, dla wszystkich zbiorów zauważalne jest, że wszystkie wartości utrzymują bardzo zbliżone względem siebie rezultaty dla kolejnych wartości wielkości zainicjalizowanych populacji rozwiązań. Ciekawą obserwacją jest, że dla zbioru *BarretNP* (rys. 34B) uśrednione wyniki miary F1 dla zainicjalizowanej populacji wynoszącej 20 rozwiązań, dla podzbioru wcześniej wyselekcyjowanego przez metodą filtrującą MI, nieznacznie przewyższają wartość wejściową dla tego zbioru.



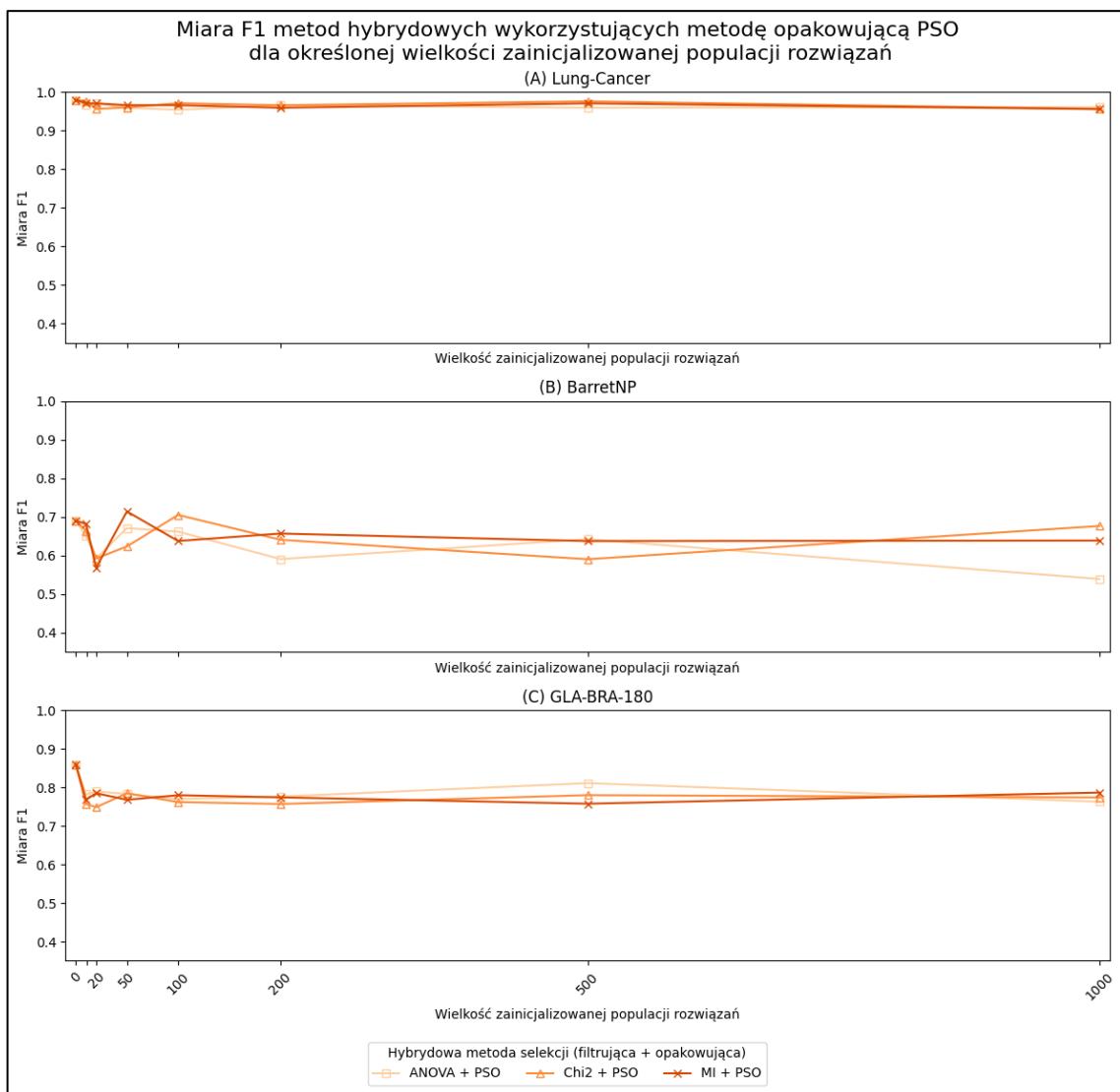
Rysunek 34. Średnia miara F1 metody hybrydowej wykorzystującej algorytm sztucznej kolonii pszczół (ABC) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów

Następnie na rysunku 35 zaprezentowano uśrednione rezultaty miary F1 dla metod hybrydowych, wykorzystujących algorytm PSO, w zależności od liczby iteracji. W przypadku zbioru *Lung-Cancer* (rys. 35A) zastosowanie algorytmu PSO, tak samo jak w przypadku algorytmu GA, wraz z kolejnymi iteracjami praktycznie nie wpływa na wydajność modelu, utrzymując dla wszystkich wartości iteracji wyniki bardzo zbliżone do wartości wejściowej. Dla zbioru *BarretNP* (rys. 35B) zauważać można, że dla 500 iteracji algorytmu podzbiorów uzyskane wcześniej za pomocą metody filtrującej MI nieznacznie przewyższają wartość wydajności wejściowej. Natomiast dla zbioru *GLA-BRA-180* (rys. 35C) algorytm PSO utrzymuje bardzo zbliżone rezultaty dla kolejnych wartości iteracji.



Rysunek 35. Średnia miara F1 metody hybrydowej wykorzystującej algorytm roju cząstek (PSO) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów

Uśrednione rezultaty miary F1 dla metod hybrydowych, wykorzystujących algorytm PSO, w zależności od wielkości zainicjalizowanej populacji rozwiązań algorytmu zaprezentowane zostały na rysunku 36. W przypadku zbioru *Lung-Cancer* (rys. 36A) kolejne badane wielkości inicjalizowanej populacji rozwiązań algorytmu, tak samo jak w przypadku analizy wpływu iteracji, nie mają znaczącego wpływu na wydajność modelu, utrzymując dla wszystkich wartości bardzo zbliżone rezultaty do wyniku wejściowego. Dla zbioru *BarretNP* (rys. 36B) zauważalne jest, że dla zainicjalizowanej populacji 50 rozwiązań, podzbiory uzyskane wcześniej za pomocą metody filtrującej MI nieznacznie przewyższają wartość wydajności wejściowej. Dodatkowo dla zainicjalizowanej populacji o wielkości 100 rozwiązań, podzbiory uzyskane wcześniej za pomocą metody filtrującej chi2 również nieznacznie przewyższają wyjściową wartość wydajności dla tego zbioru. Natomiast dla zbioru *GLA-BRA-180* (rys. 36C) algorytm PSO utrzymuje dla kolejnych wielkości zainicjalizowanych populacji rozwiązań bardzo zbliżone rezultaty dla wszystkich wcześniej wyselekcjonowanych za pomocą metod filtrujących podzbiorów.



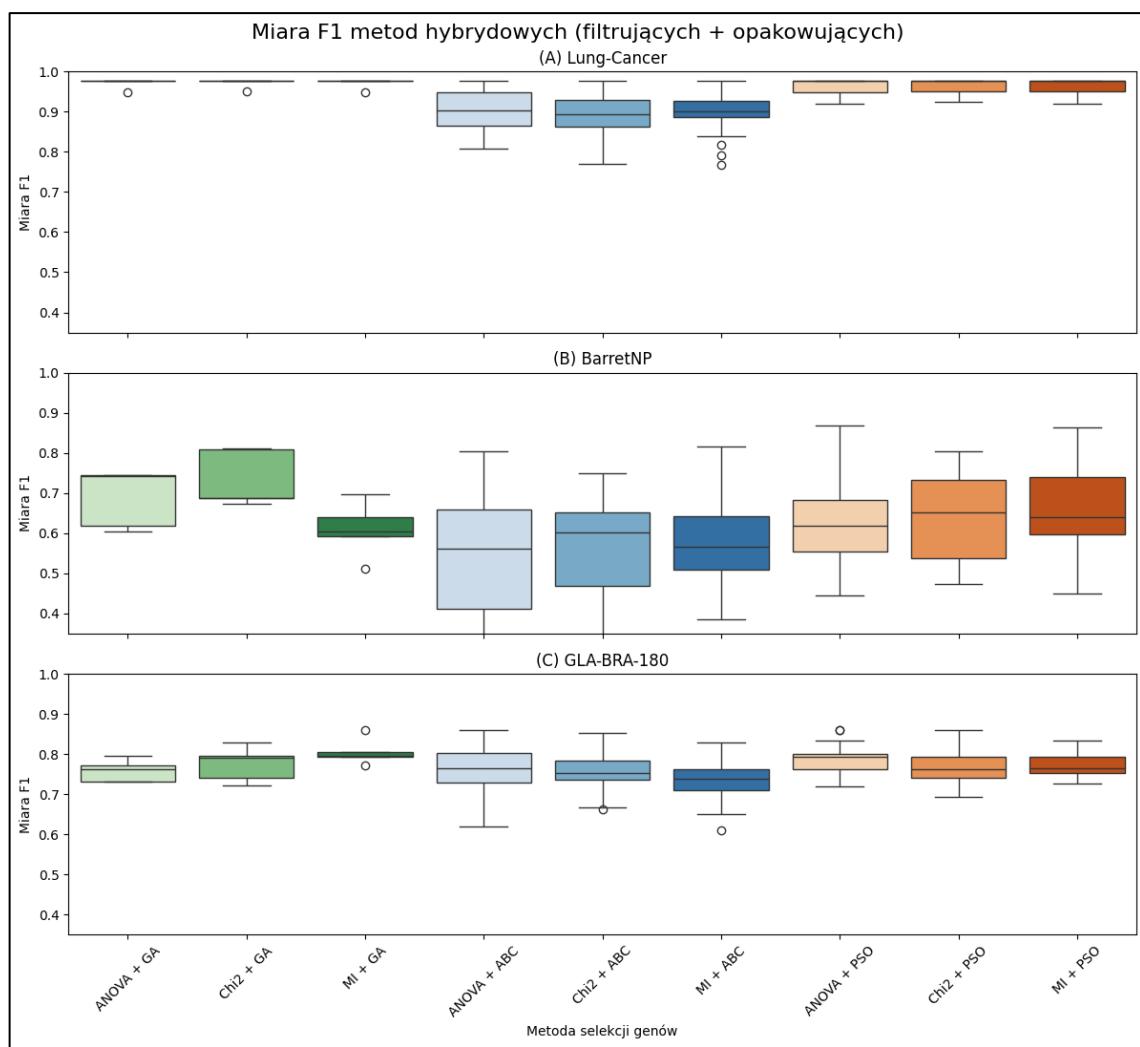
Rysunek 36. Średnia miara F1 metody hybrydowej wykorzystującej algorytm roju cząstek (PSO) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody filtrujące podzbiórów

Na rysunku 37 przedstawione zostały wszystkie uzyskane wyniki miary F1 metod hybrydowych wykorzystujących podzbiory wyselekcjonowane wcześniej za pomocą metod filtrujących. Rezultaty zostały zaprezentowane bez rozróżnienia na liczbę iteracji czy wielkość zainicjalizowanych populacji rozwiązań algorytmów w celu porównania wydajności wszystkich badanych modeli oraz identyfikacji maksymalnych uzyskanych wartości dla modeli hybrydowych.

Dla zbioru *Lung-Cancer* (rys. 37A) można zauważyć, że wszystkie badane modele filtrujące osiągnęły maksymalną wydajność na poziomie 98% (tab. 4). W przypadku tego stosunkowo łatwego zbioru wszystkie zastosowane modele hybrydowe osiągnęły identyczną wydajność jak pojedyncze metody filtrujące. Dodatkowo, metoda wykorzystująca algorytm GA charakteryzuje się najmniejszymi rozbieżnościami wyników, uzyskując wyniki zbliżone do wartości maksymalnej niezależnie od liczby iteracji tego algorytmu. Z drugiej strony, metody wykorzystujące algorytmy ABC i PSO wykazują stosunkowo większe rozbieżności, zależne od liczby iteracji i wielkości zainicjalizowanych populacji

rozwiązań algorytmów. Dla zbioru *BarretNP* (rys. 37B) zauważono, że wszystkie badane modele charakteryzują się różnymi maksymalnymi wynikami wydajności. Wszystkie metody filtrujące podczas selekcji najbardziej informacyjnych genów dla tego zbioru osiągnęły maksymalna wartość wydajności wynoszącą 69% (tab. 4), co stanowiło wejściową wartość dla metod hybrydowych. Wszystkie zastosowane algorytmy opakowujące w metodach hybrydowych spowodowały wzrost tej wydajności, a najwyższą wartość osiągnęła metoda hybrydowa ANOVA + PSO, uzyskując 87% wydajności dla tego trudnego zbioru, co stanowi znaczny wzrost względem wartości wyjściowej. Dla zbioru *GLA-BRA-180* (rys. 37C) również zauważono, że wszystkie badane modele charakteryzują się zbliżonymi maksymalnymi wynikami wydajności. Wszystkie metody filtrujące podczas selekcji najbardziej informacyjnych genów dla tego zbioru osiągnęły wartość wynoszącą 86% (tab. 4). Żaden z badanych modeli hybrydowych nie poprawił tego wyniku wejściowego, jednak dla kilku modeli udało się uzyskać dokładnie taką samą wartość. Zbiór ten charakteryzuje się największą liczbą genów spośród wszystkich badanych zbiorów, co może sugerować, że metody filtrujące odrzuciły podczas selekcji geny, które mogłyby wpłynąć na poprawę wydajności po zastosowaniu metod opakowujących.

W tabeli 8 zestawiono uzyskane maksymalne rezultaty metod hybrydowych dla wszystkich wyselekcjonowanych za pomocą metod filtrujących podzbiorów, uwzględniając wielkość wyselekcjonowanego podzbioru genów, liczbę iteracji oraz wielkość zainicjalizowanej populacji rozwiązań danego algorytmu opakowującego. Wobec zaprezentowanych danych, trudno jest jednoznacznie określić wpływ analizowanych wartości parametrów metod opakowujących. Jednak zauważać można, że wartości parametrów, które doprowadziły do uzyskania maksymalnych wartości wydajności dla poszczególnych modeli zależą zarówno od badanego zbioru, jak i od zastosowanej wcześniej metody filtrującej. W przypadku liczby iteracji zauważać można, że dla stosunkowo łatwego zbioru *Lung-Cancer* potrzebna jest mniejsza liczba iteracji do uzyskania maksymalnej wartości wydajności dla wszystkich badanych algorytmów, niż dla złożonego zbioru *BarretNP*, bądź największego badanego zbioru *GLA-BRA-180*. Dodatkowo zgodnie z uzyskanymi rezultatami można stwierdzić, że ilość wymaganych iteracji algorytmów opakowujących do uzyskania maksymalnych wartości wydajności zależy od wcześniej zastosowanej metody filtrującej. W badanym problemie największej ilości iteracji wymagały podzbiory wyselekcjonowane za pomocą metody MI, natomiast najmniejszej podzbiory wyselekcjonowane przy wykorzystaniu metody chi2. W przypadku wartości wielkości zainicjalizowanych populacji rozwiązań algorytmów metod opakowujących trudno jest określić podobne zależności.



Rysunek 37. Miara F1 zastosowanych metod hybrydowych (metody filtrujące + opakowujące)

Tabela 8. Uzyskane maksymalne rezultaty wydajności metod hybrydowych dla wszystkich wyselekcjonowanych za pomocą metod filtrujących podzbiorów

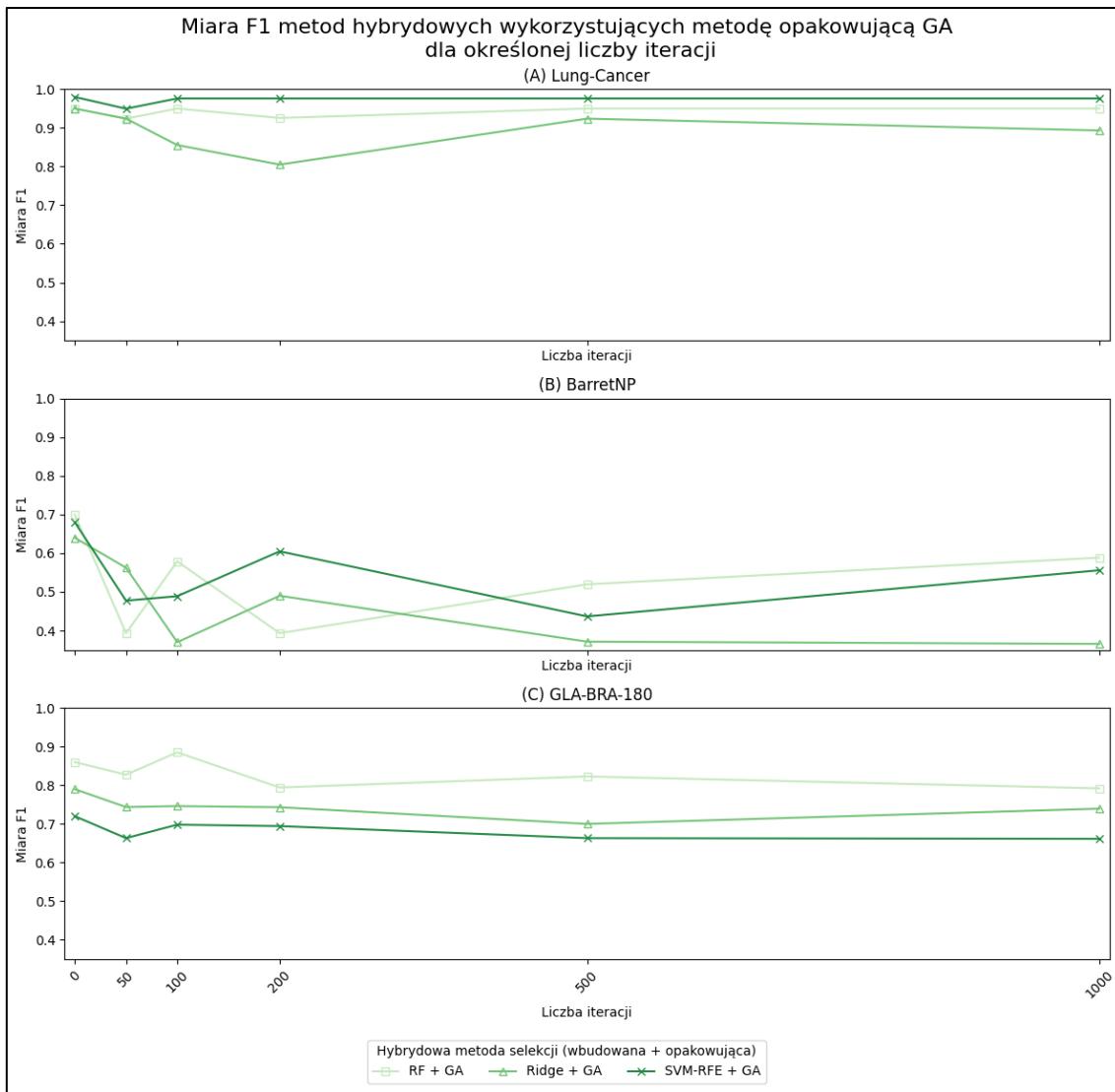
Metoda filtrująca					Metoda opakowująca					
Metoda	Zbiór danych	Liczba wyselekcjonowanych genów	Klasyfikator uzyskujący najlepsze wyniki	Uzyskana wartość miary F1	Metoda	Liczba iteracji algorytmu	Wielkość zainicjalizowanej populacji rozwiązań	Uzyskana wartość miary F1	Wielkość wyselekcjonowanego podzbioru genów	
ANOVA	<i>Lung-Cancer</i>	500	kNN	0,98	GA	50	-	0,98	263	
					ABC	500	1000	0,98	76	
					PSO	50	10	0,98	227	
ANOVA	<i>BarretNP</i>	1000	kNN	0,69	GA	50	-	0,75	507	
					ABC	50	20	0,80	337	
					PSO	500	500	0,87	270	
ANOVA	<i>GLA-BRA-180</i>	500	kNN	0,86	GA	200	-	0,80	259	
					ABC	500	10	0,86	316	
					PSO	500	50	0,86	212	
Chi2	<i>Lung-Cancer</i>	500	kNN	0,98	GA	50	-	0,98	256	
					ABC	50	20	0,98	131	
					PSO	200	100	0,98	227	
Chi2	<i>BarretNP</i>	1000	kNN	0,69	GA	200	-	0,81	528	
					ABC	500	10	0,75	147	
					PSO	500	50	0,80	492	

Chi2	<i>GLA-BRA-180</i>	500	kNN	0,86	GA ABC PSO	500 50 50	- 500 500	0,83 0,85 0,86	244 12 186
MI	<i>Lung-Cancer</i>	500	kNN	0,98	GA ABC PSO	50 100 1000	- 10 500	0,98 0,98 0,98	258 149 117
MI	<i>BarretNP</i>	1000	kNN	0,69	GA ABC PSO	100 1000 1000	- 20 100	0,70 0,82 0,86	497 77 514
MI	<i>GLA-BRA-180</i>	500	kNN	0,86	GA ABC PSO	500 1000 1000	- 500 10	0,86 0,83 0,83	257 33 214

5.3.2 Metody wbudowane + metody opakowujące

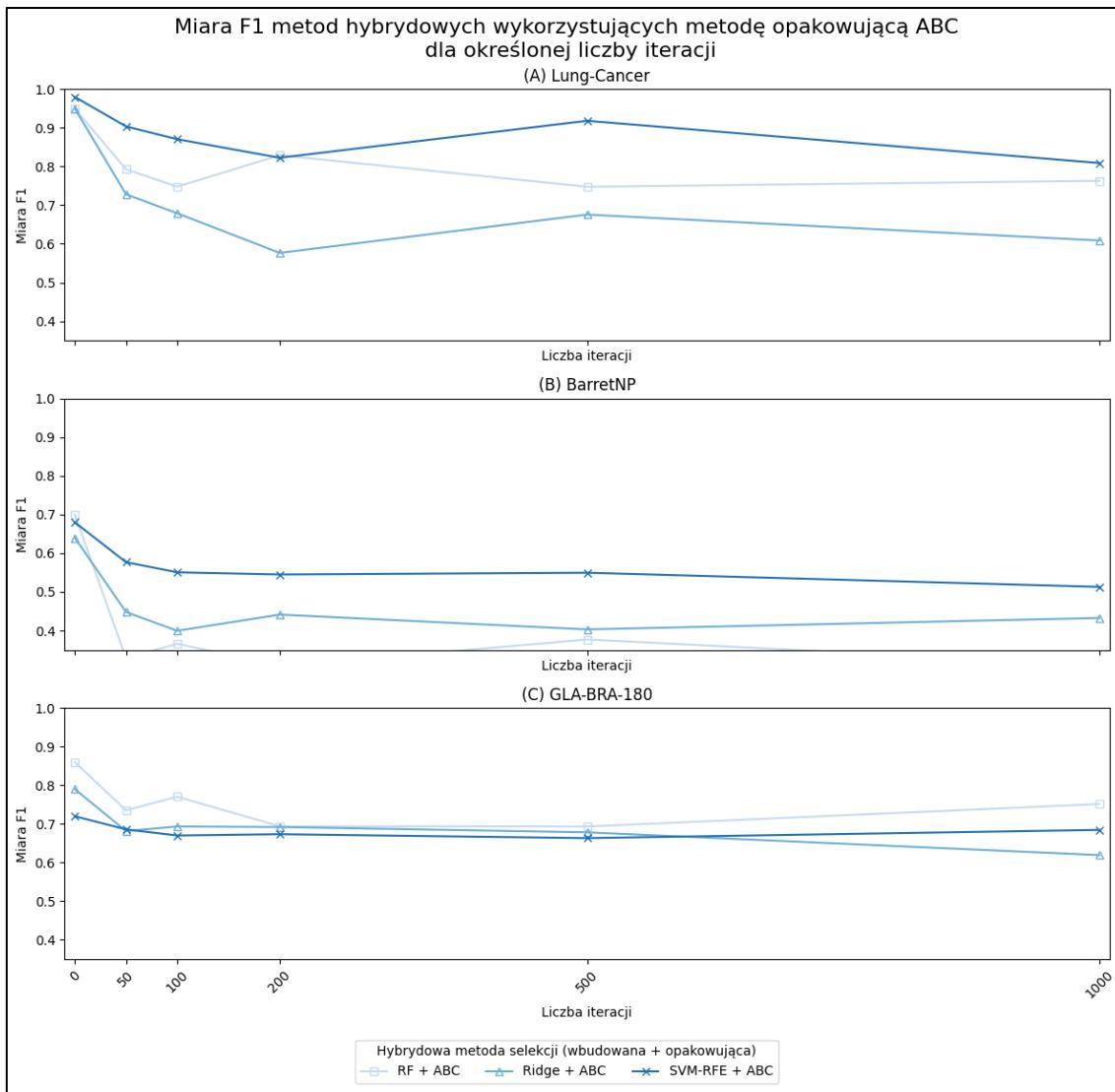
Kolejnym krokiem była weryfikacja skuteczności metod hybrydowych stanowiących połączenie podejść opakowujących ze wcześniejszą selekcją cech przy użyciu metod wbudowanych. W porównaniu do podzbiorów wyselekcjonowanych w oparciu o metody filtrujące, wejściowe wyniki wydajności dla wszystkich metod i badanych zbiorów były różne.

Na rysunku 38 zaprezentowane zostały wyniki miary F1 metody hybrydowej wykorzystującej metodę opakowującą GA dla określonej liczby iteracji wszystkich wyselekcjonowanych przez metody wbudowane podzbiorów. Dla zbioru *Lung-Cancer* (rys. 38A) zauważać można, że ilość wykonywanych iteracji algorytmu GA nie ma istotnego wpływu na wydajność modeli wykorzystujących podzbiory wyselekcjonowane wcześniej przy użyciu RF oraz SVM-RFE. Rezultaty utrzymują się na bardzo zbliżonym poziomie do wyników wejściowych. W przypadku modelu wykorzystującego podzbiór wyselekcjonowany metodą wbudowaną Ridge, zauważono, że kolejne iteracje algorytmu GA prowadzą do obniżenia wydajności. Dla zbioru *BarretNP* (rys. 38B), dla wszystkich trzech modeli zaobserwowano znaczne wahania rezultatów wydajności w zależności od liczby iteracji algorytmu GA. Warto zaznaczyć, że najgorzej wypadającym modelem była metoda wykorzystująca podzbiór wcześniejszy wyselekcjonowany przez metodę Ridge. Dla zbioru *GLA-BRA-180* (rys. 38C) stwierdzono dużą stabilność modeli, przy czym kolejne iteracje algorytmu GA nie wpływają na wydajność. Warto zauważać, że dla wszystkich iteracji wyniki poszczególnych modeli są bardzo zbliżone do wyników wejściowych, co świadczy o wysokiej jakości wyselekcjonowanych genów dla tego zbioru przy użyciu metod wbudowanych.



Rysunek 38. Miara F1 metody hybrydowej wykorzystującej algorytm genetyczny (GA) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów

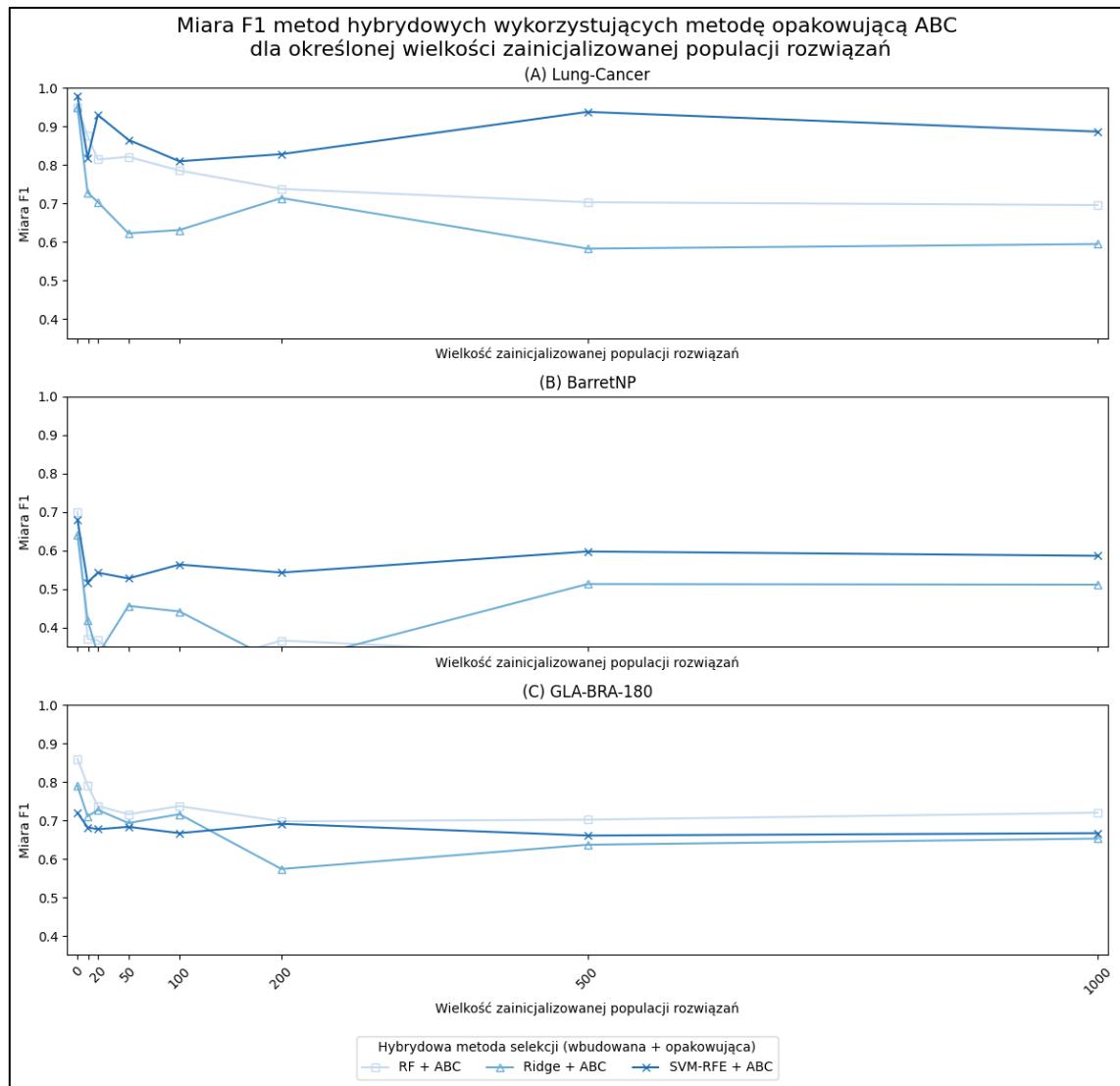
Następnie na rysunku 39 zaprezentowano uśrednione rezultaty miary F1 dla metod hybrydowych, wykorzystujących algorytm ABC, w zależności od liczby iteracji dla wszystkich wcześniej wyselekcjonowanych przez metody wbudowane podzbiorów. Dla zbioru *Lung-Cancer* (rys. 39A) zauważono, że kolejne wartości iteracji powodują spadek wydajności dla wszystkich modeli w porównaniu do wydajności wejściowych. Jedyną widoczną wzrost wydajności zaobserwowano dla modeli wykorzystujących podzbiory wyselekcjonowane przy użyciu metod SVM-RFE oraz RF dla 500 iteracji. Niemniej jednak, wartość wydajności dla tej wartości iteracji pozostaje znacznie niższa od wartości wejściowych. W przypadku pozostałych zbiorów, algorytm ABC nieco obniża średnie wartości wydajności w porównaniu do wydajności wejściowych, utrzymując dla kolejnych wartości iteracji bardzo zbliżone rezultaty.



Rysunek 39. Średnia miara F1 metody hybrydowej wykorzystującej algorytm sztucznej kolonii psozłów (ABC) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów

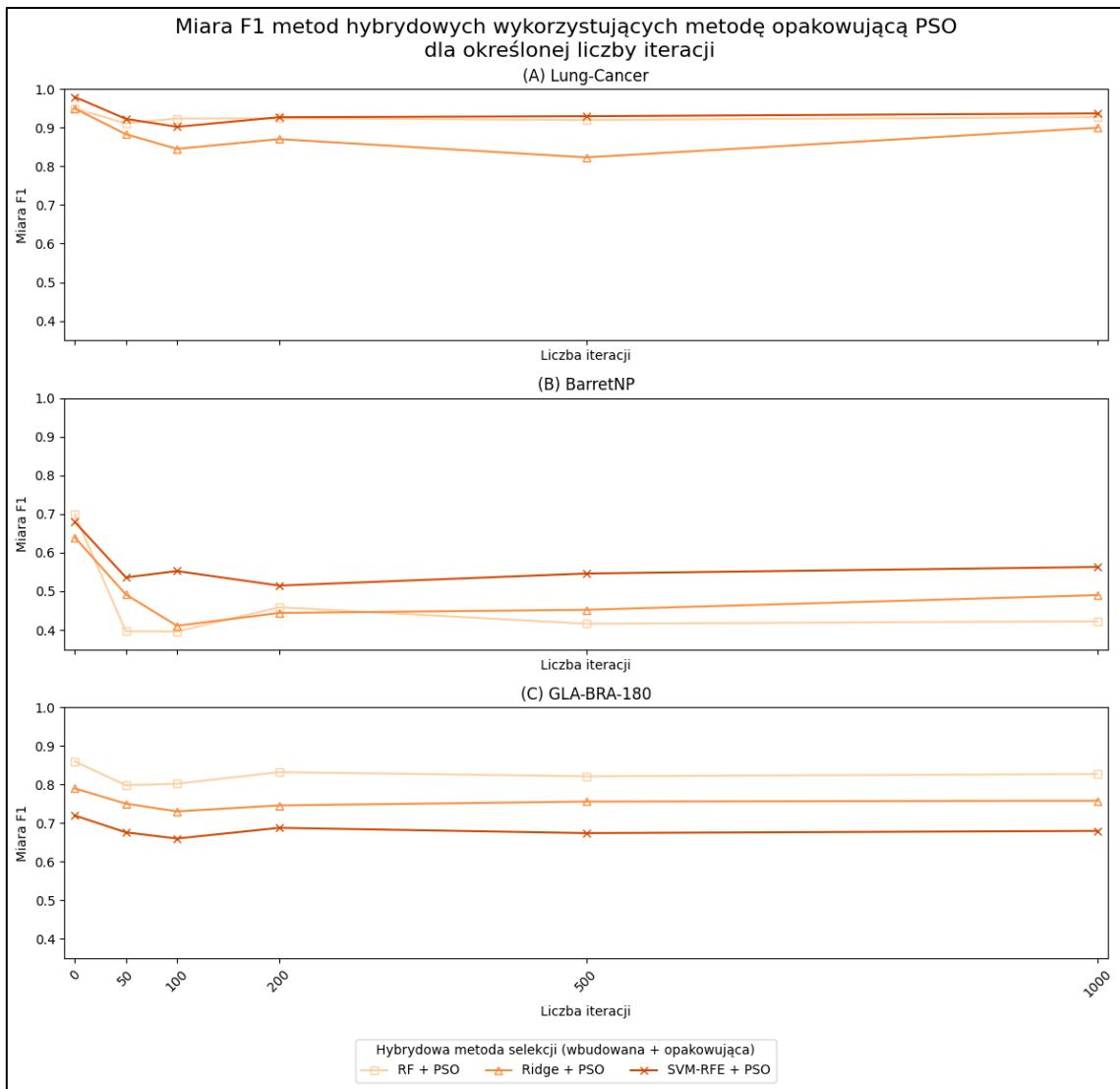
Na rysunku 40 przedstawione zostały uśrednione wyniki miary F1 dla metod hybrydowych, wykorzystujących algorytm ABC, w zależności od wielkości zainicjalizowanej populacji rozwiązań algorytmu dla wszystkich wcześniej wyselekcjonowanych przez metody wbudowane podzbiorów. Dla zbioru *Lung-Cancer* (rys. 40A), nie można jednoznacznie stwierdzić jednej tendencji zależności wielkości zainicjalizowanej populacji rozwiązań od średnich wydajności modeli. Dla podzbioru wcześniejszej wyselekcjonowanego za pomocą metody wbudowanej SVM-RFE zauważa się wzrost wydajności dla grupy 500 zainicjalizowanych rozwiązań, natomiast dla podzbioru wyselekcjonowanego za pomocą metody Ridge zauważa się spadek dla tej samej grupy zainicjalizowanych rozwiązań. Dla zbioru *BarretNP* (rys. 40B) zauważono, że uśrednione wyniki są znacznie niższe od wejściowych wartości wydajności. Ponadto, można zauważyć, że jedynym stabilnie zachowującym się modelem względem kolejnych wielkości zainicjalizowanych populacji rozwiązań jest podzióbior wcześniejszej wyselekcjonowany za pomocą metody SVM-RFE. W przypadku zbioru *GLA-BRA-180* (rys. 40C),

algorytm ABC utrzymuje dla kolejnych wielkości populacji stosunkowo zbliżone rezultaty dla wszystkich badanych modeli.

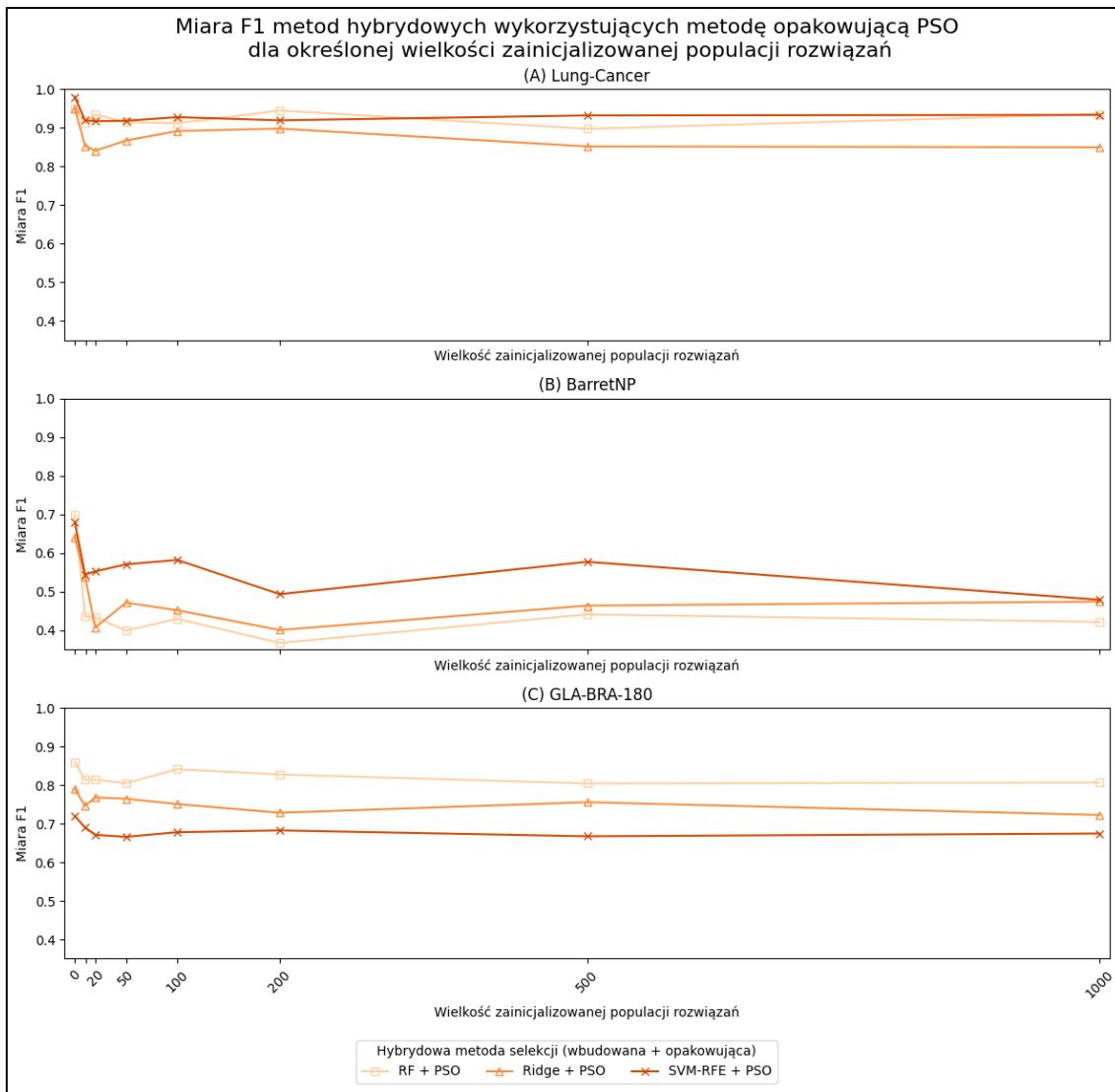


Rysunek 40. Średnia miara F1 metody hybrydowej wykorzystującej algorytm sztucznej kolonii pszczół (ABC) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów

Na rysunkach 41 i 42 przedstawiono uśrednione wyniki miany F1 dla metod hybrydowych, wykorzystujący algorytm opakowujący PSO, odpowiednio: w zależności od liczby iteracji oraz w zależności od wielkości zainicjalizowanej populacji rozwiązań algorytmu. W obu analizowanych problemach zauważono, że dla wszystkich badanych zbiorów danych, średnie wartości wydajności są nieco niższe od wartości wejściowych wszystkich wcześniej wyselekcjonowanych podzbiorów. Dodatkowo, we wszystkich przypadkach, stwierdzono, że kolejne wartości badanych parametrów tego algorytmu opakowującego nie afektują w sposób istotny na uzyskane rezultaty wydajności, dając dla badanych modeli bardzo zbliżone rezultaty.



Rysunek 41. Średnia miara F1 metody hybrydowej wykorzystującej algorytm roju cząstek (PSO) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów



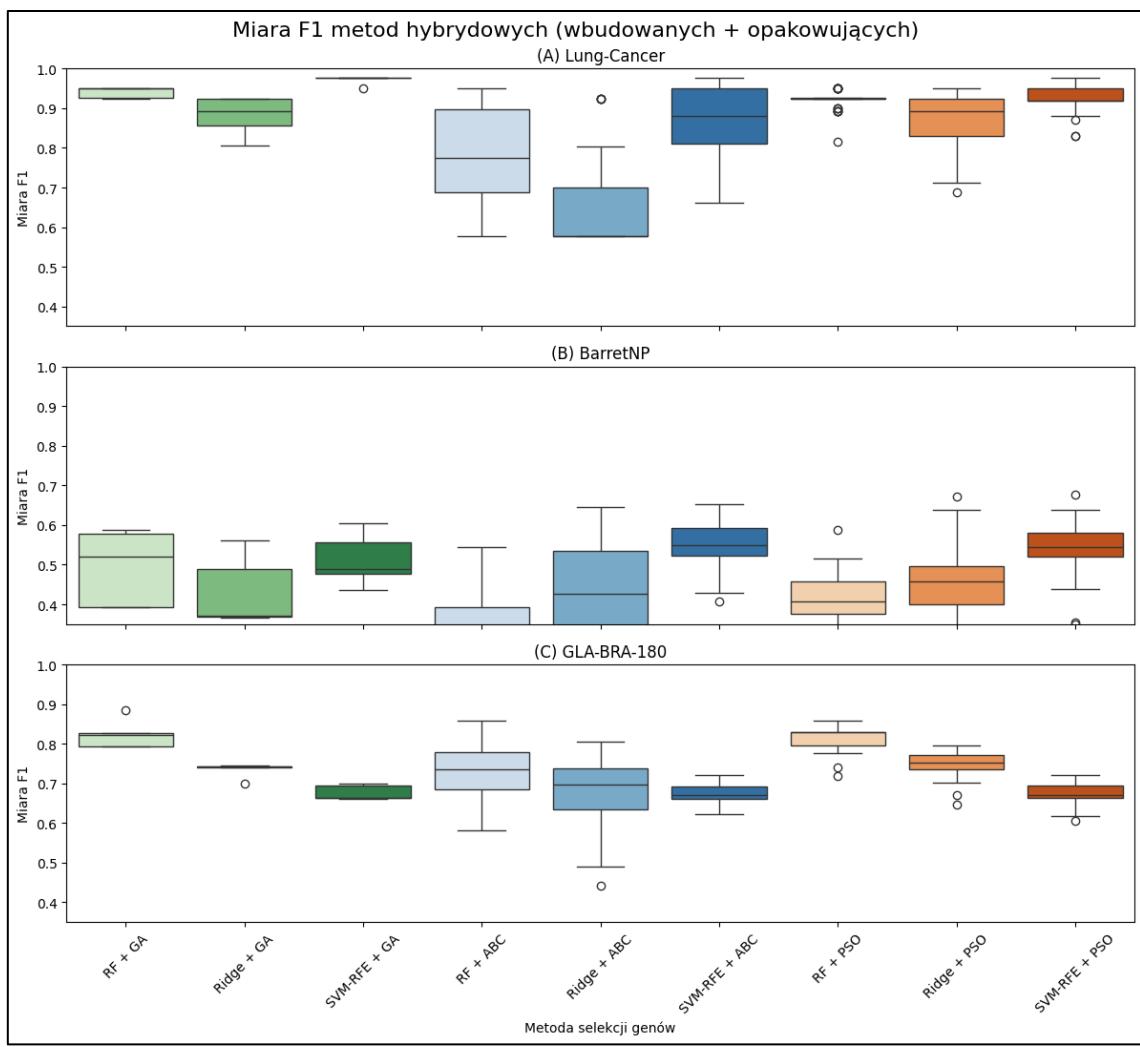
Rysunek 42. Średnia miara F1 metody hybrydowej wykorzystującej algorytm roju cząstek (PSO) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów

Na rysunku 43 przedstawiono wyniki pomiarów miary F1 dla metod hybrydowych, korzystających z wcześniej wyselekcjonowanych podzbiorów za pomocą metod wbudowanych (tab. 5-7). Rezultaty zostały przedstawione, bez podziału na liczbę iteracji czy wielkość zainicjowanych populacji rozwiązań algorytmów, co pozwoliło na porównanie wydajności wszystkich badanych modeli.

Dla zbioru *Lung-Cancer* (rys. 43A) zauważono, że wszystkie badane modele charakteryzują się różnymi maksymalnymi wynikami wydajności. W przypadku tego stosunkowo łatwego zbioru jedynie modele wykorzystujące podzbiory wyselekcjonowane wcześniej za pomocą metody wbudowanej SVM-RFE osiągnęły poziom 98% wydajności. Pozostałe modele wykazują znacznie gorsze wyniki w porównaniu do wartości wejściowych. Dodatkowo większość badanych modeli dla tego zbioru charakteryzuje się dużymi różnicami w wynikach wydajności w zależności od przyjętych parametrów algorytmów opakowujących, takich jak liczba iteracji czy wielkość zainicjalizowanych populacji

rozwiązań. W przypadku zbioru *BarretNP* (rys. 43B) zauważono, że wszystkie uzyskane za pomocą metod hybrydowych rezultaty są niższe od wyników uzyskanych za pomocą pojedynczych metod wbudowanych. Najmniejsze rozbieżności pomiędzy uzyskanymi rezultatami, niezależnie od parametrów algorytmów opakowujących, uzyskano dla podzbiorów wyselekcjonowanych wcześniej za pomocą metody SVM-RFE, przy jednoczesnym osiągnięciu najwyższych wyników dla tego zbioru wśród wszystkich metod hybrydowych. Dla zbioru *GLA-BRA-180* (rys. 43C) zauważono, że na uzyskane rezultaty metod hybrydowych większy wpływ mają wcześniej zastosowane wbudowane metody selekcji niż metody opakowujące. Najwyższe wyniki wydajności uzyskano dla podzbiorów wyselekcjonowanych za pomocą metody wbudowanej RF, natomiast najniższe dla SVM-RFE. Dodatkowo zauważono, że w przypadku zastosowania metody RF + GA dla tego zbioru udało się uzyskać 89% wydajności w porównaniu do wejściowych 86% wydajności uzyskanych w oparciu o pojedynczą metodę wbudowaną RF. Wobec tego, dla tego zbioru metoda hybrydowa RF + GA osiągnęła lepsze rezultaty niż metody hybrydowe wykorzystujące metody filtrujące.

W tabeli 9 zestawiono wszystkie uzyskane maksymalne rezultaty metod hybrydowych dla wszystkich wyselekcjonowanych za pomocą metod wbudowanych podzbiorów, uwzględniając wielkość wyselekcjonowanego podzbioru genów, liczbę iteracji oraz wielkość zainicjalizowanej populacji rozwiązań danego algorytmu opakowującego. Zgodnie z uzyskanymi rezultatami, tak samo jak w przypadku metod hybrydowych korzystających z metod filtrujących, trudno jest jednoznacznie wskazać wpływ badanych wartości parametrów metod opakowujących ze względu na dużą różnorodność parametrów dla maksymalnych wartości wydajności. Jednakże, zauważalna jest zależność dla algorytmów ABC i PSO, że im większe są wartości iteracji algorytmów, tym mniejsze zainicjalizowane populacje rozwiązań są potrzebne w celu osiągnięcia maksymalnych wartości wydajności modeli hybrydowych, niezależnie od wcześniej zastosowanej metody wbudowanej. Dodatkowo zauważono, że w przypadku metod hybrydowych wykorzystujących podzbiory wyselekcjonowane w oparciu o wbudowaną metodę selekcji SVM-RFE, algorytmy opakowujące potrzebują mniejszej ilości iteracji w celu osiągnięcia maksymalnych wartości wydajności modeli w porównaniu do pozostałych metod opakowujących. Może być to związane z tym, że w przypadku selekcji za pomocą wbudowanej metody SVM-RFE wyselekcjonowane podzbiory są najmniejszymi zbiorami ze wszystkich dla każdego z badanych zbiorów. Ponadto, w odniesieniu do zastosowanych algorytmów opakowujących zauważać można, że w metodach hybrydowych wykorzystujących metody wbudowane, zastosowanie algorytmów opakowujących ABC lub PSO w porównaniu do algorytmu GA pozwala uzyskać nieco wyższe wydajności modeli przy jednoczesnym wyselekcjonowaniu odpowiednio mniejszej ilości genów w podzbiorach.



Rysunek 43. Miara F1 zastosowanych metod hybrydowych (metody wbudowane + opakowujące)

Tabela 9. Uzyskane maksymalne rezultaty wydajności metod hybrydowych dla wszystkich wyselekcjonowanych za pomocą metod wbudowanych podzbiorów

Metoda wbudowana						Metoda opakowująca					
Metoda	Zbiór danych	Liczba wyselekcjonowanych genów	Wartość lambda / Liczba drzew	Klasyfikator uzyskujący najlepsze wyniki	Uzyskana wartość miary F1	Metoda	Liczba iteracji algorytmu	Wielkość zainicjalizowanej populacji rozwiązań	Uzyskana wartość miary F1	Wielkość wyselekcjonowanego podzbioru genów	
Ridge	<i>Lung-Cancer</i>	500	10 ⁻²	SVM	0,95	GA	500	-	0,92	257	
						ABC	500	10	0,92	271	
						PSO	100	1000	0,95	255	
Ridge	<i>BarretNP</i>	500	10	kNN	0,64	GA	50	-	0,56	245	
						ABC	50	1000	0,65	8	
						PSO	50	10	0,67	251	
Ridge	<i>GLA-BRA-180</i>	1000	5	kNN	0,79	GA	100	-	0,75	473	
						ABC	200	100	0,81	86	
						PSO	500	50	0,80	406	
RF	<i>Lung-Cancer</i>	500	1000	SVM	0,95	GA	1000	-	0,95	283	
						ABC	1000	50	0,95	259	
						PSO	1000	20	0,95	253	
RF	<i>BarretNP</i>	200	500	SVM	0,70	GA	1000	-	0,59	114	
						ABC	500	20	0,54	102	
						PSO	200	20	0,59	101	

RF	<i>GLA-BRA-180</i>	500	50	kNN	0,86	GA ABC PSO	100 1000 200	- 10 200	0,89 0,86 0,86	234 157 198
SVM-RFE	<i>Lung-Cancer</i>	50	-	SVM	0,98	GA ABC PSO	100 50 50	- 100 1000	0,98 0,98 0,98	48 35 28
SVM-RFE	<i>BarretNP</i>	50	-	RF	0,68	GA ABC PSO	200 200 50	- 20 100	0,60 0,65 0,68	48 34 22
SVM-RFE	<i>GLA-BRA-180</i>	100	-	RF	0,72	GA ABC PSO	100 100 50	- 200 1000	0,72 0,72 0,72	80 73 48

Analiza wyników wskazuje, że metody hybrydowe wykorzystujące podzbiory wcześniej wyselekcjonowanych metod wbudowanych uzyskują ogólnie gorsze wyniki w porównaniu do metod hybrydowych korzystających z podzbiorów wyselekcjonowanych metodami filtrującymi. Jedynie dla największego badanego zbioru danych *GLA-BRA-180*, zauważono, że metoda hybrydowa wykorzystująca metodę wbudowaną osiągnęła nieznaczną, kilkuprocentową przewagę nad maksymalną wydajnością metod hybrydowych wykorzystujących metody filtrujące.

Ustalenie optymalnych parametrów dla algorytmów opakowujących wykorzystanych w metodach hybrydowych jest bardzo trudne do osiągnięcia. W przypadku metod hybrydowych wykorzystujących podzbiory wyselekcjonowane wcześniej za pomocą metod filtrujących wartości parametrów algorytmów opakowujących zależą od badanego zbioru oraz od zastosowanej metody filtrującej. Z drugiej strony wartości optymalnych parametrów metod hybrydowych wykorzystujących podzbiory wyselekcjonowane za pomocą metod wbudowanych zależą w mniejszym stopniu od struktury badanych zbiorów danych, a większe znaczenie mają wzajemne wartości optymalnych parametrów. Jednak mimo tego, należy zaznaczyć, że każdy zbiór jest unikatowy pod względem charakterystyki. Dlatego też, nawet jeśli dana metoda osiągnie dobre rezultaty na jednym zbiorze, niekoniecznie będzie równie skuteczna na innym. Sprawia to, że dopasowanie parametrów algorytmów opakowujących staje się problematyczne, jednak warto jest dążyć do jego rozwiązania. Ta praca pokazała, że nawet dla tak trudnych zbiorów danych jak *BarretNP* oraz *GLA-BRA-180*, możliwe było w obu przypadkach osiągnięcie wydajności wynoszącej blisko 90% (odpowiednio rys. 37 i 43) poprzez zastosowanie odpowiednich metod hybrydowych z optymalnymi parametrami algorytmu. Te wyniki podkreślają znaczenie starannego dostosowywania parametrów do specyfiki konkretnego zadania klasyfikacji, a szczególnie klasyfikacji wieloklasowej.

6 Wnioski i plan dalszych badań

Podsumowując niniejszą pracę, należy podkreślić złożoność przeprowadzonych obliczeń. W miarę postępów realizacji okazało się, że obliczenia stały się na tyle czasochłonne, iż niezbędne było wykorzystanie klastra obliczeniowego, który pozwolił na przeprowadzenie około 2000 kombinacji obliczeń, doprowadzając w skrajnym przypadku do jednoczesnego wykorzystania 190 rdzeni procesora.

Na podstawie przeprowadzonych badań można wysnuć wnioski, dotyczące oceny poszczególnych metod selekcji najbardziej informacyjnych genów oraz wpływu badanych parametrów zastosowanych w tych metodach. W szczególności stwierdzono, że:

- Metody filtrujące wymagają większej ilości genów w porównaniu do metod wbudowanych, w celu osiągnięcia tak samo wysokich wydajności modeli dla złożonych zbiorów danych.
- Dla metody wbudowanej regresji grzbietowej stwierdzono, że wraz ze wzrostem złożoności zbioru wymagana jest wyższa wartość parametru regularizacyjnego λ w celu osiągnięcia wyższych wyników wydajności.
- Zastosowanie algorytmów opakowujących w metodach hybrydowych na wcześniej wyselekcyjowanych na podstawie metod filtrujących podzbiorach pozwala zwiększyć wydajność modelu przy jednoczesnym ograniczeniu wielkości podzbioru.
- Stwierdzono, że parametry algorytmów opakowujących w metodach hybrydowych zastosowanych na wcześniej wyselekcyjowanych podzbiorach w oparciu o metody filtrujące zależą zarówno od struktury badanego zbioru, jak i od zastosowanej wcześniej metody filtrującej.
- Wyselekcyjowanie podzbiorów o większej ilości genów przez metody filtrujące, w porównaniu do mniejszej ilości genów wyselekcyjowanych w oparciu o metody wbudowane ma lepszy wpływ na rezultaty metod hybrydowych.
- Stwierdzono, że parametry algorytmów opakowujących sztucznej kolonii pszczół (ABC) i roju częstek (PSO) w metodach hybrydowych zastosowanych na wyselekcyjowanych podzbiorach przy wykorzystaniu metod wbudowanych, są uzależnione od wzajemnych optymalnych wartości parametrów: im większa jest liczba iteracji, tym wymagana jest mniejsza populacja zainicjalizowanych rozwiązań algorytmu.
- Metody opakowujące wykorzystujące podzbiory wyselekcyjowane w oparciu o metodę wbudowaną SVM-RFE nie są w stanie poprawić wydajności modeli, ze względu na bardzo ograniczoną liczbę genów w tych podzbiorach.
- Zgodnie z uzyskanyimi rezultatami, nie można wskazać jednej uniwersalnej metody selekcji najbardziej informacyjnych cech, która byłaby optymalna dla wszystkich zbiorów danych, a jej wybór zależy od specyfiki danego zbioru.

Uzyskane wyniki stanowią podstawę do wskazania dalszych badań. Biorąc po uwagę fakt, że przy przeprowadzeniu badań o bardzo dużej liczbie kombinacji parametrów metod selekcji najbardziej informacyjnych genów udało się osiągnąć satysfakcyjujące rezultaty wydajności dla tak złożonych zbiorów danych, obiecującym planem byłoby zastosowanie optymalizacji globalnej tych parametrów, co w znaczny sposób mogłoby przyspieszyć poszukiwania optymalnych wartości. Oprócz tego, zastosowanie bardziej agresywnego podejścia redukcji wymiarowości dla metod hybrydowych, mogłoby stanowić odpowiednią podstawę do zrozumienia znaczenia wybranych genów u osób zajmujących się doborem odpowiednich terapii w leczeniu schorzeń. Dodatkowo istotnym elementem badań byłaby optymalizacja parametrów wykorzystanych klasyfikatorów, co mogłoby w znaczący sposób wpływać na uzyskane rezultaty wydajności.

7 Bibliografia

- [1] Cancer: Key Facts, World Health Organization, 2022 [dostęp 25 lipca 2023], Dostępny w Internecie: www.who.int/news-room/fact-sheets/detail/cancer
- [2] Differences in cancer incidence and mortality across the globe, World Cancer Research Fund International, 2023 [dostęp 25 lipca 2023], Dostępny w Internecie: www.wcrf.org/differences-in-cancer-incidence-and-mortality-across-the-globe/
- [3] Chial H., Genetic regulation of cancer, Nature Education, 2008, 1, 67.
- [4] Hipfner D., Cohen S., Connecting proliferation and apoptosis in development and disease, Nature Reviews Molecular Cell Biology, 2004, 5, 805.
- [5] Wang Z., Gerstein M., Snyder M., RNA-Seq: a revolutionary tool for transcriptomics, Nature Reviews Genetics, 2009, 10, 57.
- [6] Aebersold R., Mann M., Mass spectrometry-based proteomics, Nature, 2003, 422, 198.
- [7] The Structure, Function, and Applications of GeneChip Microarrays, Affymetrix, 2005.
- [8] Manufacturing of GeneChip Microarrays and Building Models, Affymetrix, 2005.
- [9] Dalma-Weiszhausz D.D. I in., The Affymetrix GeneChip® Platform: An Overview, Methods in Enzymology, 2006, 410, 3.
- [10] Saeys Y., Inza I., Larrañaga P., A review of feature selection techniques in bioinformatics, Bioinformatics, 2007, 23, 2507.
- [11] Almazrua H., Alshamlan H., A Comprehensive Survey of Recent Hybrid Feature Selection Methods in Cancer Microarray Gene Expression Data, IEEE, 2022, 10, 71427.
- [12] Shukla A. K., Tripathi D., Identification of potential biomarkers on microarray data using distributed gene selection approach, Mathematical Biosciences, 2019, 315.
- [13] What is mutual information?, QuantDare, 2021 [dostęp 5 sierpnia 2023], Dostępny w Internecie: www.quantdare.com/what-is-mutual-information/
- [14] Ghosh K. K. i in., Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data, Expert Systems with Applications, 2021, 169.
- [15] Pirooznia M., Yang J.Y., Yang M.Q., A comparative study of different machine learning methods on microarray gene expression data, BMC Genomics, 2008, 9.
- [16] Kumar M. i in., Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor, Procedia Computer Science, 2015, 54, 301.
- [17] Cuevas A., Febrero M., Fraiman R., An anova test for functional data, Computational Statistics & Data Analysis, 2004, 47, 111.
- [18] Kohavi R., John G. H., Wrappers for feature subset selection, Artificial Intelligence, 1997, 97, 273.
- [19] Nicholas P. i in., A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction, Frontiers in Bioinformatics, 2022, 2.
- [20] McCall J., Genetic Algorithms for Modelling and Optimisation, Journal of Computational and Applied Mathematics, 2005, 184, 205.
- [21] Babatunde O. H. i in., A Genetic Algorithm-Based Feature Selection, International Journal of Electronics Communication and Computer Engineering, 2014, 5, 899.
- [22] Karaboga D., AN IDEA BASED ON HONEY BEE SWARM FOR NUMERICAL OPTIMIZATION, Raport techniczny, Uniwersytet Eryciyes, 2005.
- [23] Servet K. M., Babalik A., Improved Artificial Bee Colony Algorithm for Continuous Optimization Problems, Journal of Computer and Communications, 2014, 2, 108.
- [24] Karaboga D., Basturk B., On the Performance of Artificial Bee Colony (ABC) Algorithm, Applied Soft Computing, 2008, 8, 687.

- [25] Tomera M. Zastosowanie algorytmów rojowych do optymalizacji parametrów w modelach układów regulacji, Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej, 2015, 46, 97.
- [26] Ye Z. i in., Feature Selection Based on Adaptive Particle Swarm Optimization with Leadership Learning, Computational intelligence and neuroscience, 2022, 1.
- [27] He Y. i in., The Parameters Selection of PSO Algorithm influencing In performance of Fault Diagnosis, 2016, 63, 2019.
- [28] Ahmad I., Feature Selection Using Particle Swarm Optimization in Intrusion Detection, International Journal of Distributed Sensor Networks, 2015, 11.
- [29] Ardjani F., Sadouni K., Benyettou M., Optimization of SVM MultiClass by Particle Swarm (PSO-SVM), 2010 2nd International Workshop on Database Technology and Applications, 2010.
- [30] Chandrashekhar G., Sahin F., A survey on feature selection methods, Computers & Electrical Engineering, 2014, 40, 16.
- [31] Debjani P. i in., Predictive Systems: Role of Feature Selection in Prediction of Heart Disease, Journal of Physics, 2019, 1, 1372.
- [32] Dissanayake K., Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms, Applied Computational Intelligence and Soft Computing, 2021, 1, 1.
- [33] Muthukrishnan R., Rohini R., LASSO: A feature selection technique in predictive modeling for machine learning, IEEE International Conference on Advances in Computer Applications (ICACA), 2016.
- [34] Płoński P., Zastosowanie wybranych metod przekształcania i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji, Rozprawa doktorska, Politechnika Warszawska, 2016.
- [35] Díaz-Uriarte R., Alvarez de Andrés S., Gene selection and classification of microarray data using random forest, BMC Bioinformatics, 2006, 7, 3.
- [36] Qi Y., Random Forest for Bioinformatics, Ensemble Machine Learning, 2012.
- [37] Rustam Z., Kharis S. A. A., Comparison of Support Vector Machine Recursive Feature Elimination and Kernel Function as Feature Selection Using Support Vector Machine for Lung Cancer Classification, Journal of Physics, 2020, 1442.
- [38] Tang Y., Zhang Y.-Q., Huang Z., Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4, 365.
- [39] Yuanyuan D., Wilkins D., Improving the Performance of SVM-RFE to Select Genes in Microarray Data, BMC Bioinformatics, 2006, 7, 12.
- [40] Kumari B., Swarnkar T., Filter versus wrapper feature subset selection in large dimensionality micro array: A review, International Journal of Computer Science and Information Technologies, 2011, 2, 1048.
- [41] Joyce J., Bayes' Theorem, The Stanford Encyclopedia of Philosophy, 2021.
- [42] Parsian M., Data Algorithms: Recipes for Scaling Up With Hadoop Spark, O'Reilly Media, 2015.
- [43] Tan P.-N i in., Support vector machine (SVM), Introduction to Data Mining, 2006.
- [44] Geron A., Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (3rd. ed.), O'Reilly Media, 2022.
- [45] Shadeed I., Alwan J., Abd D., The effect of gamma value on support vector machine performance with different kernels, International Journal of Electrical and Computer Engineering (IJECE), 2020, 5, 10.
- [46] Telnoni P. A., Budiawan R., Qana'a M., Comparison of Machine Learning Classification Method on Text-based Case in Twitter, 2019 International Conference on ICT for Smart Society (ICISS), 2019.

- [47] Markoulidakis I. i in., Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem, *Technologies*, 2021, 9, 81.
- [48] Accuracy, precision, and recall in multi-class classification, *Evidently AI*, 2022 [dostęp 12 października 2023], Dostępny w Internecie: www.evidentlyai.com/classification-metrics/multi-class-metrics
- [49] Margherita G., Bagli E., Visani G., Metrics for Multi-Class Classification: An Overview, *ArXiv*, 2020.
- [50] Understanding Micro and Macro Averages in Multiclass Multilabel Problems, Krystian's Safjan Blog, 2021 [dostęp 12 października 2023], Dostępny w Internecie: www.safjan.com/micro-and-macro-averages-in-multiclass-multilabel-problems/
- [51] Takahashi K., Yamamoto K., Kuchiba A., Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores, *Applied Intelligence*, 2022, 52, 4961.
- [52] Changyong F. i in., Log-transformation and its implications for data analysis, *Shanghai archives of psychiatry*, 2014, 26, 105.
- [53] Aittokallio T., Dealing with missing values in large-scale studies: microarray data imputation and beyond, *Briefings in Bioinformatics*, 2010, 11, 253.
- [54] Statnikov A. i in., A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*, 2005, 21, 631.
- [55] Chang C.-C., Lin C.-J., LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2007, 27, 3.
- [56] Oshiro T., Perez P., Baranauskas J. A., How Many Trees in a Random Forest? *MLDM*, 2012, 7376, 154.
- [57] Uddin S., Haque I., Lu H., Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction, *Scientific Reports*, 2022, 12.

8 Spis rysunków

Rysunek 1. Schemat budowy mikromacierzy [7].....	5
Rysunek 2. Proces wytwarzania mikromacierzy [8]	6
Rysunek 3. Schemat przebiegu pomiarów macierzowych [8].....	8
Rysunek 4. Przykładowe dane mikromacierzowe.....	8
Rysunek 5. Schemat metod filtrujących.....	10
Rysunek 6. Schemat metod opakowujących	13
Rysunek 7. Schemat algorytmu genetycznego	15
Rysunek 8. Schemat algorytmu sztucznej kolonii pszczół.....	18
Rysunek 9. Schemat algorytmu roju cząstek.....	20
Rysunek 10. Schemat metod wbudowanych	21
Rysunek 11. Schemat metod hybrydowych	25
Rysunek 12. Macierz pomyłek dla problemu klasyfikacji binarnej (A) oraz klasyfikacji wieloklasowej (B) [47]	29
Rysunek 13. Przykładowy podział czteroklasowego zbioru danych przedstawiający potencjalną przydział klas przewidzianych do klas rzeczywistych [48].....	29
Rysunek 14. Dokładność klasyfikatora dla przykładowego, wieloklasowego zbioru [48]	30
Rysunek 15. Precyza klasyfikatora dla danej klasy dla przykładowego, wieloklasowego zbioru [48]	31
Rysunek 16. Czułość klasyfikatora dla danej klasy dla przykładowego, wieloklasowego zbioru [48]	32
Rysunek 17. Dostępność danych dla poszczególnych klas w badanych zbiorach danych.....	36
Rysunek 18. Schemat metodyki badań.....	37
Rysunek 19. Histogram wybranego genu ze zbioru GLA-BRA-180 przed przeprowadzeniem transformacji logarytmicznej (A) i po transformacji (B).....	38
Rysunek 20. Dostępność danych dla poszczególnych klas w badanych zbiorach danych po podziale na grupy treningowe i walidacyjne	40
Rysunek 21. Schemat metodyki badań metod filtrujących	42
Rysunek 22. Schemat metodyki badań metod wbudowanych.....	43
Rysunek 23. Schemat metodyki badań metod hybrydowych	44
Rysunek 24. Miara F1 metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	47
Rysunek 25. Miara F1 badanych metod filtrujących.....	48
Rysunek 26. Średnia miara F1 metody Ridge dla określonej wartości lambda oraz zastosowanego klasyfikatora	51
Rysunek 27. Średnia miara F1 metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora.....	52

Rysunek 28. Średnia miara F1 metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora.....	54
Rysunek 29. Średnia miara F1 metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	55
Rysunek 30. Miara F1 metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora.....	57
Rysunek 31. Miara F1 dla wszystkich zastosowanych metod wbudowanych	59
Rysunek 32. Miara F1 metody hybrydowej wykorzystującej algorytm genetyczny (GA) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów	61
Rysunek 33. Średnia miara F1 metody hybrydowej wykorzystującej algorytm sztucznej kolonii pszczół (ABC) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów.....	62
Rysunek 34. Średnia miara F1 metody hybrydowej wykorzystującą algorytm sztucznej kolonii pszczół (ABC) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów	63
Rysunek 35. Średnia miara F1 metody hybrydowej wykorzystującą algorytm roju częstek (PSO) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów	64
Rysunek 36. Średnia miara F1 metody hybrydowej wykorzystującą algorytm roju częstek (PSO) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody filtrujące podzbiorów	65
Rysunek 37. Miara F1 zastosowanych metod hybrydowych (metody filtrujące + opakowujące).....	67
Rysunek 38. Miara F1 metody hybrydowej wykorzystującą algorytm genetyczny (GA) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów	71
Rysunek 39. Średnia miara F1 metody hybrydowej wykorzystującą algorytm sztucznej kolonii pszczół (ABC) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów.....	72
Rysunek 40. Średnia miara F1 metody hybrydowej wykorzystującą algorytm sztucznej kolonii pszczół (ABC) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów	73
Rysunek 41. Średnia miara F1 metody hybrydowej wykorzystującą algorytm roju częstek (PSO) dla określonej liczby iteracji na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów	74
Rysunek 42. Średnia miara F1 metody hybrydowej wykorzystującą algorytm roju częstek (PSO) dla określonej wielkości zainicjalizowanej populacji rozwiązań algorytmu na podstawie wyselekcjonowanych przez metody wbudowane podzbiorów	75
Rysunek 43. Miara F1 zastosowanych metod hybrydowych (metody wbudowane + opakowujące) ...	77
Rysunek 44. Dokładność metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora.....	91
Rysunek 45. Precyzja metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora.....	92

Rysunek 46. Czułość metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	93
Rysunek 47. Dokładność zastosowanych metod filtrujących.....	94
Rysunek 48. Precyzja zastosowanych metod filtrujących	95
Rysunek 49. Czułość zastosowanych metod filtrujących.....	96
Rysunek 50. Średnia dokładność metody regresji grzbietowej dla określonej wartości lambda oraz zastosowanego klasyfikatora	97
Rysunek 51. Średnia precyzja metody regresji grzbietowej dla określonej wartości lambda oraz zastosowanego klasyfikatora	98
Rysunek 52. Średnia czułość metody regresji grzbietowej dla określonej wartości lambda oraz zastosowanego klasyfikatora	99
Rysunek 53. Średnia dokładność metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	100
Rysunek 54. Średnia precyzja metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora.....	101
Rysunek 55. Średnia czułość metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora.....	102
Rysunek 56. Średnia dokładność metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora	103
Rysunek 57. Średnia precyzja metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora	104
Rysunek 58. Średnia czułość metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora	105
Rysunek 59. Średnia dokładność metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora.....	106
Rysunek 60. Średnia precyzja metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	107
Rysunek 61. Średnia czułość metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	108
Rysunek 62. Dokładność metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	109
Rysunek 63. Precyzja metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	110
Rysunek 64. Czułość metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora	111
Rysunek 65. Dokładność dla wszystkich zastosowanych metod wbudowanych	112
Rysunek 66. Precyzja dla wszystkich zastosowanych metod wbudowanych	113
Rysunek 67. Czułość dla wszystkich zastosowanych metod wbudowanych	114

9 Spis tabel

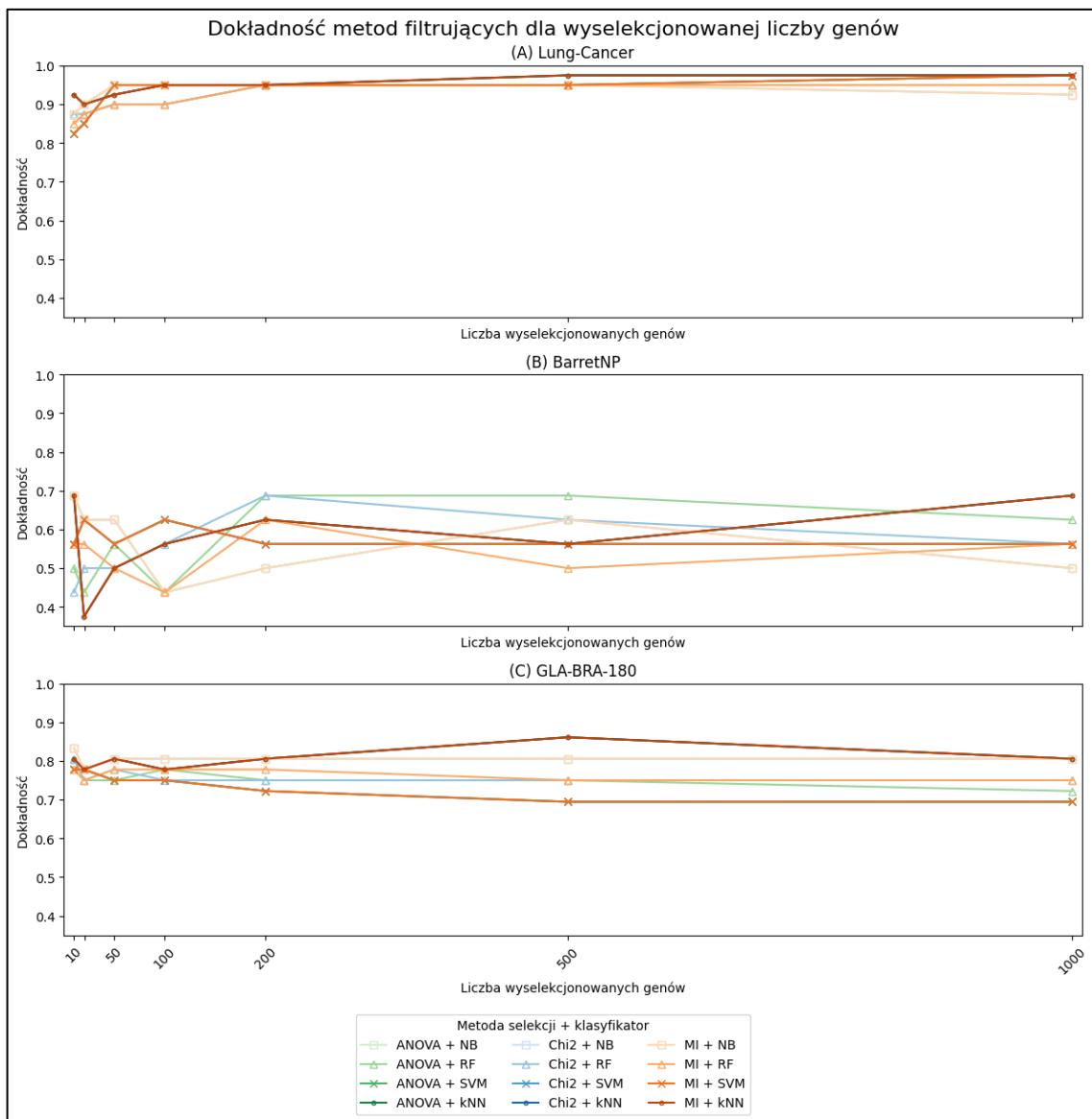
Tabela 1. Zalety i ograniczenia metod selekcji cech [11, 19, 40]	23
Tabela 2. Zestawienie wykorzystanych zbiorów danych.....	35
Tabela 3. Zestawienie badanych zbiorów danych po wstępnej selekcji cech różnicujących.....	41
Tabela 4. Wyselekcjonowane podzbiory za pomocą metod filtrujących charakteryzujące się najwyższą wydajnością.....	49
Tabela 5. Wyselekcjonowane podzbiory za pomocą metody Ridge charakteryzujące się najwyższą wydajnością.....	53
Tabela 6. Wyselekcjonowane podzbiory za pomocą metody lasu losowego charakteryzujące się najwyższą wydajnością	56
Tabela 7. Wyselekcjonowane podzbiory za pomocą metody SVM-RFE charakteryzujące się najwyższą wydajnością.....	58
Tabela 8. Uzyskane maksymalne rezultaty wydajności metod hybrydowych dla wszystkich wyselekcjonowanych za pomocą metod filtrujących podzbiorów	68
Tabela 9. Uzyskane maksymalne rezultaty wydajności metod hybrydowych dla wszystkich wyselekcjonowanych za pomocą metod wbudowanych podzbiorów	78

10 Spis załączników

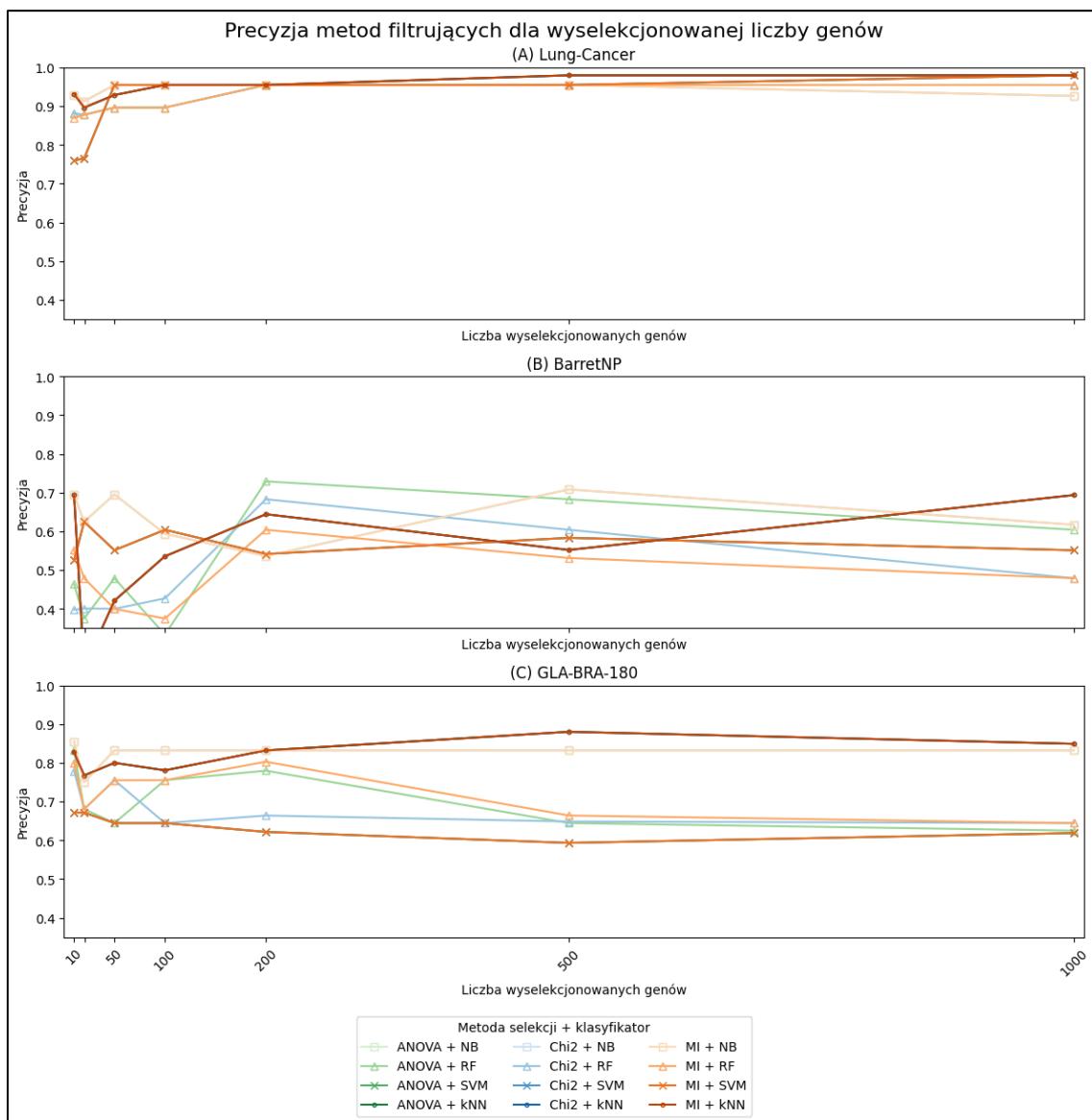
11	Załączniki	Error! Bookmark not defined.
11.1	Załącznik nr 1	Error! Bookmark not defined.
11.2	Załącznik nr 2	Error! Bookmark not defined.

11 Załączniki

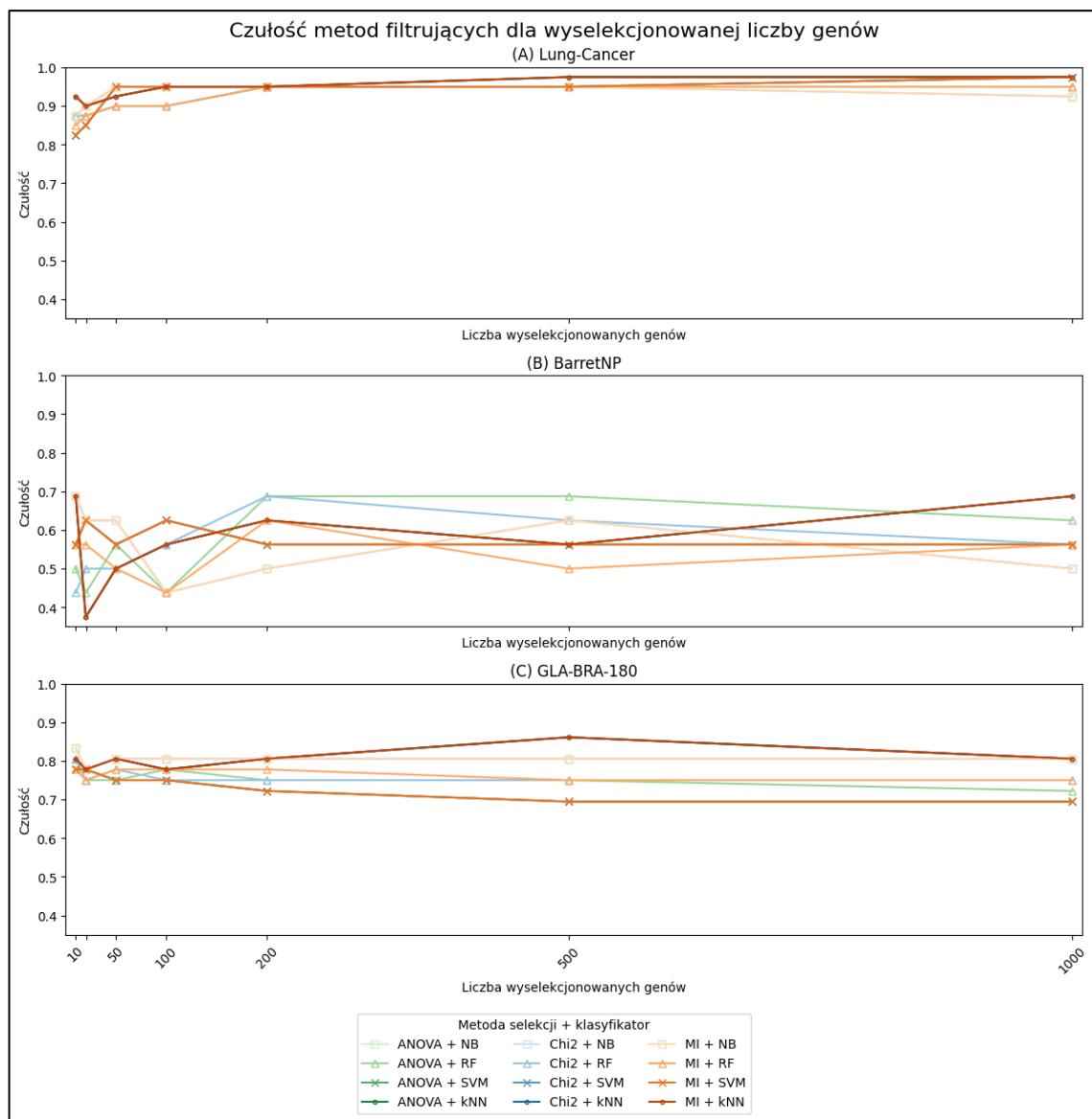
11.1 Załącznik nr 1



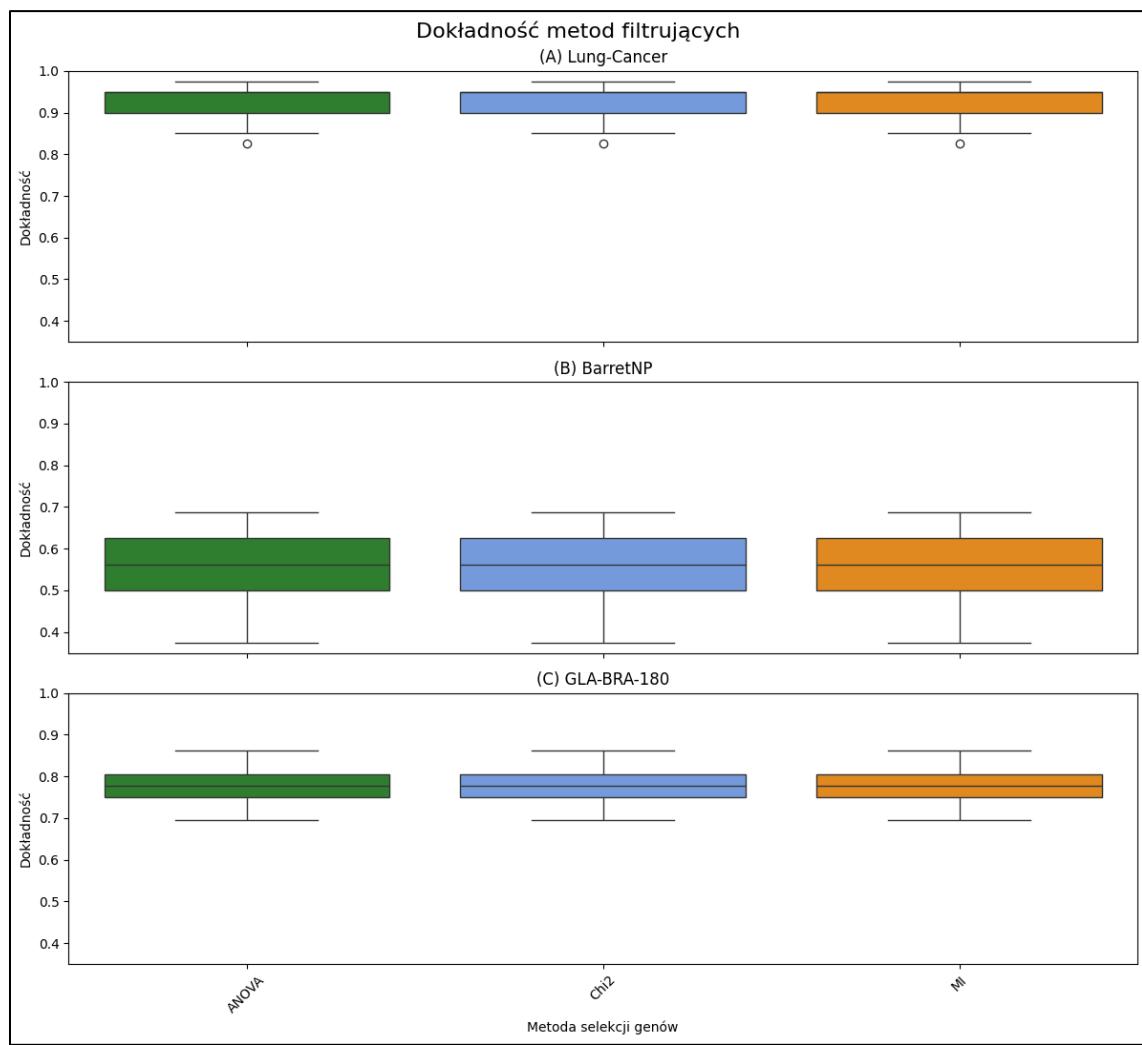
Rysunek 44. Dokładność metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



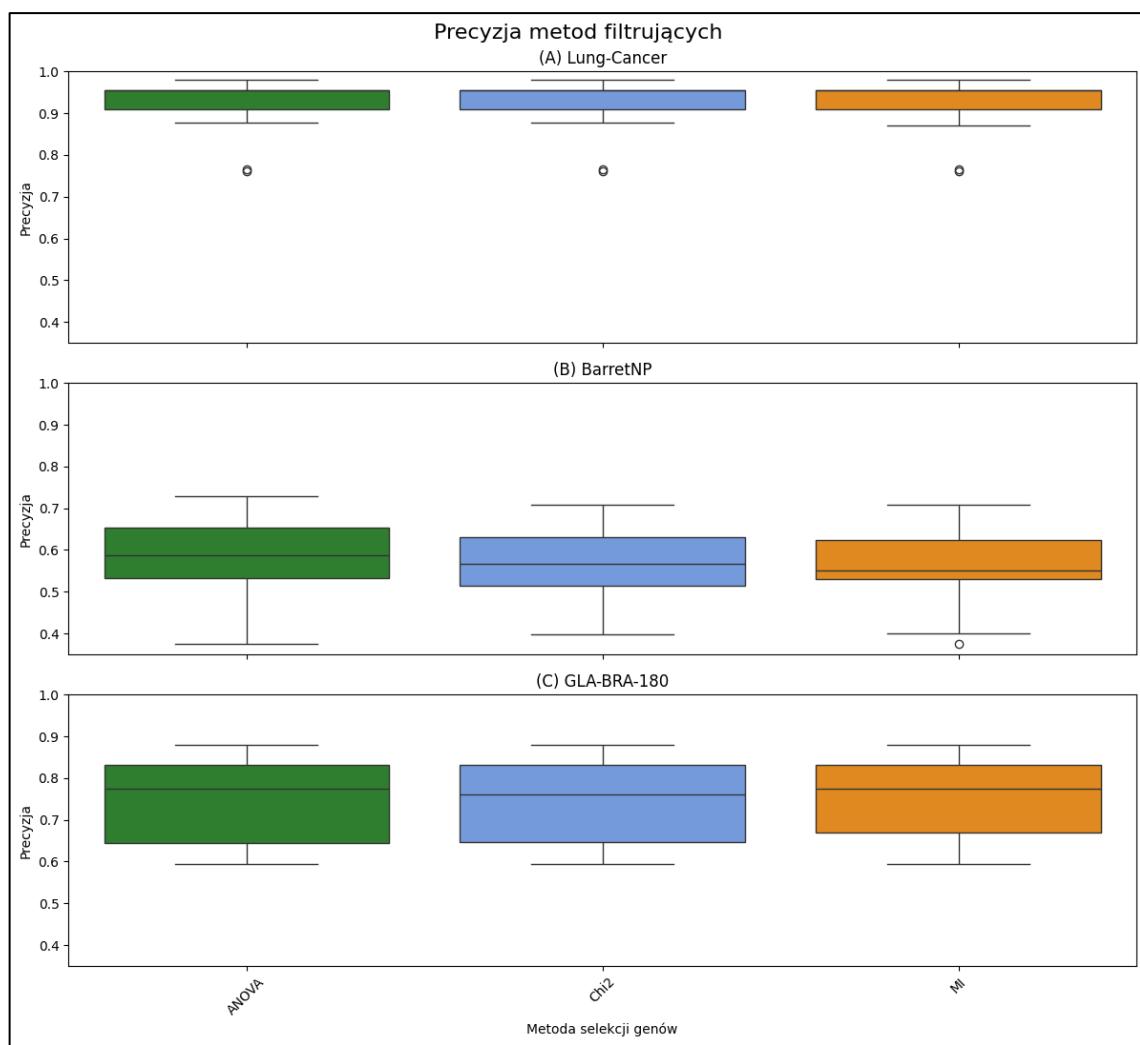
Rysunek 45. Precyza metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



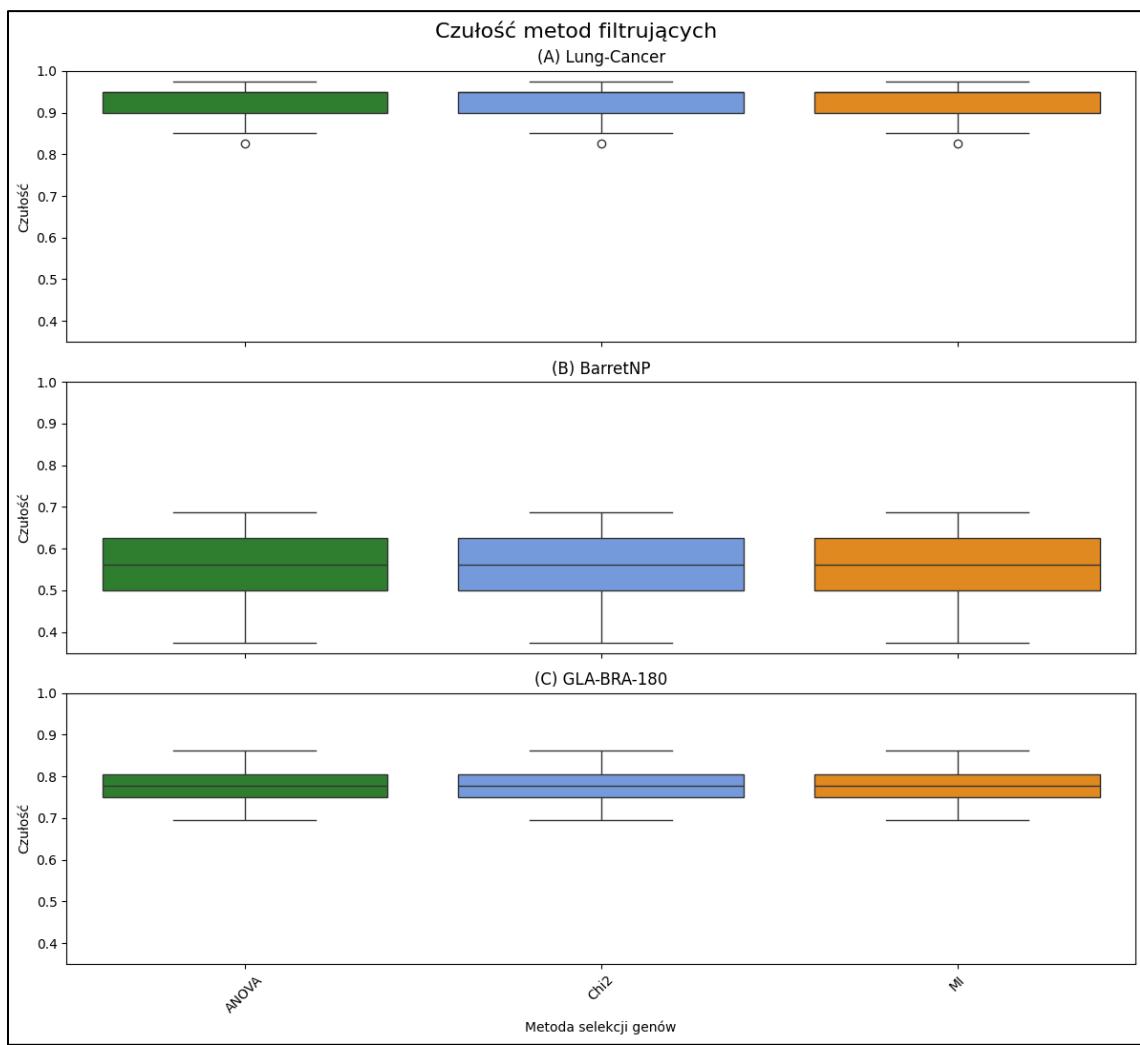
Rysunek 46. Czułość metod filtrujących dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



Rysunek 47. Dokładność zastosowanych metod filtrujących

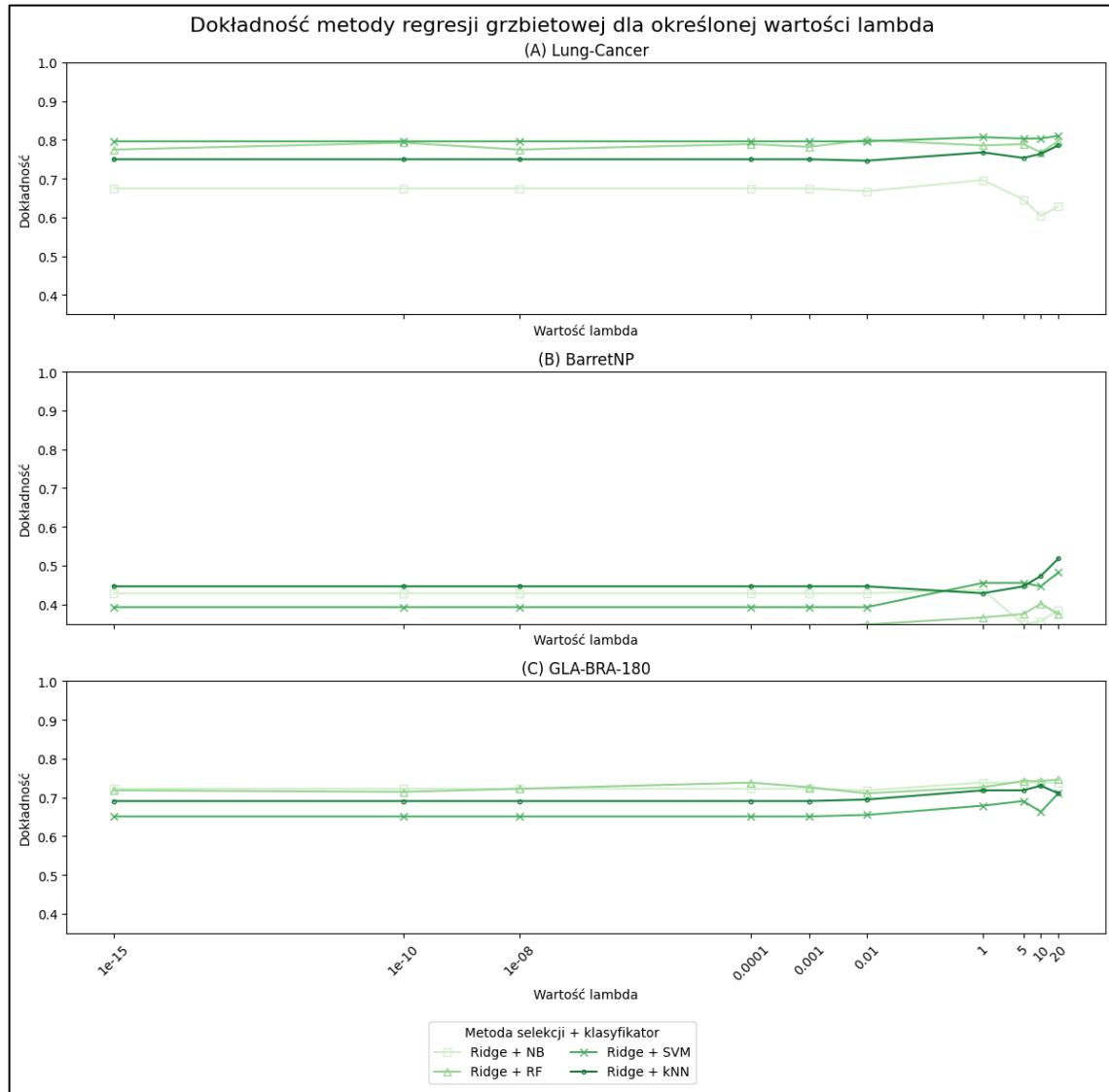


Rysunek 48. Precyzaja zastosowanych metod filtrujących

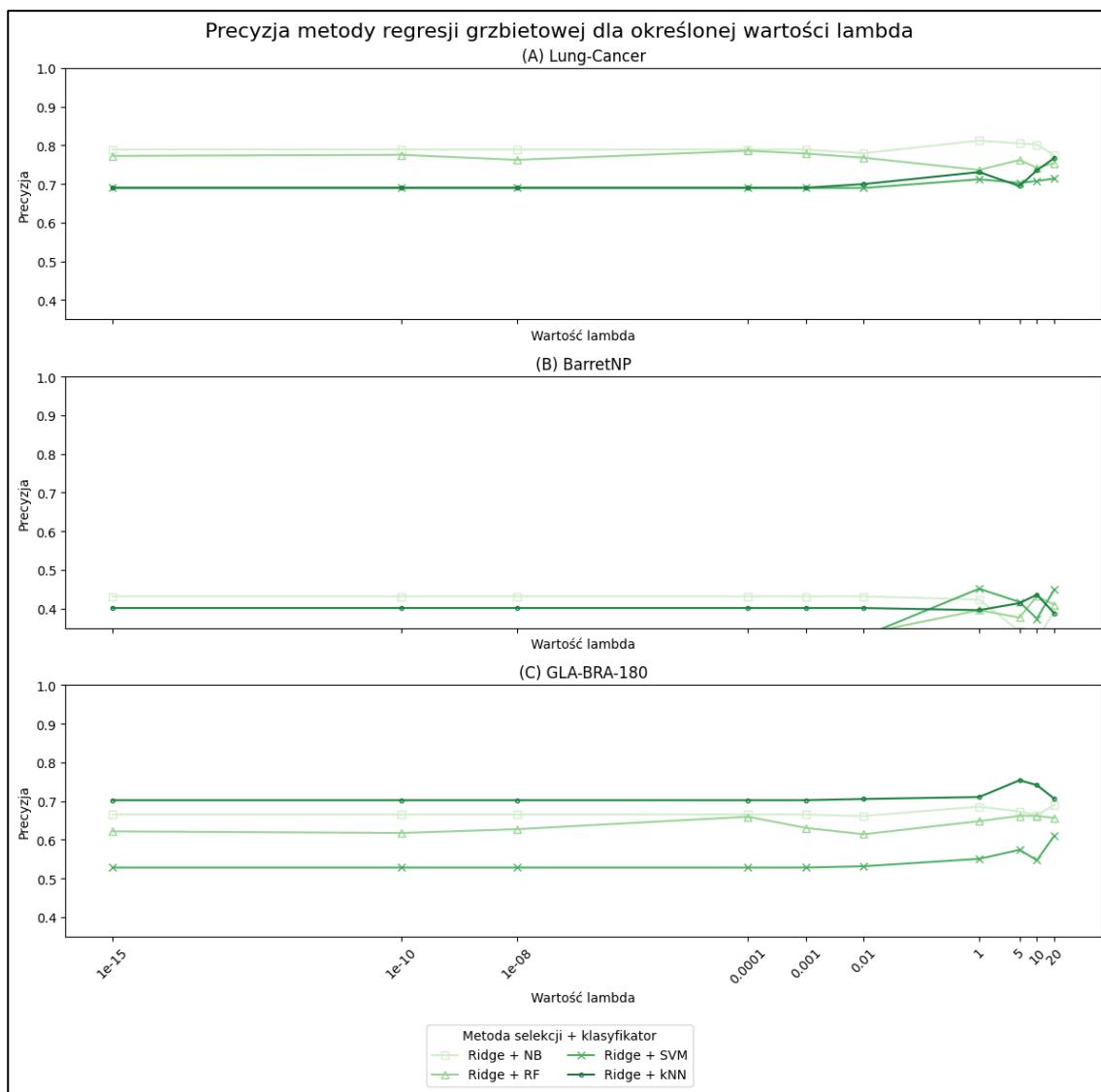


Rysunek 49. Czułość zastosowanych metod filtrujących

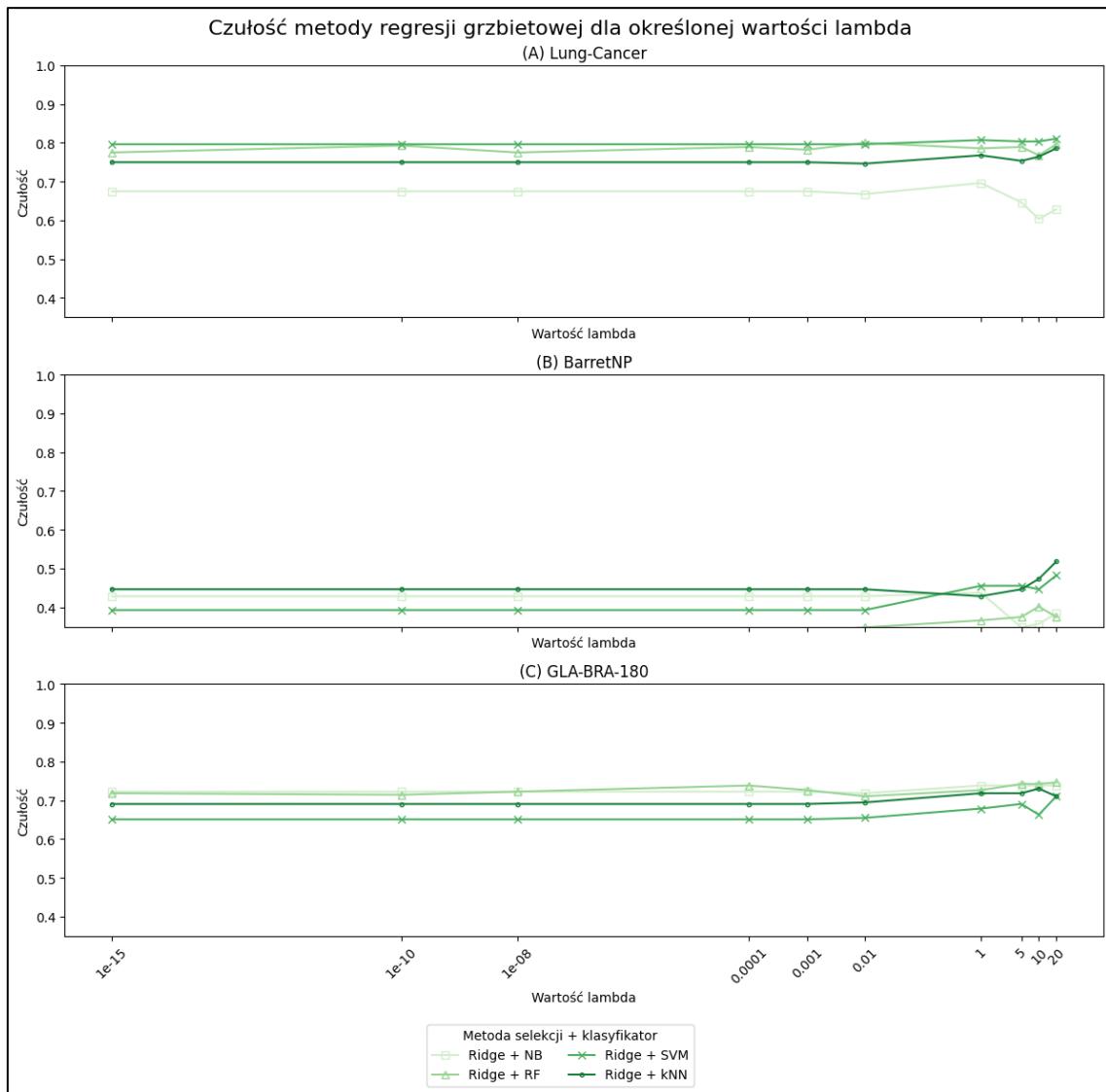
11.2 Załącznik nr 2



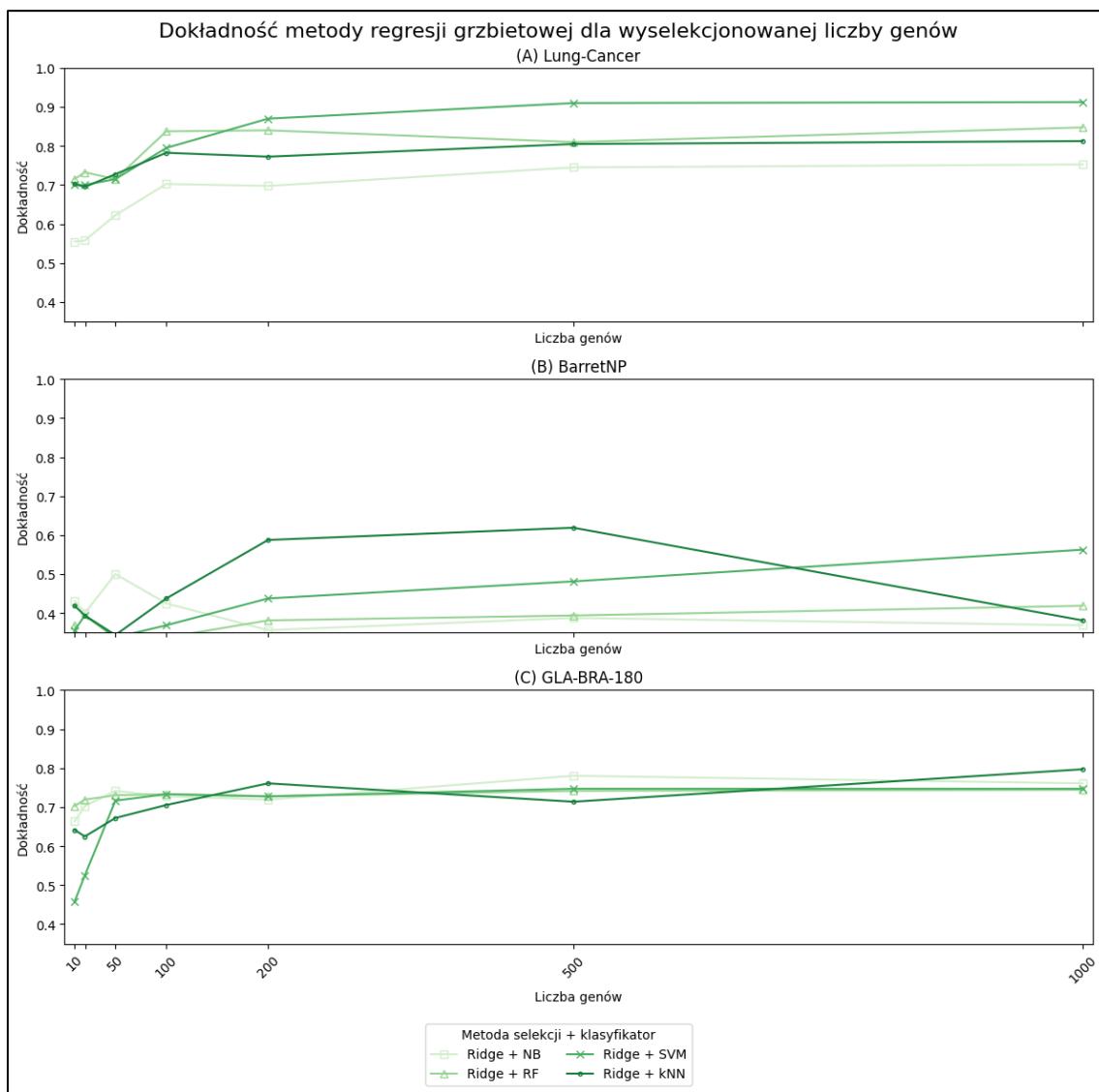
Rysunek 50. Średnia dokładność metody regresji grzbietowej dla określonej wartości lambda oraz zastosowanego klasyfikatora



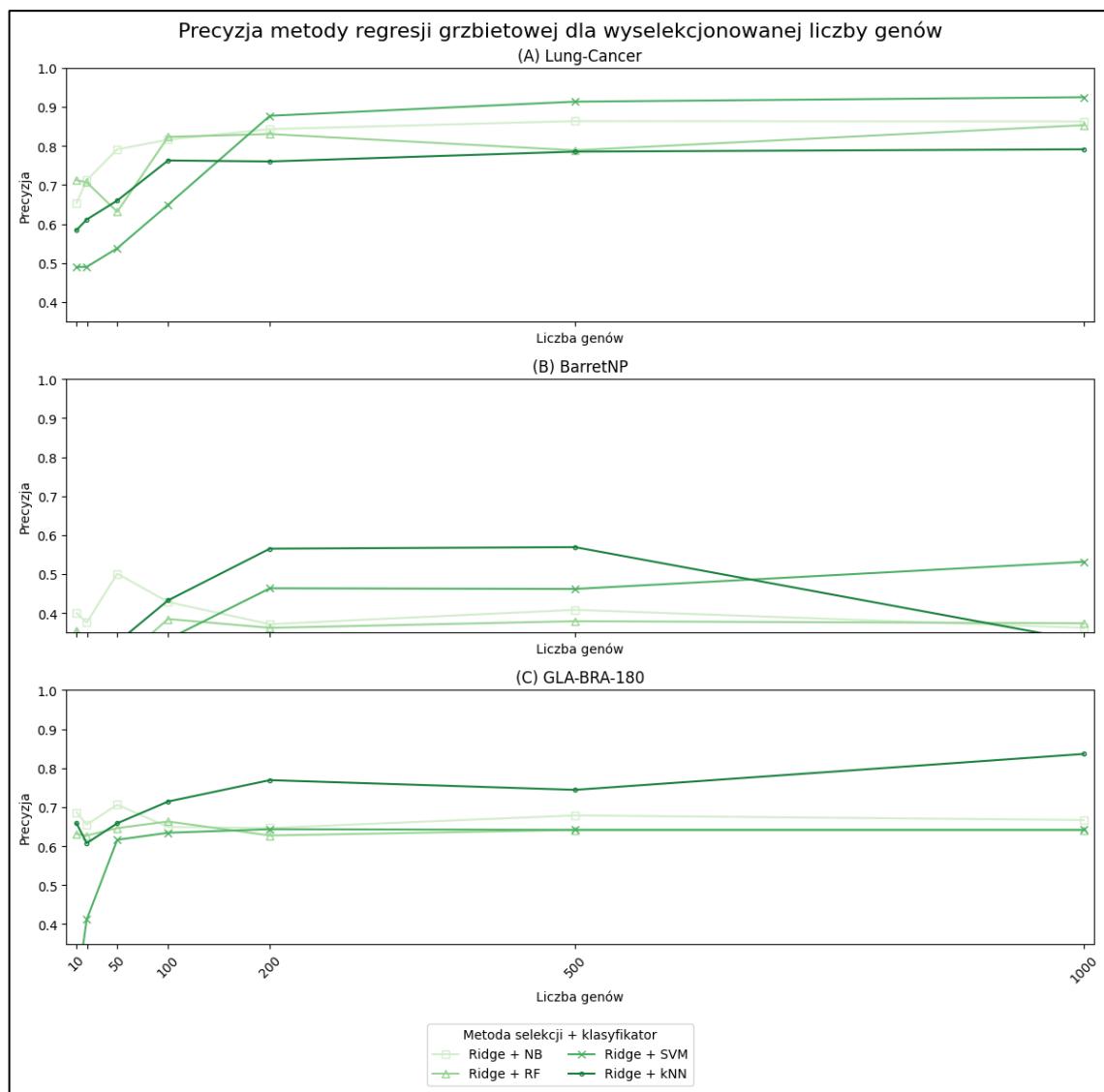
Rysunek 51. Średnia precyza metody regresji grzbietowej dla określonej wartości lambda oraz zastosowanego klasyfikatora



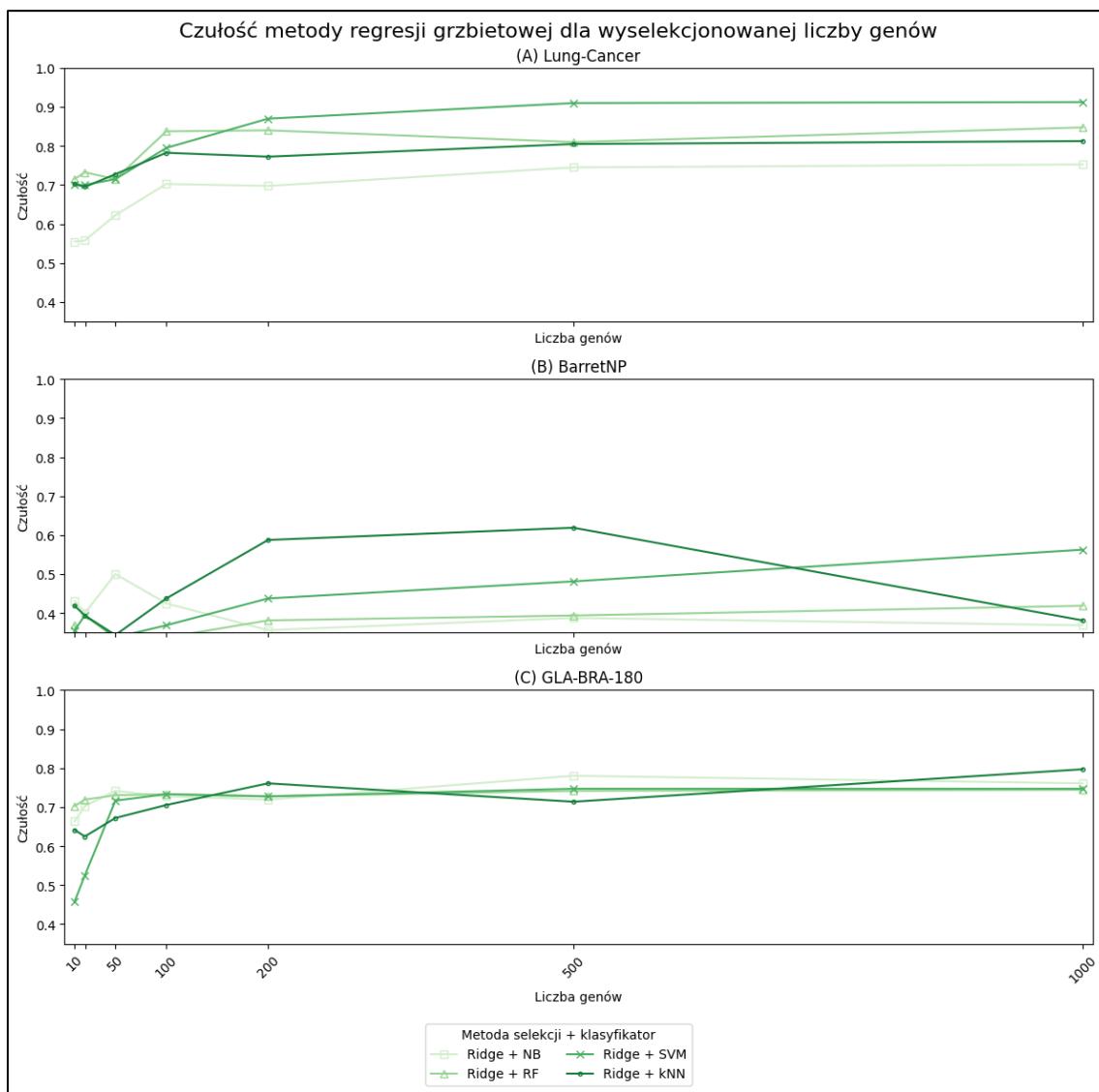
Rysunek 52. Średnia czułość metody regresji grzbietowej dla określonej wartości lambda oraz zastosowanego klasyfikatora



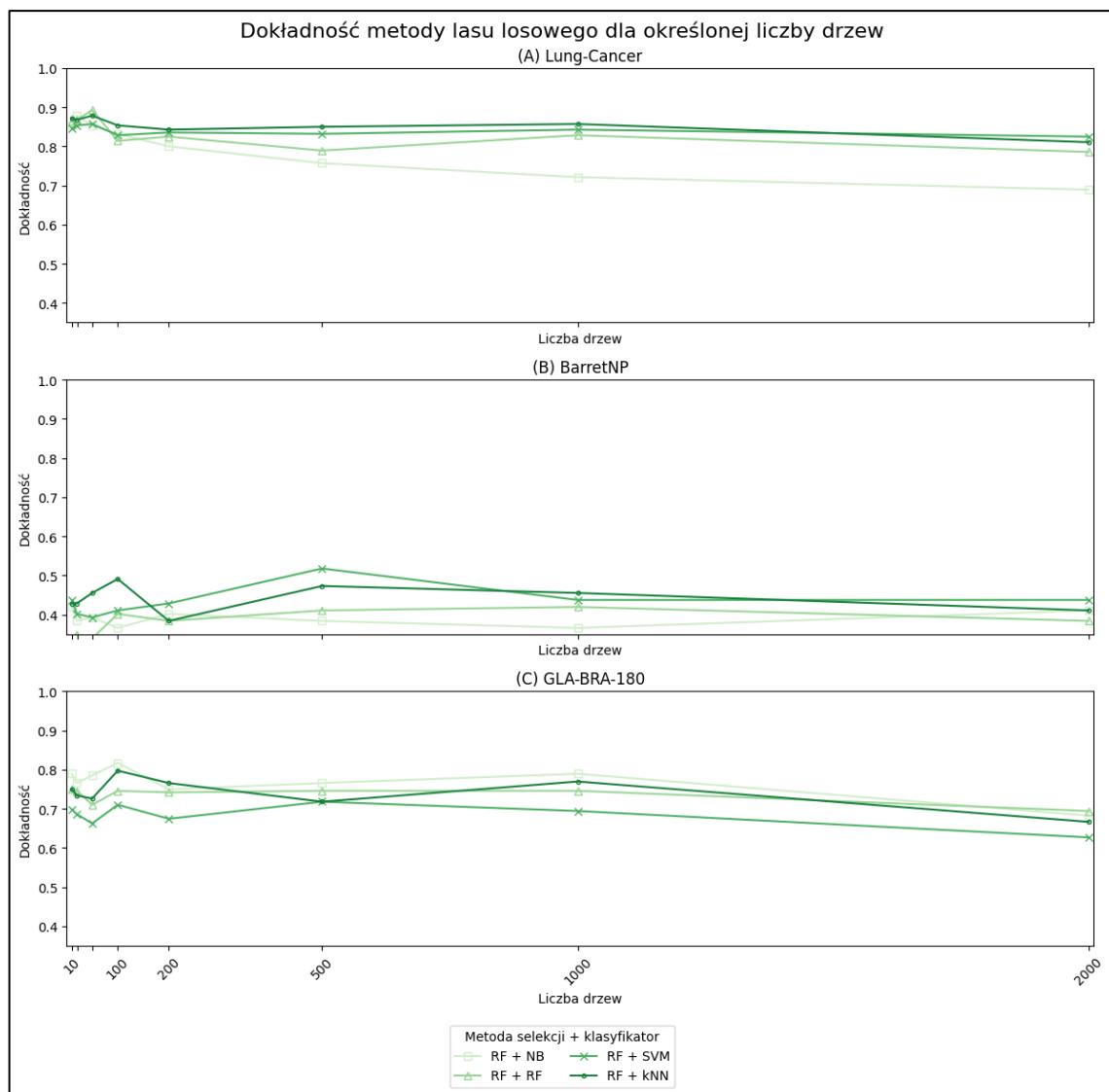
Rysunek 53. Średnia dokładność metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



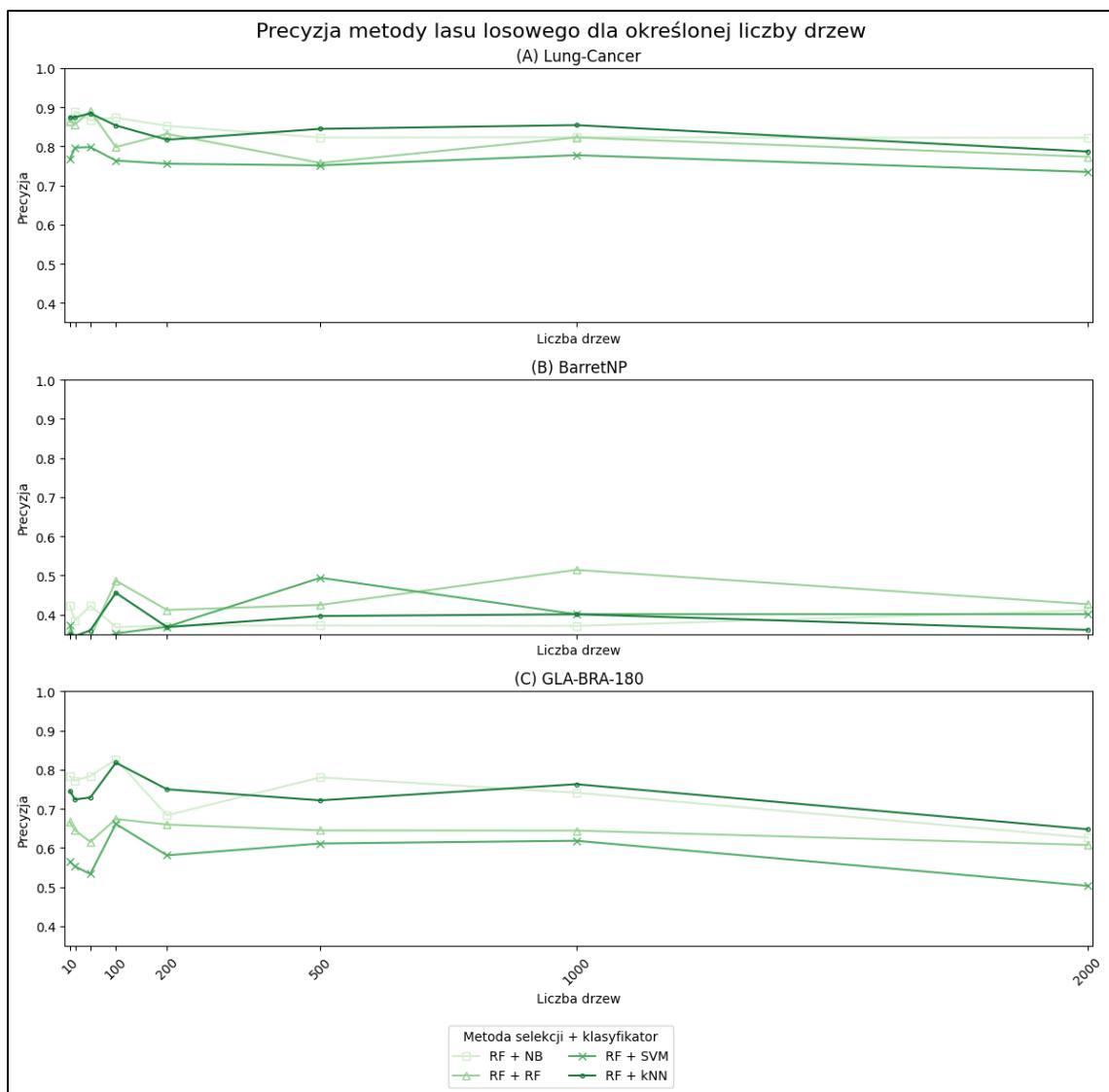
Rysunek 54. Średnia precyzaja metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



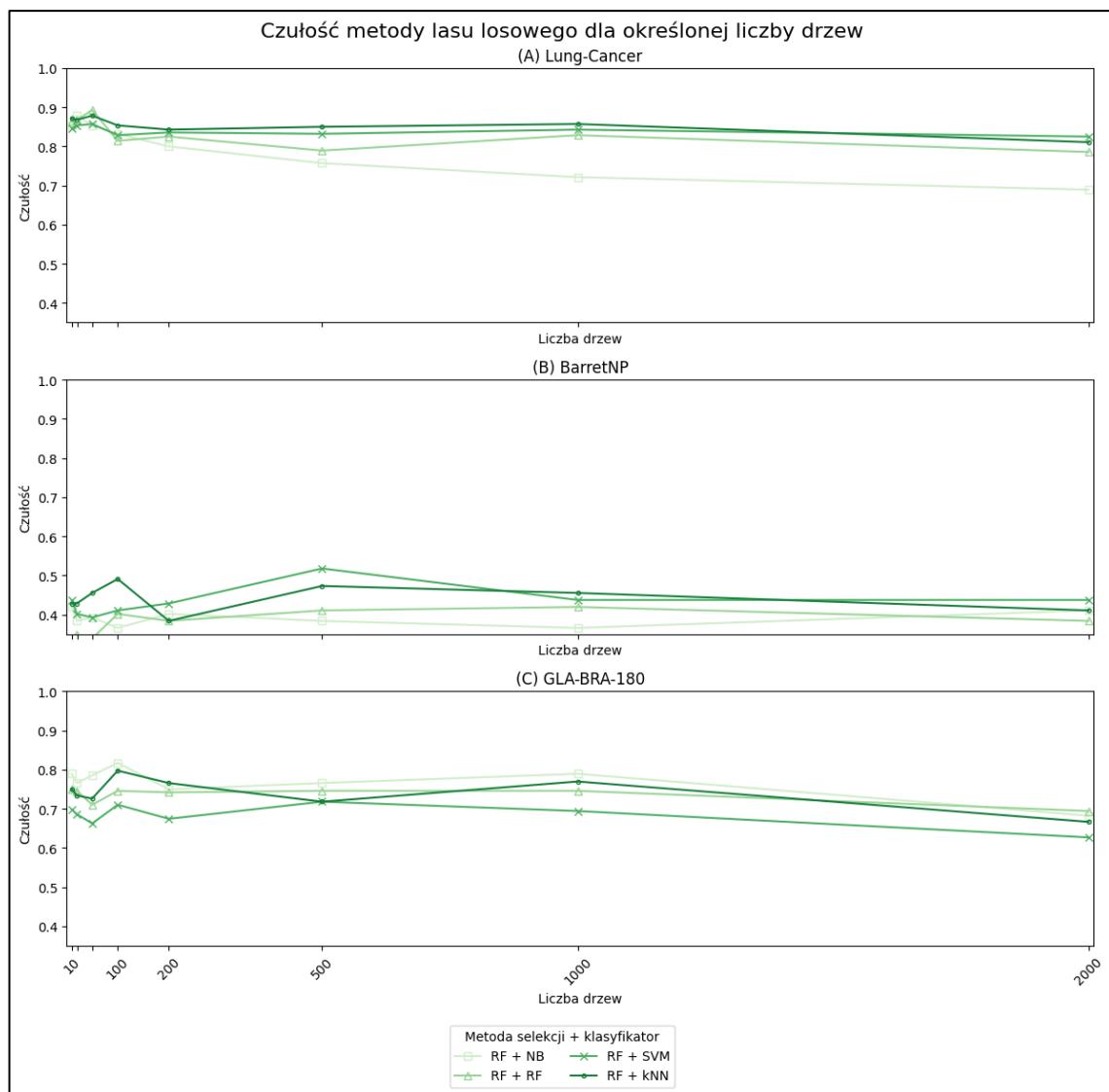
Rysunek 55. Średnia czułość metody regresji grzbietowej dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



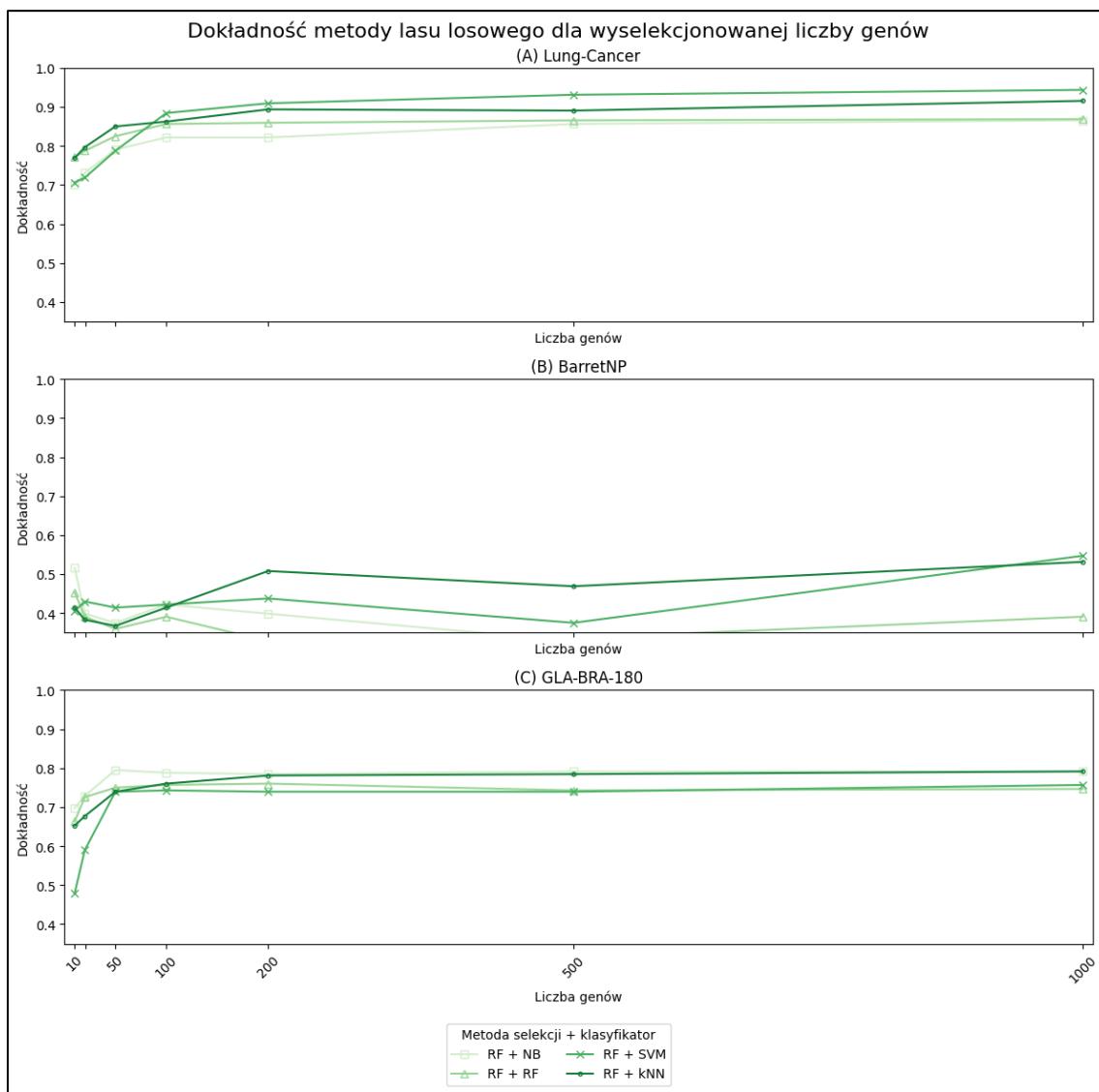
Rysunek 56. Średnia dokładność metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora



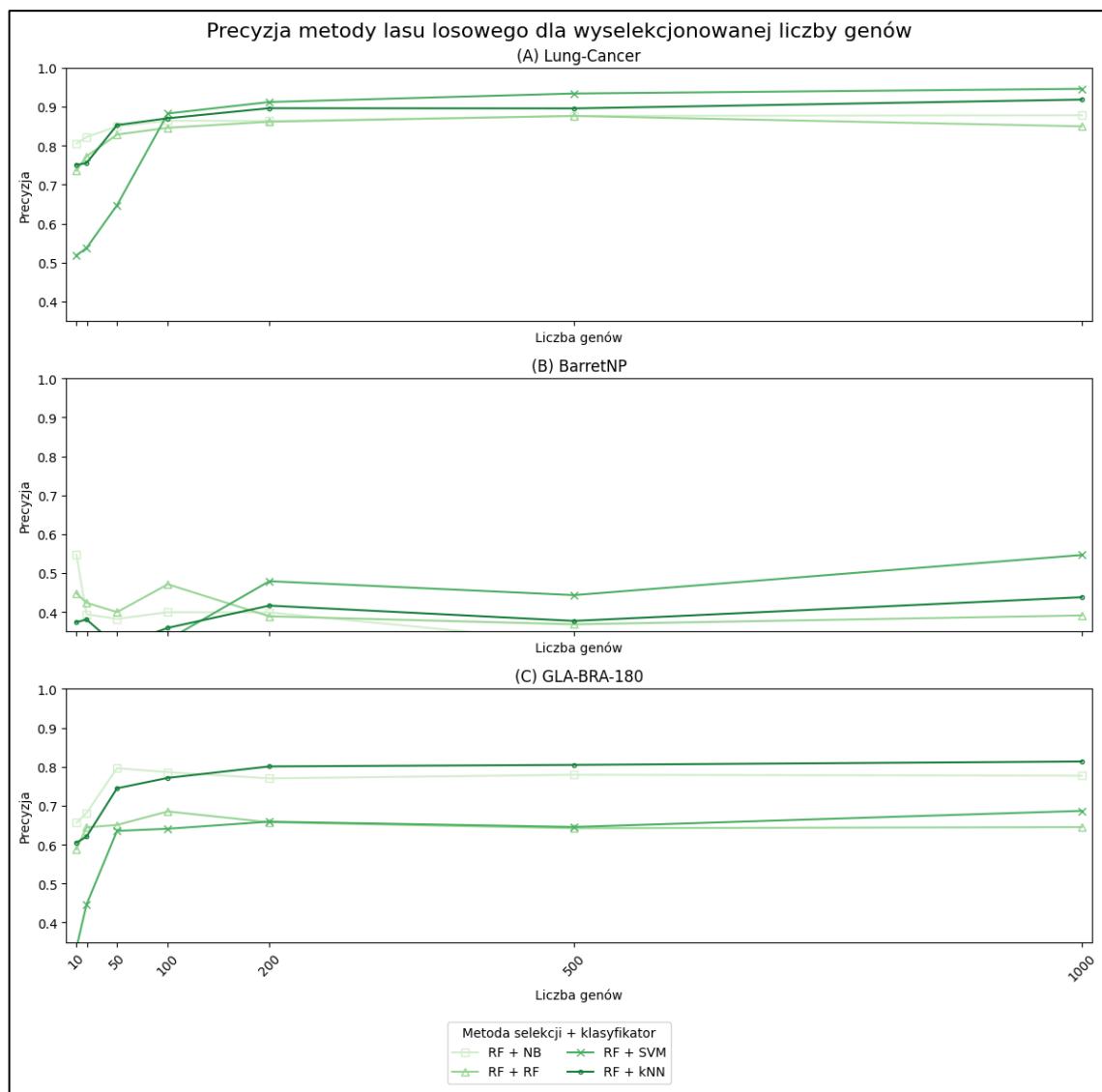
Rysunek 57. Średnia precyza metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora



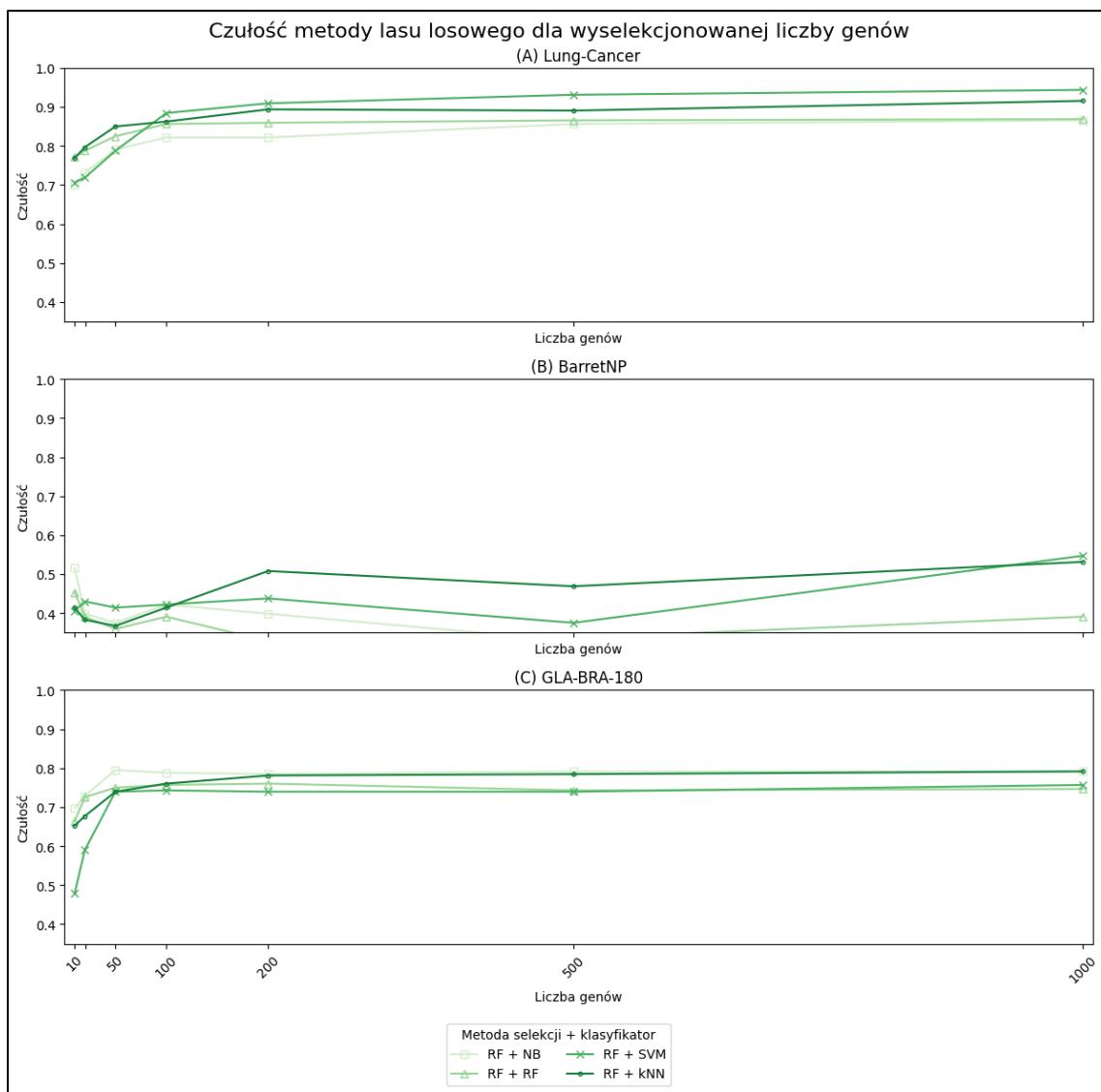
Rysunek 58. Średnia czułość metody lasu losowego dla określonej populacji drzew oraz zastosowanego klasyfikatora



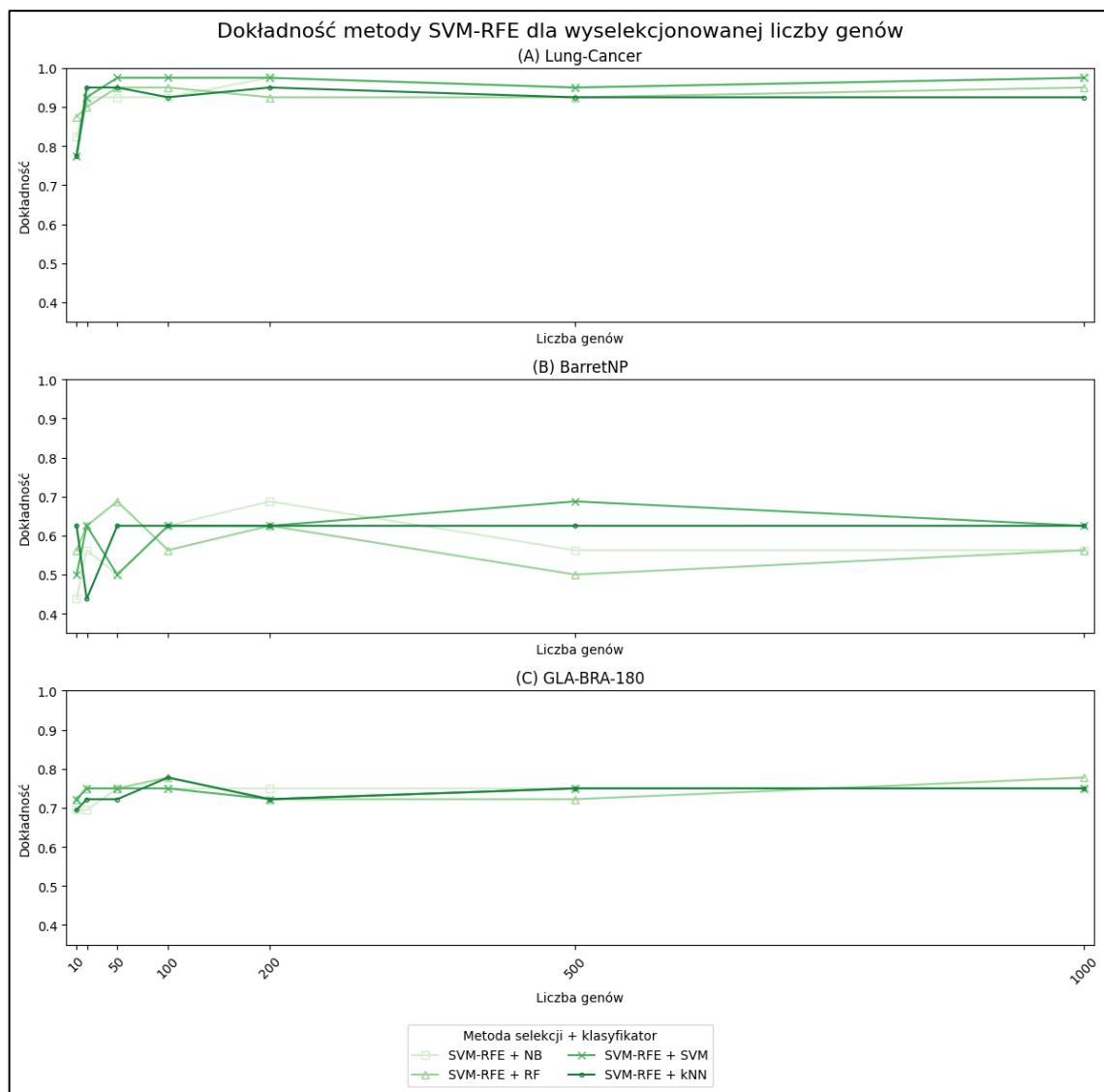
Rysunek 59. Średnia dokładność metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



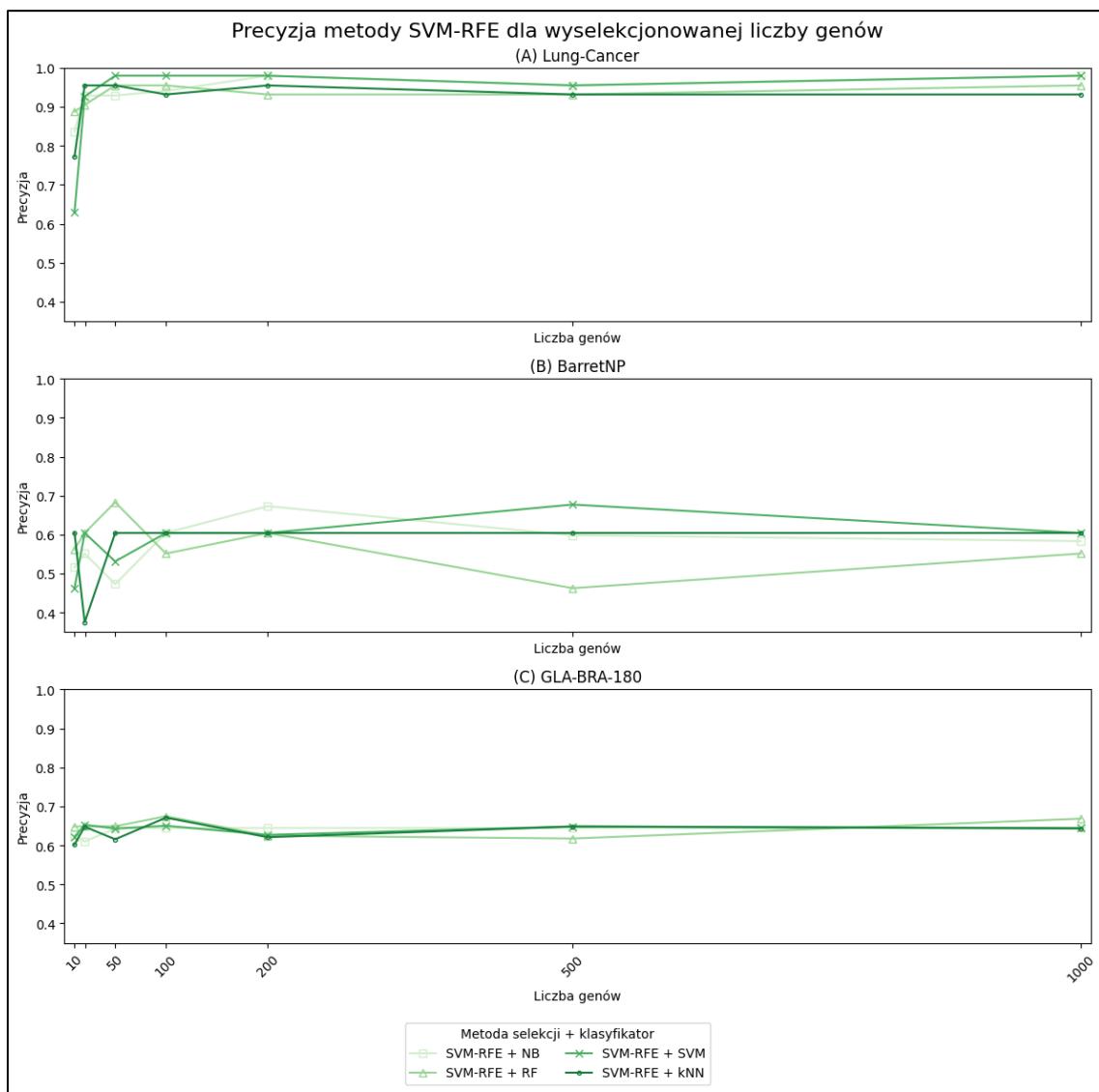
Rysunek 60. Średnia precyza metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



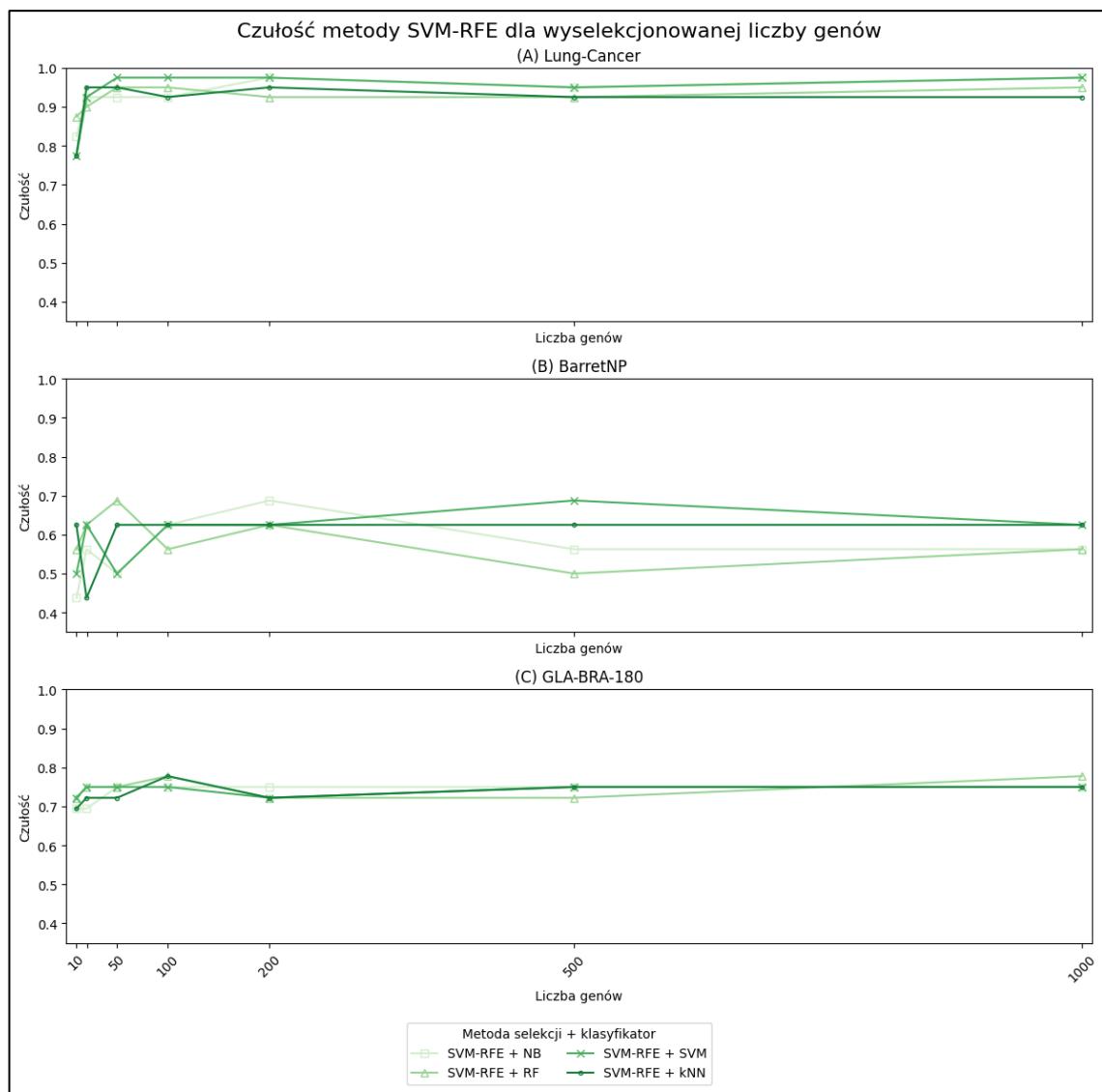
Rysunek 61. Średnia czułość metody lasu losowego dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



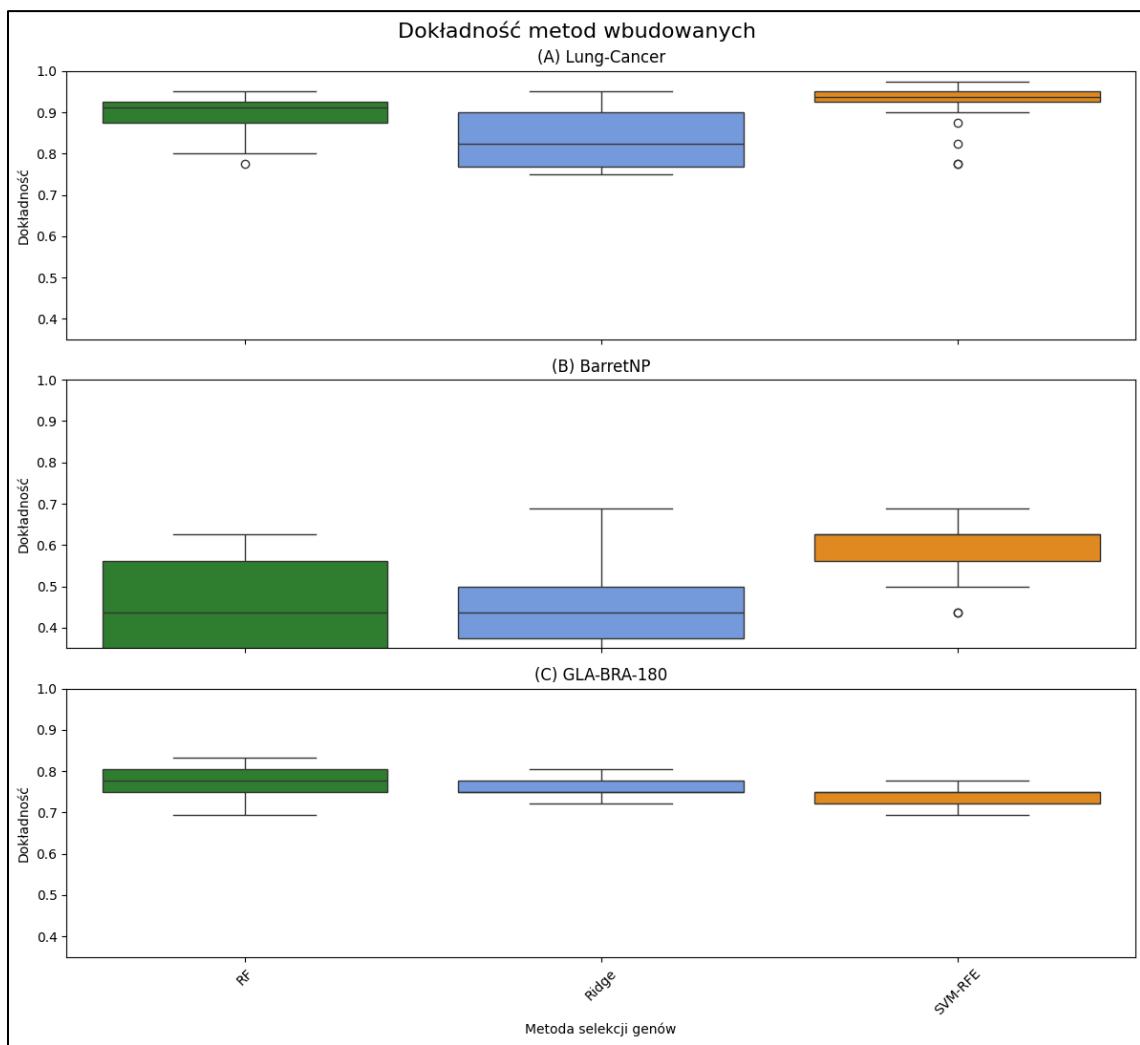
Rysunek 62. Dokładność metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



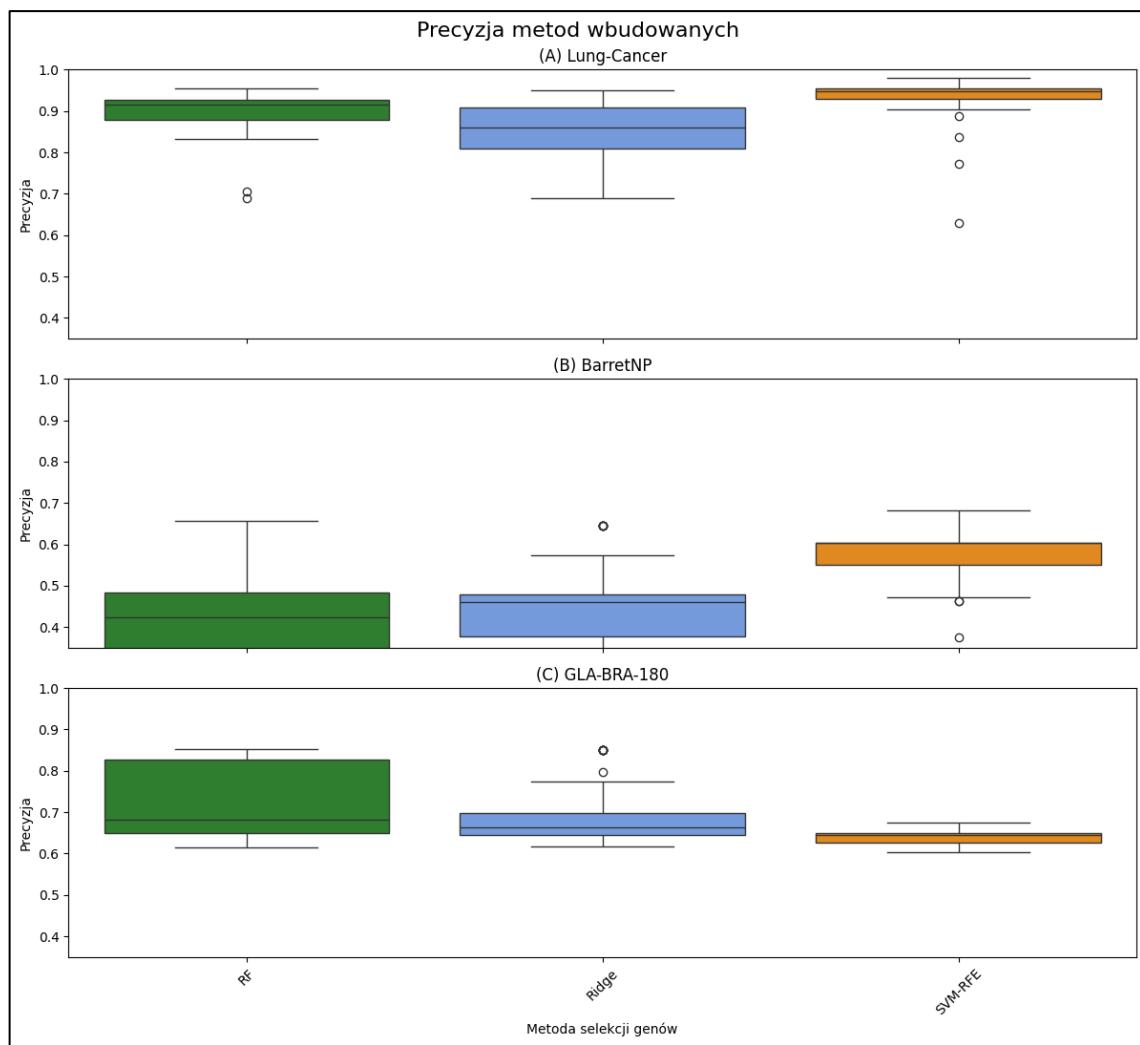
Rysunek 63. Precyza metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



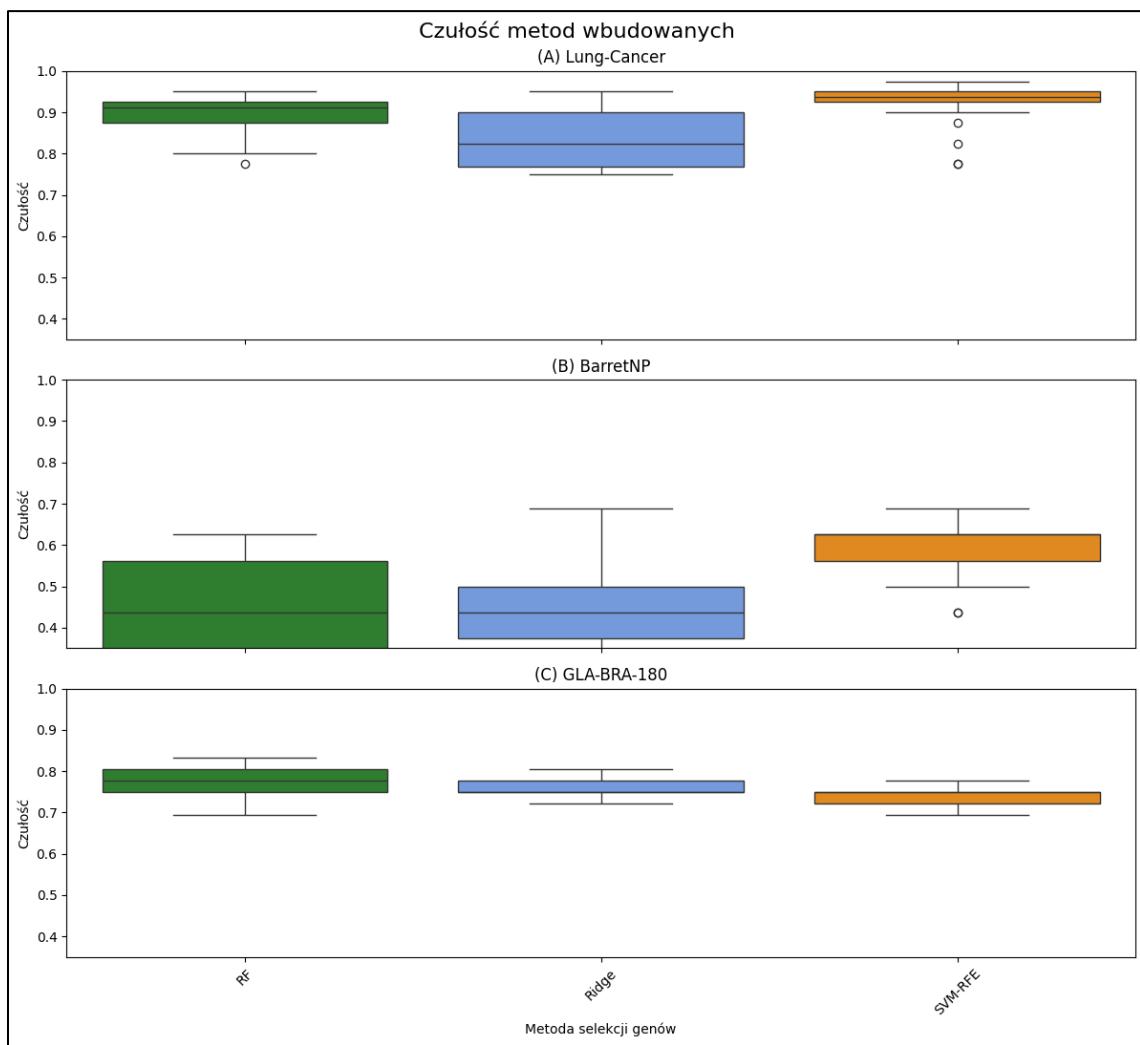
Rysunek 64. Czułość metody SVM-RFE dla określonej liczby wyselekcjonowanych genów oraz zastosowanego klasyfikatora



Rysunek 65. Dokładność dla wszystkich zastosowanych metod wbudowanych



Rysunek 66. Precyza dla wszystkich zastosowanych metod wbudowanych



Rysunek 67. Czułość dla wszystkich zastosowanych metod wbudowanych