Selekcja Cech i Modele Zespołowe w Klasyfikacji Stopnia Złośliwości Glejaka

Dominik Banach, Marek Kacprzak

8 czerwca 2025

Streszczenie

W tym projekcie podjęliśmy się odtworzenia i oceny zaawansowanej metodologii klasyfikacji stopnia złośliwości glejaka, opierając się na pracy naukowców National Institutes of Health (NIH) z roku 2022 [1]. Generalnie, naszym celem było sprawdzenie, czy ich wieloetapowe podejście, łączące selekcję cech i modele zespołowe, jest skuteczne i powtarzalne. Proces opierał się na hierarchicznej selekcji cech, gdzie cztery różne techniki (IV, RFE, Lasy Losowe, LASSO) „głosowały" na najważniejsze predyktory. Następnie, na podstawie tych wybranych cech, zbudowaliśmy i oceniliśmy 16 modeli zespołowych z wykorzystaniem mechanizmu miękkiego głosowania. Całą analizę przeprowadziliśmy na zbiorze 839 pacjentów w ramach 10-krotnej walidacji krzyżowej. Najlepszy z naszych modeli hybrydowych osiągnął średnią dokładność 87.1%, co jest wynikiem niemal identycznym z 87.6% zgłoszonym w oryginalnym badaniu. Potwierdza to powtarzalność zaprezentowanej metodologii.

Słowa kluczowe: uczenie maszynowe, klasyfikacja, glejak, nowotwór mózgu, gen, mutacja, selekcja cech, modele zespołowe, soft voting, Information Value (IV), RFE, LASSO, Lasy Losowe, AdaBoost, SVM, KNN, Regresja Logistyczna, TCGA.

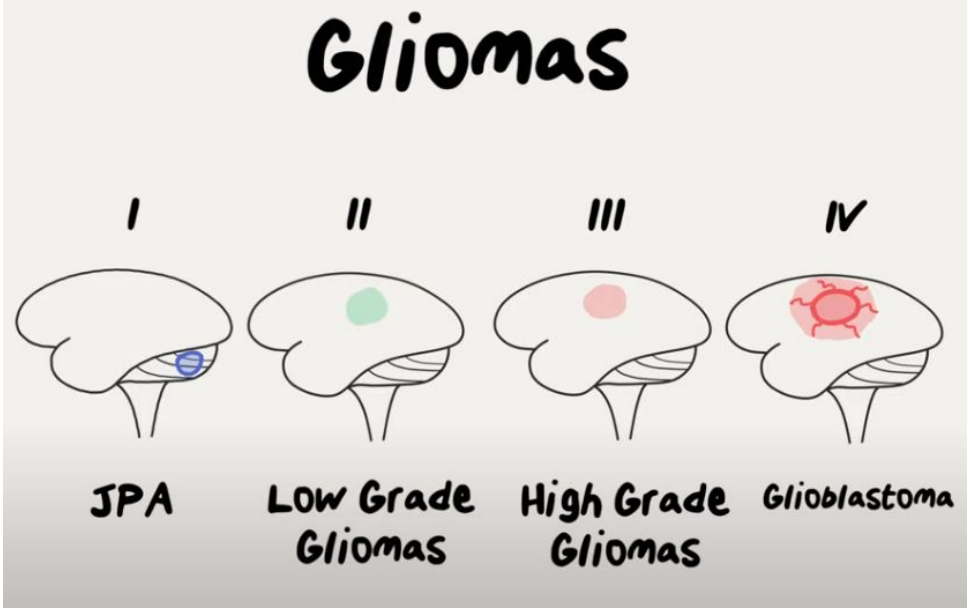
Spis treści

1 Wprowadzenie

Glejaki to jeden z poważniejszych problemów we współczesnej onkologii. Są to najczęstsze i najbardziej zróżnicowane nowotwory mózgu u dorosłych, wywodzące się z komórek glejowych. Kluczowe jest, aby dokładnie ocenić stopień ich złośliwości, ponieważ od tej diagnozy zależy całe dalsze leczenie i rokowania pacjenta.

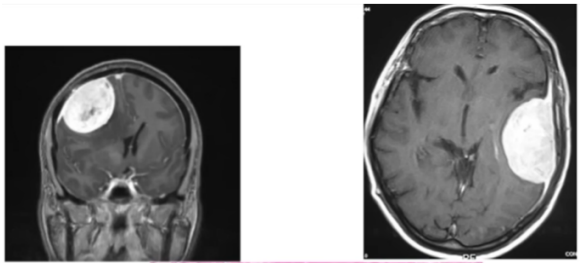
Klasyfikacja glejaków według Światowej Organizacji Zdrowia (WHO) wyróżnia cztery stopnie złośliwości (Rysunek 1). W praktyce klinicznej i w naszym projekcie najważniejsze jest podział na dwie główne grupy:

- **Glejaki o niższym stopniu złośliwości (LGG – Low-Grade Gliomas)**, obejmujące stopnie I i II. Charakteryzują się one wolniejszym wzrostem i stosunkowo lepszym rokowaniem.
- **Glejaki o wysokim stopniu złośliwości (HGG – High-Grade Gliomas)**, obejmujące stopnie III i IV. Najbardziej agresywną formą jest glejak wielopostaciowy (GBM – Glioblastoma Multiforme), klasyfikowany jako stopień IV.

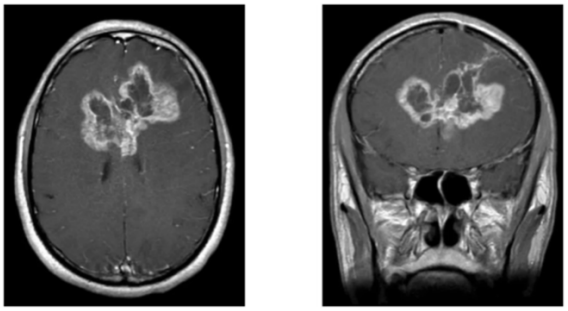


Rysunek 1. Schematyczne przedstawienie stopni złośliwości glejaków według WHO. (źródło: https://youtu.be/HHlvpxS9YgY)

Różnice między LGG a GBM są fundamentalne i dotyczą zarówno wyglądu guza (Rysunek 4), jak i podejścia terapeutycznego, rokowań oraz kosztów leczenia.



Rysunek 2. Obraz rezonansu magnetycznego dla guza o niższym stopniu złośliwości (LGG). Guz jest wyraźnie odgraniczony.



Rysunek 3. Obraz rezonansu magnetycznego dla glejaka wielopostaciowego (GBM). Widoczny jest naciekający, nieregularny charakter guza.

Rysunek 4. Porównanie wizualne glejaka niskiego i wysokiego stopnia w badaniu MRI, czyli rezonansie magnetycznym. (źródło: https://youtu.be/N2EsvFkUTH8)

W przypadku LGG, podstawą leczenia jest jak najszersza resekcja chirurgiczna, po której często stosuje się obserwację lub, w przypadkach wysokiego ryzyka, radioterapię i chemioterapię. Rokowania są stosunkowo dobre, a przeżycie może sięgać nawet kilkunastu lat, zwłaszcza przy obecności korzystnych mutacji genetycznych.

Z kolei GBM to zupełnie inne wyzwanie. Ze względu na jego naciekający charakter, całkowite usunięcie chirurgiczne jest praktycznie niemożliwe. Standardem leczenia jest agresywna radiochemioterapia po operacji. Mimo to, rokowania są bardzo złe – mediana przeżycia pacjentów wynosi zaledwie około 15 miesięcy, a 5-letnie przeżycie nie przekracza 10%. Leczenie jest przy tym niezwykle kosztowne.

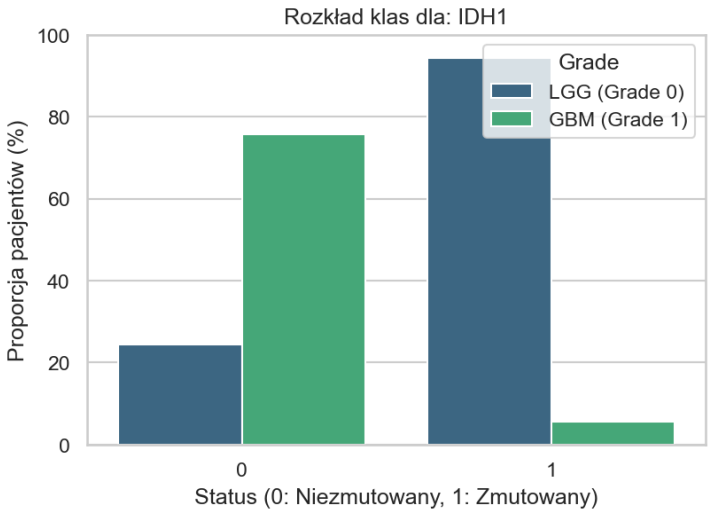
1.1 Rola markerów molekularnych w powstawaniu glejaków

Glejaki powstają w wyniku nagromadzenia się mutacji genetycznych w komórkach glejowych, co prowadzi do utraty kontroli nad ich podziałami i wzrostem. Zmiany te nie są przypadkowe i często dotyczą genów, które pełnią kluczowe role w specyficznych szlakach biologicznych. Zrozumienie funkcji tych genów pozwala na podział glejaków na molekularne podtypy, co ma fundamentalne znaczenie dla rokowania i wyboru terapii. Poniżej przedstawiono najważniejsze z analizowanych genów, pogrupowane według ich podstawowych funkcji w komórce.

1.1.1 Kluczowe geny definiujące podtypy glejaków

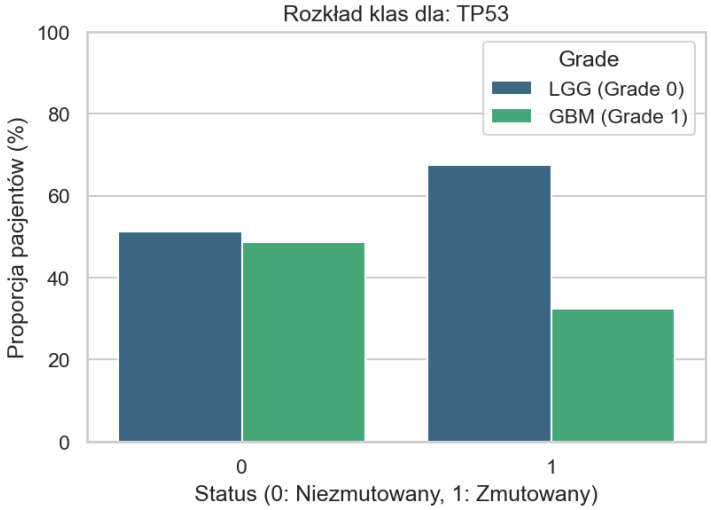
Pewne mutacje są tak fundamentalne, że stały się podstawą nowoczesnej klasyfikacji glejaków.

- `IDH1` **(Dehydrogenaza izocytrynianowa 1):** To jeden z najważniejszych markerów. Jego mutacja jest wydarzeniem inicjującym dla większości glejaków o niższym stopniu złośliwości (LGG). Co ciekawe, zmutowany enzym `IDH1` nie przestaje działać, lecz nabywa nową, toksyczną funkcję – zaczyna produkować tzw. onkometabolit, który całkowicie zmienia metabolizm komórki. Paradoksalnie, guzy z tą mutacją, mimo że są nowotworowe, charakteryzują się znacznie lepszym rokowaniem.



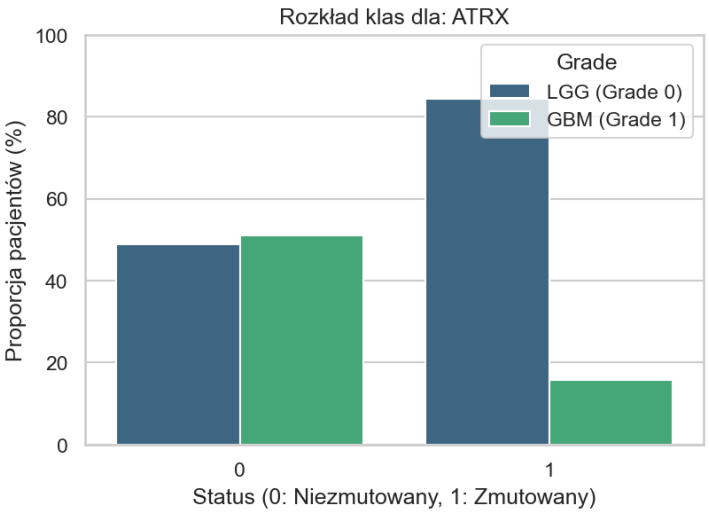
Rysunek 5. Rozkład klas dla genu `IDH1`.

- `TP53` **(Tumor protein p53):** Nazywany "strażnikiem genomu". Jego prawidłowa funkcja polega na wykrywaniu uszkodzeń DNA i zatrzymywaniu podziału komórki w celu ich naprawy lub, w razie potrzeby, uruchomieniu jej programowanej śmierci (apoptozy). Mutacja w genie `TP53` to jeden z najczęstszych kroków w rozwoju wielu nowotworów, w tym glejaków. Powoduje ona oporność na leczenie i przyczynia się do agresywności guza.



Rysunek 6. Rozkład klas dla genu `TP53`.

- `ATRX` **(ATRX chromatin remodeler):** Gen ten jest kluczowy dla utrzymania stabilności chromosomów. Jego mutacje prowadzą do niestabilności genomu. Często współwystępują z mutacjami `IDH1` i `TP53`, definiując konkretny podtyp gwiaździaków rozlanych (jedna z form LGG) o lepszym rokowaniu niż guzy bez mutacji `IDH1`.

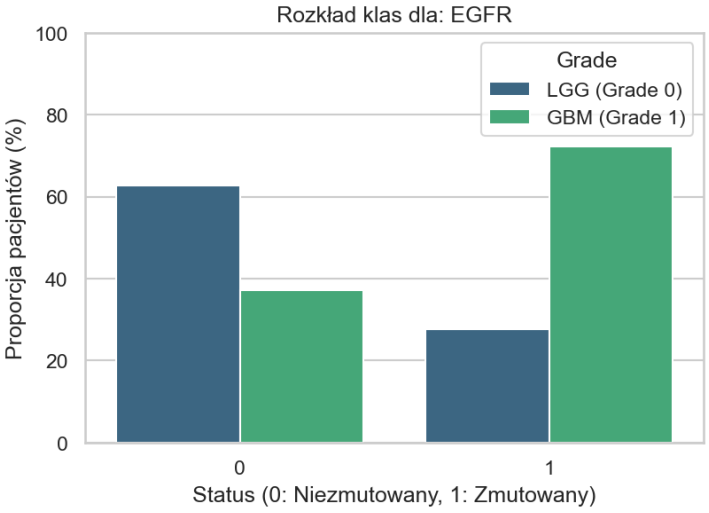


Rysunek 7. Rozkład klas dla genu `ATRX`.

1.1.2 Geny związane z sygnalizacją i wzrostem komórki

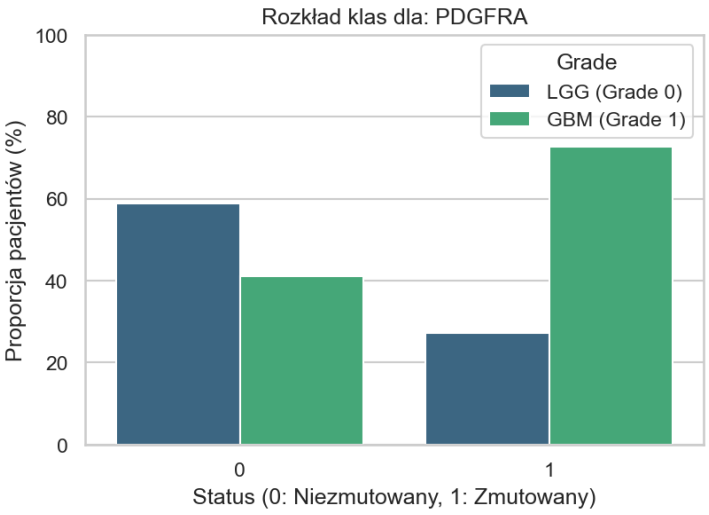
Ta grupa genów działa jak system komunikacji w komórce, przekazując sygnały do wzrostu i podziału. Ich mutacje często prowadzą do nieustannego wysyłania sygnału do wzrostu i podziału.

- `EGFR` **(Epidermal Growth Factor Receptor):** W glejakach, zwłaszcza GBM, gen ten jest często zwielokrotniony (ulega amplifikacji), co prowadzi do nadprodukcji receptora na powierzchni komórki i ciągłej stymulacji do podziałów. Jest to cecha charakterystyczna dla tzw. klasycznego podtypu GBM.



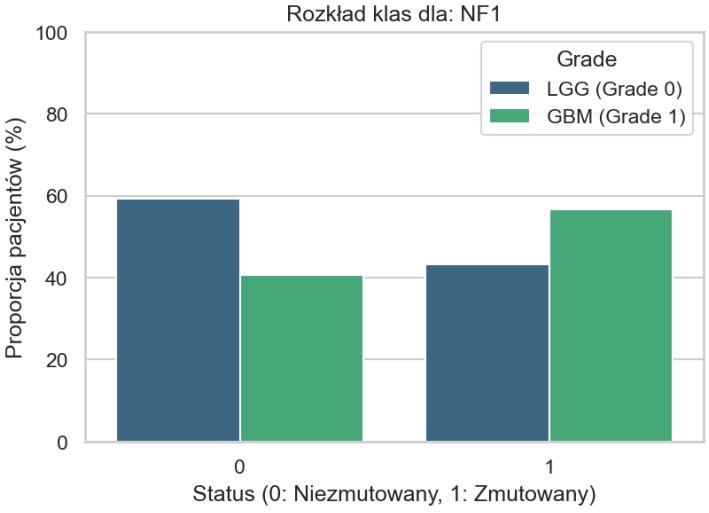
Rysunek 8. Rozkład klas dla genu `EGFR`.

- `PDGFRA` **(Platelet-derived growth factor receptor alpha):** Działa podobnie do EGFR, będąc receptorem dla innego czynnika wzrostu.



Rysunek 9. Rozkład klas dla genu `PDGFRA`.

- `NF1` **(Neurofibromin 1):** W przeciwieństwie do powyższych, jest to gen supresorowy, który hamuje szlaki sygnałowe. Jego mutacja lub utrata powoduje utratę kontroli nad hamowaniem podziałów.

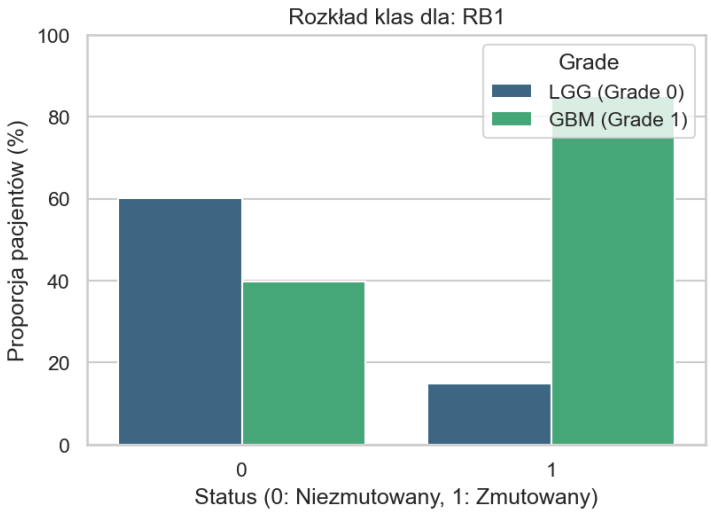


Rysunek 10. Rozkład klas dla genu `NF1`.

1.1.3 Geny regulujące cykl komórkowy i ekspresję genów

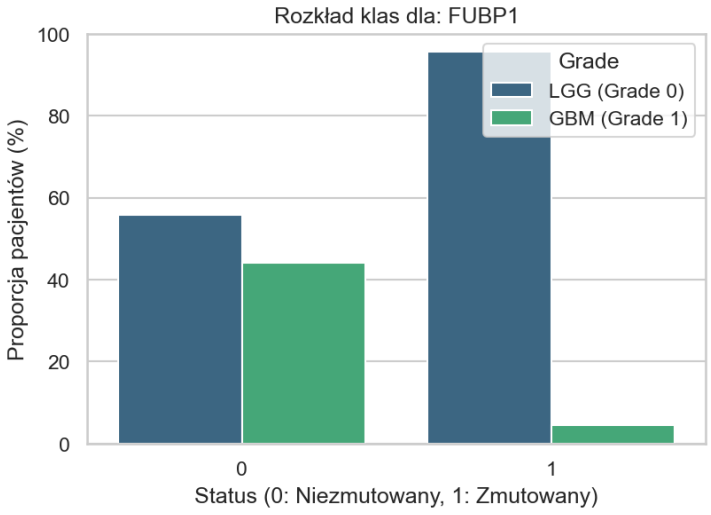
Ta grupa odpowiada za kontrolę poszczególnych faz cyklu komórkowego oraz za to, które geny w komórce są aktywne.

- `RB1` **(RB transcriptional corepressor 1):** Jeden z najważniejszych genów supresorowych, który jest regulatorem cyklu komórkowego. Decyduje, czy komórka może rozpocząć proces podziału. Jego mutacja jest jednym z klasycznych mechanizmów utraty kontroli nad gwałtownym rozrostem.



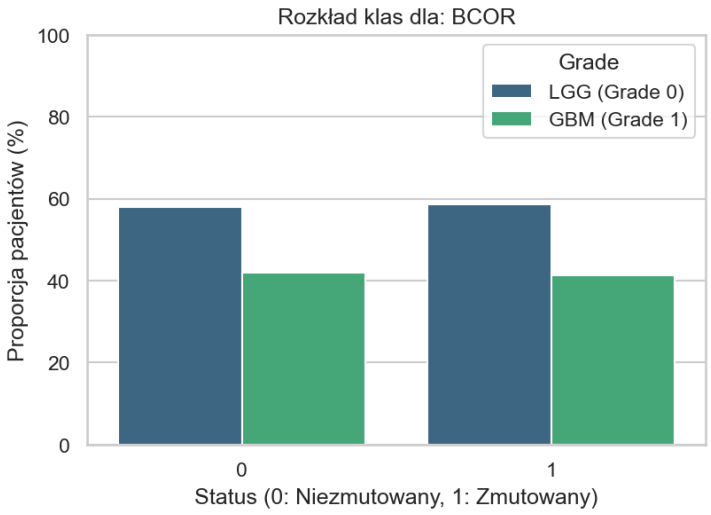
Rysunek 11. Rozkład klas dla genu `RB1`.

- `FUBP1` **(Far upstream element binding protein 1):** Gen, którego produkt białkowy bierze udział w regulacji ekspresji innych genów. Jego mutacja może zaburzyć całą dynamikę genetyczną komórki.



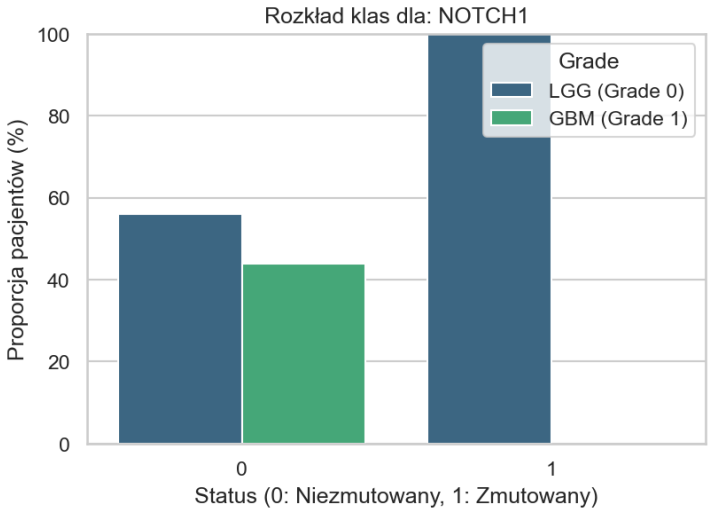
Rysunek 12. Rozkład klas dla genu `FUBP1`.

- `BCOR` **(BCL6 Corepressor):** Podobnie jak FUBP1, jest to gen zaangażowany w regulację aktywności innych genów.



Rysunek 13. Rozkład klas dla genu `BCOR`.

- `NOTCH1` **(Notch receptor 1):** Kolejny receptor, którego szlak sygnałowy jest kluczowy dla komunikacji międzykomórkowej i rozwoju komórek.



Rysunek 14. Rozkład klas dla genu `NOTCH1`.

1.1.4 Pozostałe istotne geny w puli

W naszym zbiorze danych znajduje się również wiele innych genów, których mutacje przyczyniają się do rozwoju glejaka, choć ich rola może być bardziej zróżnicowana. Należą do nich m.in. `CIC`, `PTEN`, `PIK3CA`, `PIK3R1` (zaangażowane w szlaki sygnałowe), `SMARCA4` (remodelowanie chromatyny) czy `IDH2` (działający podobnie do `IDH1`).

1.2 Cel badania w kontekście biologii glejaków

Jak widać, obraz molekularny glejaków jest niezwykle złożony. Pacjent może mieć mutacje w wielu z tych genów jednocześnie, a ich kombinacja decyduje o charakterze guza. Ręczna analiza tak wielu cech jest trudna i nieefektywna.

Właśnie dlatego celem naszego projektu jest weryfikacja metodologii, która potrafi automatycznie "nauczyć się", które z tych wszystkich genów i cech klinicznych są najważniejsze. Chcemy

sprawdzić, czy podejście oparte na hierarchicznej selekcji cech i modelach zespołowych, zaproponowane przez Tasci et al. (2022) [1], jest w stanie skutecznie przetworzyć tę złożoność i stworzyć model, który z wysoką dokładnością odróżni glejaki o niskim i wysokim stopniu złośliwości, bazując na ich unikalnej charakterystyce molekularnej.

2 Wstępna analiza danych

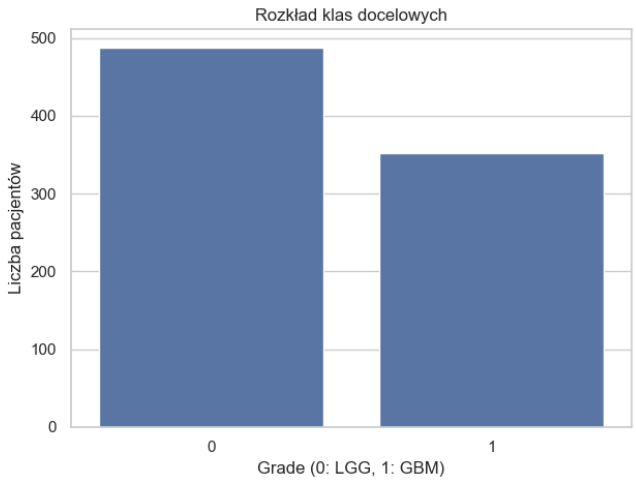
2.1 Charakterystyka zbioru i zmiennych

Analizowany zbiór danych pochodzi z repozytorium UCI Machine Learning i jest oparty na danych z projektu The Cancer Genome Atlas (TCGA). Składa się on z 839 obserwacji (pacjentów) i 24 zmiennych, w tym jednej zmiennej docelowej i 23 zmiennych objaśniających. Zmienne objaśniające dzielą się na 3 cechy kliniczne oraz 20 cech molekularnych opisujących binarnie status mutacji poszczególnych genów.

Zmienną docelową jest `Grade`, która określa stopień złośliwości guza. Tabela 1 przedstawia rozkład klas w zbiorze, a Rysunek 15 wizualizuje go graficznie.

Tabela 1. Rozkład klas w zmiennej docelowej 'Grade'.

Klasa (Grade)	Opis	Liczebność	Proporcja (%)
0	LGG (Glejak o niższym stopniu złośliwości)	487	58.05
1	GBM (Glejak wielopostaciowy)	352	41.95



Rysunek 15. Rozkład liczebności klas w zmiennej docelowej.

2.2 Statystyki opisowe

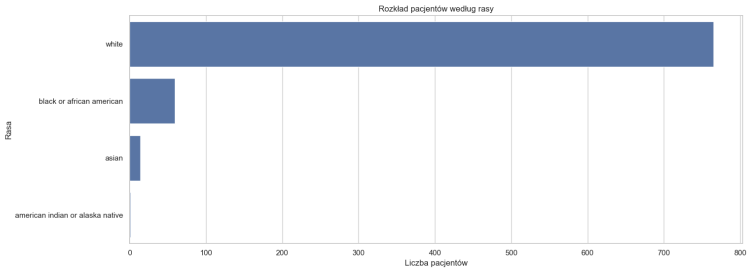
Wśród zmiennych objaśniających jedyną cechą ciągłą jest wiek pacjenta w momencie diagnozy. Pozostałe zmienne mają charakter kategorialny. Tabela 2 prezentuje kluczowe statystyki opisowe dla wieku. Rozkład tej zmiennej jest niemal idealnie symetryczny, na co wskazuje wartość skośności bliska zeru (0.07).

Tabela 2. Statystyki opisowe dla zmiennej numerycznej 'Wiek w momencie diagnozy'.

Zmienna	Średnia	Mediana	Minimum	Maksimum	Odch. stand.	Skośność
Wiek w momencie diagnozy	50.94	51.55	14.42	89.29	15.70	0.07

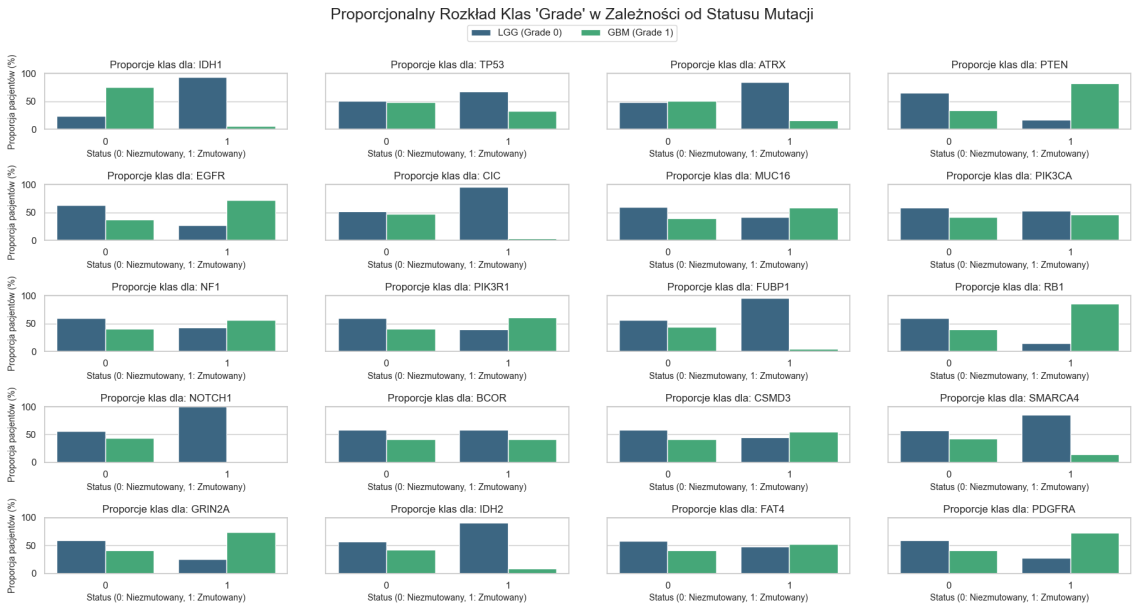
2.3 Wizualizacja danych

Wizualizacja zawiera rozkłady cech klinicznych (Rysunek 16 i Rysunek 17) oraz związek między statusem mutacji genów a klasą guza (Rysunek 18).



Rysunek 16. Rozkład pacjentów według rasy.

Rysunek 17. Analiza zmiennej 'Wiek w momencie diagnozy': histogram rozkładu (po lewej) oraz boxplot przedstawiający rozkład wieku względem klasy guza (po prawej).



Rysunek 18. Proporcjonalny rozkład klas 'Grade' (LGG/GBM) w zależności od statusu mutacji dla 20 analizowanych genów. Każdy słupek sumuje się do 100%.

2.4 Wstępne wnioski z analizy

Wstępna analiza wizualna, a w szczególności analiza proporcjonalnego rozkładu klas (Rysunek 18), pozwala na zidentyfikowanie cech o potencjalnie dużej i małej mocy predykcyjnej, jeszcze przed zastosowaniem formalnych metod selekcji.

Jednym z najsilniejszych predyktorów w analizowanym zbiorze wydaje się być status mutacji genu `IDH1`. Obserwacje z wykresu są jednoznaczne:

- W grupie pacjentów posiadających **mutację genu** `IDH1` (słupek dla wartości 1), niemal 100% przypadków to glejaki o niższym stopniu złośliwości (LGG, Grade 0).
- W grupie pacjentów **bez tej mutacji** (słupek dla wartości 0), proporcje drastycznie się odwracają, a zdecydowaną większość stanowią pacjenci z glejakiem wielopostaciowym (GBM, Grade 1).

Tak wyraźna separacja klas sugeruje, że status mutacji genu `IDH1` jest kluczową cechą do rozróżniania analizowanych typów guzów i prawdopodobnie zostanie wysoko oceniony przez wszystkie metody selekcji cech.

Przykładem cechy o potencjalnie niskiej mocy predykcyjnej jest gen `PIK3CA`. Z odpowiedniego subplotu na Rysunku 18 można odczytać, że:

- Proporcje klas LGG i GBM są bardzo zbliżone do siebie, niezależnie od tego, czy gen jest zmutowany, czy nie.
- W obu grupach – zarówno z mutacją, jak i bez niej – rozkład procentowy klas oscyluje wokół ogólnego rozkładu w całym zbiorze (ok. 58% LGG do 42% GBM).

Sugeruje to, że informacja o posiadaniu mutacji w genie `PIK3CA` wnosi niewiele nowej wiedzy do procesu klasyfikacji i prawdopodobnie nie będzie kluczowym predyktorem w finalnym modelu.

3 Metodologia badawcza

Zastosowaliśmy wieloetapowe podejście do analizy danych i budowy modelu klasyfikacyjnego, opierając się na metodologii przedstawionej przez autorów pracy[1].

3.1 Walidacja i przygotowanie danych

Wszystkie etapy selekcji cech oraz trenowania i oceny modeli przeprowadzono z użyciem 10-krotnej walidacji krzyżowej. Takie podejście zapewnia bardziej rzetelną ocenę wydajności modelu, minimalizując ryzyko nadmiernego dopasowania do konkretnego podziału danych.

Przed przystąpieniem do selekcji cech i modelowania, w każdej fałdzie walidacji krzyżowej dane treningowe i walidacyjne poddawano następującym krokom preprocessingu:

- Kodowanie zmiennej kategorialnej `Race` na wartości numeryczne.
- Standaryzacja (Z-score) zmiennej ciągłej `Age_at_diagnosis`, polegająca na odjęciu średniej i podzieleniu przez odchylenie standardowe. Parametry standaryzacji (średnia i odchylenie standardowe) były obliczane wyłącznie na zbiorze treningowym danej fałdy, a następnie stosowane do transformacji zarówno zbioru treningowego, jak i walidacyjnego.

3.2 Metody selekcji cech

W celu wyboru optymalnego podzbioru cech, w każdej fałdzie walidacji krzyżowej zastosowano cztery niezależne metody. Ostateczny zestaw cech determinowany był przez mechanizm głosowania.

3.2.1 Weight of Evidence (WOE) / Information Value (IV)

Information Value (IV) jest miarą mocy predykcyjnej pojedynczej cechy. Wyższa wartość IV wskazuje na silniejszą zdolność cechy do rozróżniania między klasami. WOE dla każdej kategorii i danej cechy jest obliczane jako logarytm stosunku dystrybucji klasy pozytywnej (GBM) do negatywnej (LGG):

$$\text{WOE}_i = \ln\left(\frac{\%\text{ pacjentów z GBM w kategorii } i}{\%\text{ pacjentów z LGG w kategorii } i}\right)$$

Następnie, IV dla całej cechy jest sumą ważoną WOE dla wszystkich jej kategorii. W każdej fałdzie obliczono IV dla wszystkich cech, przy czym dla cechy ciągłej `Age_at_diagnosis` zastosowano dyskretyzację na 5 kwantyli. Cecha otrzymywała "głos", jeśli jej IV przekroczyło próg **0.02**.

3.2.2 Rekurencyjna Eliminacja Cech (RFE)

RFE[1] to iteracyjna metoda selekcji, która rozpoczyna od pełnego zestawu cech i w każdym kroku trenuje model bazowy, a następnie usuwa najsłabszą cechę, aż do osiągnięcia pożądanej ich liczby. Jako estymatora w RFE użyto modelu Regresji Logistycznej z parametrami `solver='liblinear'` i `max_iter=200`. W każdej fałdzie RFE wybierało **12 cech** o najwyższej istotności.

[1] Więcej informacji na temat metody: [9]

3.2.3 Ważność cech z Lasów Losowych

Lasy Losowe[2], będące zespołem drzew decyzyjnych, naturalnie dostarczają miary ważności cech, najczęściej opartej na średnim spadku nieczystości Giniego (Mean Decrease in Impurity). W każdej fałdzie trenowano model Lasów Losowych (`n_estimators=100`), a następnie wybierano cechy, których obliczona ważność przekraczała próg ustalony jako **połowa średniej ważności** wszystkich cech.

3.2.4 Selekcja z użyciem LASSO (L1)

LASSO[3] to metoda regresji z regularyzacją L1, która dodaje do funkcji kosztu karę za wielkość współczynników, co ma zdolność do zerowania współczynników mniej istotnych cech. W każdej fałdzie trenowano model Regresji Logistycznej z karą L1. Kluczowym hiperparametrem jest odwrotność siły regularyzacji `C`, który ustalono na **0.1**, co oznacza stosunkowo silną regularyzację. Wybierano cechy, których współczynniki w modelu LASSO były różne od zera.

3.2.5 Schemat głosowania

Po zastosowaniu wszystkich czterech metod selekcji w danej fałdzie, tworzono finalny zestaw cech. Zgodnie z podejściem z referencyjnego artykułu [1], do tego zestawu wybierano każdą cechę, która otrzymała co najmniej jeden "głos"(została wskazana jako istotna) od którejkolwiek z czterech zastosowanych metod.

3.3 Modele klasyfikacyjne

Po etapie selekcji cech, w każdej fałdzie walidacji krzyżowej, na zredukowanym zbiorze cech trenowano pięć indywidualnych modeli bazowych. Celem było nie tylko uzyskanie ich indywidualnej wydajności, ale przede wszystkim zebranie predykcji prawdopodobieństw, które posłużą do budowy modeli zespołowych. Hiperparametry dla tych modeli dobrano tak, aby były zgodne z użytymi w artykule Tasci et al. (2022) [1].

3.3.1 Regresja Logistyczna (LR) [4]

To prosty, ale skuteczny model, który modeluje prawdopodobieństwo przynależności obserwacji do danej klasy.

- `penalty='l2'`: Zastosowano regularyzację L2 (Ridge), która zapobiega przeuczeniu poprzez dodanie kary za wysokie wartości współczynników modelu.
- `C=1.0`: Parametr ten jest odwrotnością siły regularyzacji. Wartość 1.0 oznacza standardowy, umiarkowany poziom regularyzacji.
- `solver='liblinear'`: Algorytm optymalizacyjny, który dobrze sprawdza się dla mniejszych zbiorów danych i jest kompatybilny z regularyzacją L1 i L2.

3.3.2 Maszyna Wektorów Nośnych (SVM) [5]

Model, który szuka optymalnej hiperpłaszczyzny separującej klasy w wielowymiarowej przestrzeni cech.

- `C=1.0`: Parametr regularyzacji. Kontroluje kompromis między maksymalizacją marginesu separującego a minimalizacją błędów klasyfikacji.

[2]Opis algorytmu: [7]
[3]Opis metody statystycznej: [10]

- `kernel='rbf'`: Zastosowano jądro RBF (Radial Basis Function), które pozwala na znalezienie nieliniowej granicy decyzyjnej, mapując dane na przestrzeń o wyższym wymiarze.
- `gamma='scale'`: Parametr jądra RBF. Określa, jak daleko sięga wpływ pojedynczej obserwacji treningowej. Wartość 'scale' jest dobrym, domyślnym wyborem, dostosowującym się do danych.
- `probability=True`: Kluczowy parametr, który sprawia, że model SVM jest w stanie generować predykcje prawdopodobieństw, niezbędne do mechanizmu soft voting.

3.3.3 k-Najbliższych Sąsiadów (KNN) [6]

Prosty algorytm instancyjny, który klasyfikuje nową obserwację na podstawie klasy większości jej "k"najbliższych sąsiadów w przestrzeni cech.

- `n_neighbors=5`: Liczba sąsiadów branych pod uwagę przy podejmowaniu decyzji o klasyfikacji.
- `metric='minkowski'`: Metryka używana do obliczania odległości między obserwacjami. Przy domyślnym parametrze p=2, jest to standardowa odległość euklidesowa.

3.3.4 Lasy Losowe (RF) [7]

Model zespołowy składający się z wielu drzew decyzyjnych. Ostateczna decyzja podejmowana jest na podstawie głosowania większościowego drzew.

- `n_estimators=100`: Liczba drzew decyzyjnych w "lesie".
- `criterion='gini'`: Funkcja używana do mierzenia jakości podziału w drzewie (tzw. nieczystość Giniego).
- `min_samples_split=2`: Minimalna liczba obserwacji wymagana do dokonania podziału w węźle drzewa.
- `min_samples_leaf=1`: Minimalna liczba obserwacji, jaka musi znaleźć się w liściu (węźle końcowym) drzewa.

3.3.5 AdaBoost [8]

Algorytm boostingowy, który sekwencyjnie buduje słabe klasyfikatory (np. proste drzewa), przy czym każdy kolejny model skupia się na błędach popełnionych przez poprzedników.

- `n_estimators=50`: Liczba słabych klasyfikatorów do zbudowania w sekwencji.
- `learning_rate=1.0`: Parametr, który określa, jak bardzo każdy kolejny model wpływa na ostateczny wynik zespołu. Wartość 1.0 oznacza, że modele wnoszą swój pełny wkład.

Należy zaznaczyć, że w artykule podano również parametr `algorithm='SAMME.R'`, jednak w nowszych wersjach biblioteki `scikit-learn` stał się on przestarzały i nie ma już wpływu na działanie algorytmu, dlatego został pominięty w naszej finalnej implementacji.

3.3.6 Model hybrydowy (Soft Voting)

Mechanizm miękkiego głosowania (soft voting) jest techniką budowy modeli zespołowych, która agreguje predykcje z kilku różnych klasyfikatorów. W odróżnieniu od głosowania twardego (gdzie liczy się tylko głosy na konkretną klasę), soft voting uśrednia przewidywane przez modele bazowe **prawdopodobieństwa** przynależności do każdej z klas. Ostateczna predykcja zespołu to ta klasa, która uzyskała najwyższe średnie prawdopodobieństwo. W ramach projektu utworzono i oceniono 16 unikalnych modeli hybrydowych.

3.4 Miary oceny modeli

Do oceny wydajności zarówno modeli bazowych, jak i zespołowych, wykorzystano sześć standardowych metryk klasyfikacji binarnej:

- **Dokładność (ACC):** Proporcja poprawnie sklasyfikowanych obserwacji do wszystkich obserwacji. Jest to najbardziej ogólna miara skuteczności.
- **AUC (Area Under the ROC Curve):** Pole pod krzywą ROC. Mierzy zdolność modelu do rozróżniania między klasą pozytywną a negatywną. Wartość 1.0 oznacza idealny klasyfikator, a 0.5 klasyfikator losowy.
- **F1-Score:** Średnia harmoniczna precyzji i czułości. Jest to dobra miara dla niezbalansowanych zbiorów danych, ponieważ uwzględnia obie metryki.
- **Precyzja (PRE):** Stosunek prawdziwie pozytywnych predykcji do wszystkich predykcji pozytywnych. Odpowiada na pytanie: "jaki odsetek pacjentów, których model oznaczył jako GBM, faktycznie ma GBM?".
- **Czułość (REC):** Stosunek prawdziwie pozytywnych predykcji do wszystkich rzeczywistych przypadków pozytywnych. Odpowiada na pytanie: "jaki odsetek pacjentów, którzy faktycznie mają GBM, został poprawnie wykryty przez model?".
- **Swoistość (SPEC):** Stosunek prawdziwie negatywnych predykcji do wszystkich rzeczywistych przypadków negatywnych. Odpowiada na pytanie: "jaki odsetek pacjentów, którzy faktycznie mają LGG, został poprawnie wykryty przez model?".

4 Wyniki

W tej sekcji przedstawiono rezultaty przeprowadzonej hierarchicznej selekcji cech oraz wyniki indywidualnych modeli bazowych i modeli zespołowych.

4.1 Wyniki selekcji cech

Proces selekcji cech został przeprowadzony w każdej z 10 fałd walidacji krzyżowej przy użyciu czterech niezależnych metod. Tabela 3 prezentuje średnie wartości Information Value dla wszystkich analizowanych cech.

Tabela 3. Średnie wartości Information Value (IV) dla cech (po 10 fałdach).

Cecha	Średnie IV
IDH1	2.8009
Age-at-diagnosis	1.3727
CIC	0.6395
PTEN	0.6091
ATRX	0.4795
NOTCH1	0.3095
EGFR	0.2449
FUBP1	0.2151
RB1	0.1761
TP53	0.1095
IDH2	0.0689
Race	0.0654
GRIN2A	0.0591
MUC16	0.0579
SMARCA4	0.0493
PDGFRA	0.0437
PIK3R1	0.0420
NF1	0.0324
Gender	0.0153
CSMD3	0.0105
FAT4	0.0057
PIK3CA	0.0040
BCOR	0.0007

Tabela 4 podsumowuje średnią liczbę cech wybieranych przez każdą z metod po finalnym głosowaniu.

Tabela 4. Średnia liczba cech wybranych przez poszczególne metody selekcji po głosowaniu (po 10 fałdach).

Metoda Selekcji	Średnia liczba wybranych cech
Metoda IV (próg > 0.02)	18.00
Metoda RFE (top 12)	12.00
Metoda Random Forest (próg > 0.5*średnia)	8.90
Metoda LASSO (C=0.1)	4.40
Po Głosowaniu (min. 1 głos)	18.60

Najważniejszym wynikiem etapu selekcji jest informacja o tym, które cechy były konsekwentnie wybierane przez mechanizm głosowania we wszystkich fałdach. Częstość wyboru poszczególnych cech przedstawiono w Tabeli 5.

Tabela 5. Częstość wyboru poszczególnych cech w mechanizmie głosowania (po 10 fałdach).

Cecha	Liczba fałd (z 10), w których wybrano
IDH2	10
TP53	10
CIC	10
IDH1	10
SMARCA4	10
EGFR	10
NOTCH1	10
GRIN2A	10
Race	10
NF1	10
PTEN	10
RB1	10
MUC16	10
FUBP1	10
PDGFRA	10
Age at diagnosis	10
PIK3R1	10
ATRX	10
Gender	5
CSMD3	1

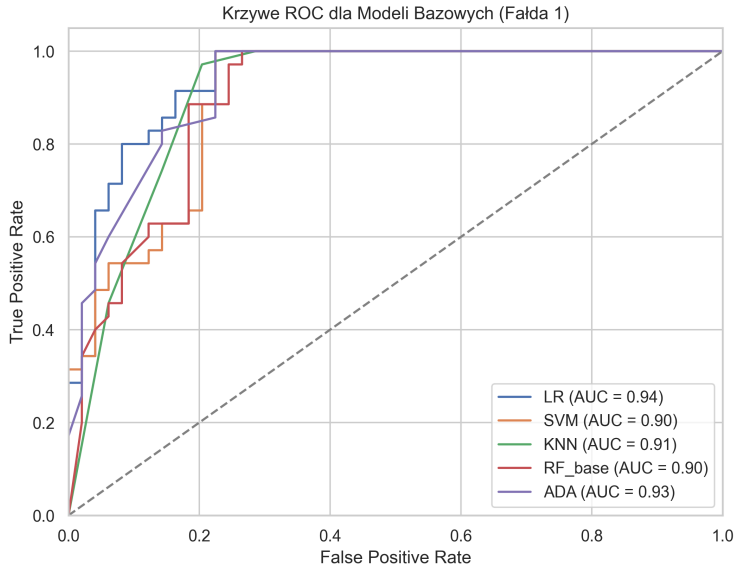
Jak widać, 18 cech zostało wybranych we wszystkich 10 fałdach, co wskazuje na ich wysoką stabilność i potencjalne znaczenie dla modelu predykcyjnego. Cecha `Gender` została wybrana w połowie przypadków, a `CSMD3` tylko raz. Cechy `FAT4`, `PIK3CA` oraz `BCOR`, które miały najniższe wartości IV, nie zostały wybrane w procesie głosowania w żadnej z fałd.

4.2 Wyniki indywidualnych modeli bazowych

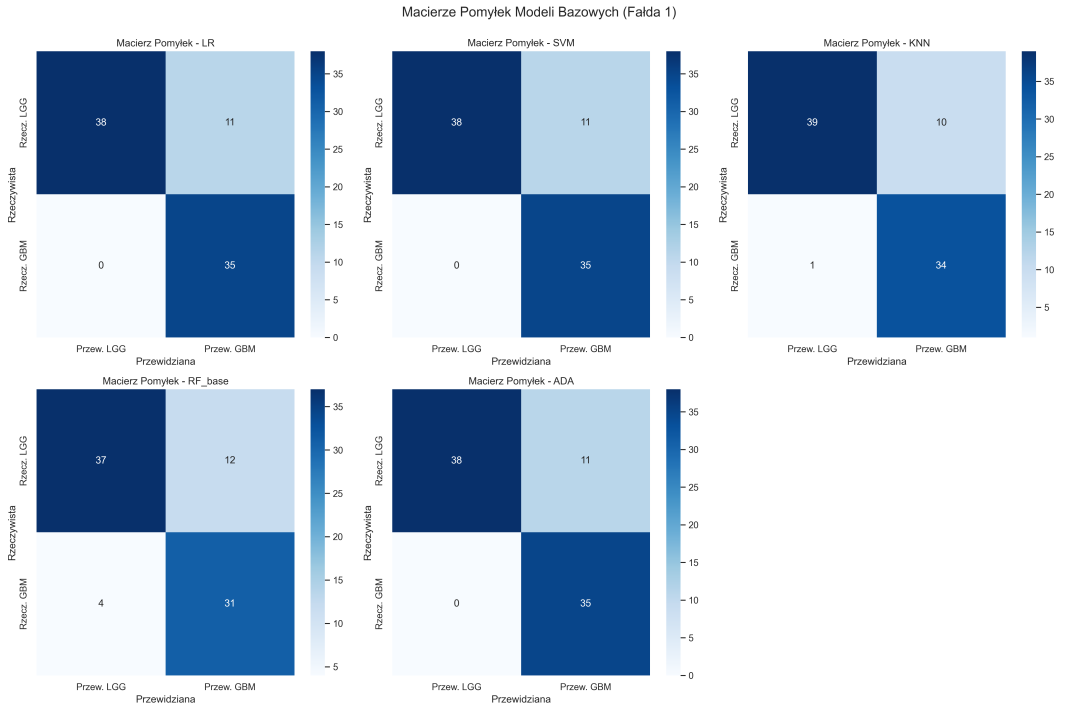
Po etapie selekcji cech, pięć modeli bazowych zostało wytrenowanych i ocenionych w każdej z fałd walidacji krzyżowej. Tabela 6 przedstawia uśrednione wyniki metryk wydajności dla tych modeli. Dodatkowo, dla pierwszej fałdy walidacji, wygenerowano krzywe ROC (Rysunek 19), macierze pomyłek (Rysunek 20) oraz wizualizację przykładowego drzewa z modelu Lasu Losowego (Rysunek 21).

Tabela 6. Średnie metryki wydajności dla indywidualnych modeli bazowych (po 10 fałdach walidacji krzyżowej). Wartości przedstawiono jako 'średnia ± odchylenie standardowe'. Wartości pogrubione wskazują najlepszy wynik w danej kolumnie.

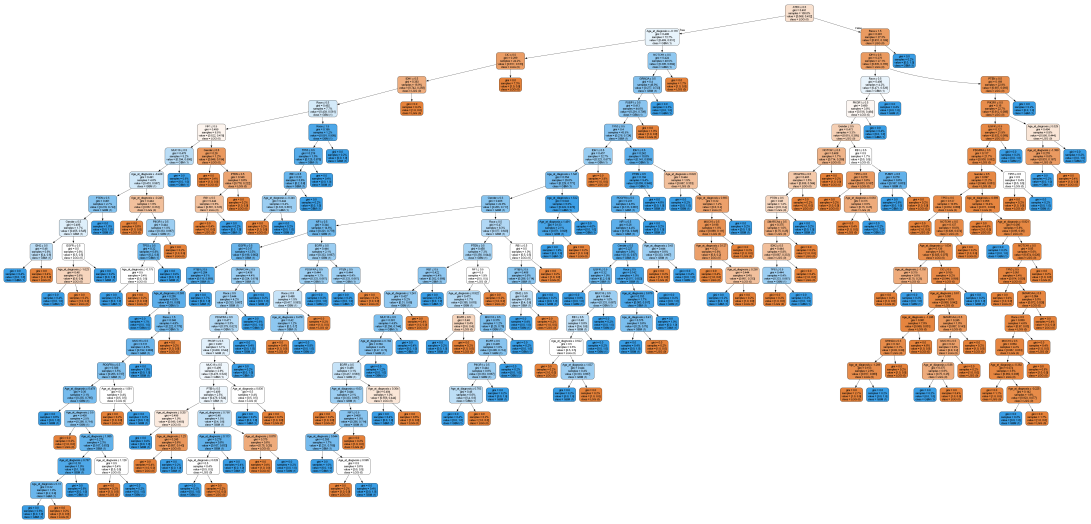
Model	ACC (Śr. ± Std)	AUC (Śr. ± Std)	F1 (Śr. ± Std)	PRE (Śr. ± Std)	REC (Śr. ± Std)	SPEC (Śr. ± Std)
LR	0.870 ± 0.040	**0.921 ± 0.030**	0.855 ± 0.044	0.809 ± 0.058	0.909 ± 0.053	0.842 ± 0.057
SVM	**0.872 ± 0.040**	0.908 ± 0.032	**0.856 ± 0.047**	**0.817 ± 0.059**	0.903 ± 0.065	**0.850 ± 0.056**
KNN	0.840 ± 0.036	0.896 ± 0.029	0.814 ± 0.042	0.799 ± 0.053	0.833 ± 0.065	0.846 ± 0.052
RF_base	0.830 ± 0.028	0.904 ± 0.025	0.799 ± 0.027	0.798 ± 0.057	0.804 ± 0.036	0.848 ± 0.056
ADA	0.866 ± 0.038	0.915 ± 0.026	0.853 ± 0.040	0.799 ± 0.059	**0.918 ± 0.043**	0.830 ± 0.061



Rysunek 19. Krzywe ROC dla modeli bazowych na danych walidacyjnych pierwszej fałdy.



Rysunek 20. Macierze pomyłek dla modeli bazowych na danych walidacyjnych pierwszej fałdy.



Rysunek 21. Wizualizacja przykładowego drzewa decyzyjnego z modelu Lasu Losowego (pierwsza fałda), wytrenowanego na cechach wybranych w procesie głosowania.

4.3 Wyniki modeli hybrydowych

Po ocenie indywidualnych modeli bazowych, przystąpiono do budowy i oceny 16 modeli zespołowych, zgodnie z kombinacjami przedstawionymi w pracy [1]. Modele te wykorzystywały mechanizm miękkiego głosowania (soft voting) na podstawie predykcji prawdopodobieństw uzyskanych z odpowiednich modeli bazowych, wytrenowanych na cechach wybranych w procesie hierarchicznej selekcji. Tabela 7 przedstawia uśrednione wyniki metryk wydajności dla wszystkich 16 modeli zespołowych, posortowane malejąco według średniej dokładności (ACC).

Tabela 7. Średnie metryki wydajności dla modeli zespołowych (soft voting) po 10 fałdach walidacji krzyżowej. Wartości przedstawiono jako 'średnia ± odchylenie standardowe'. Wartości pogrubione wskazują najlepszy wynik w danej kolumnie.

Model Zespołowy	ACC (Śr. ± Std)	AUC (Śr. ± Std)	F1 (Śr. ± Std)	PRE (Śr. ± Std)	REC (Śr. ± Std)	SPEC (Śr. ± Std)
Ens13-LR-SVM-RF-ADA	**0.871 ± 0.038**	0.923 ± 0.022	**0.856 ± 0.042**	0.812 ± 0.060	**0.909 ± 0.054**	0.844 ± 0.060
Ens11-LR-SVM-KNN-RF	**0.871 ± 0.040**	0.919 ± 0.026	0.854 ± 0.043	**0.820 ± 0.062**	0.895 ± 0.051	0.854 ± 0.060
Ens1-LR-SVM-KNN	0.870 ± 0.039	0.918 ± 0.028	0.854 ± 0.042	0.813 ± 0.063	0.903 ± 0.045	0.846 ± 0.061
Ens2-LR-SVM-RF	0.870 ± 0.042	0.922 ± 0.024	0.854 ± 0.046	0.815 ± 0.064	0.901 ± 0.054	0.848 ± 0.062
Ens12-LR-SVM-KNN-ADA	0.869 ± 0.036	0.919 ± 0.027	0.854 ± 0.037	0.810 ± 0.059	0.906 ± 0.044	0.842 ± 0.059
Ens9-SVM-RF-ADA	0.869 ± 0.036	0.919 ± 0.023	0.851 ± 0.038	0.819 ± 0.062	0.892 ± 0.052	0.852 ± 0.062
Ens16-LR-SVM-KNN-RF-ADA	0.869 ± 0.043	0.919 ± 0.025	0.852 ± 0.045	0.815 ± 0.065	0.898 ± 0.051	0.848 ± 0.064
Ens5-LR-KNN-ADA	0.868 ± 0.036	0.919 ± 0.026	0.852 ± 0.038	0.811 ± 0.064	0.901 ± 0.046	0.844 ± 0.062
Ens3-LR-SVM-ADA	0.868 ± 0.036	**0.925 ± 0.028**	0.852 ± 0.040	0.805 ± 0.053	**0.909 ± 0.053**	0.838 ± 0.053
Ens14-LR-KNN-RF-ADA	0.866 ± 0.036	0.918 ± 0.024	0.847 ± 0.040	0.819 ± 0.056	0.881 ± 0.056	**0.856 ± 0.054**
Ens6-LR-RF-ADA	0.865 ± 0.031	0.923 ± 0.021	0.847 ± 0.034	0.812 ± 0.049	0.889 ± 0.048	0.848 ± 0.048
Ens8-SVM-KNN-ADA	0.864 ± 0.039	0.916 ± 0.026	0.847 ± 0.041	0.811 ± 0.067	0.892 ± 0.049	0.844 ± 0.066
Ens15-SVM-KNN-RF-ADA	0.864 ± 0.038	0.916 ± 0.025	0.846 ± 0.041	0.813 ± 0.060	0.886 ± 0.060	0.848 ± 0.059
Ens7-SVM-KNN-RF	0.862 ± 0.033	0.913 ± 0.028	0.841 ± 0.037	0.816 ± 0.056	0.872 ± 0.060	0.854 ± 0.055
Ens4-LR-KNN-RF	0.859 ± 0.033	0.917 ± 0.026	0.838 ± 0.036	0.816 ± 0.057	0.867 ± 0.058	0.854 ± 0.056
Ens10-KNN-RF-ADA	0.851 ± 0.038	0.912 ± 0.025	0.827 ± 0.044	0.810 ± 0.059	0.850 ± 0.066	0.852 ± 0.056

Tabela 8. Wyniki obliczeniowe autorów pracy [1]. Wartości pogrubione wskazują najlepsze wyniki w danej kolumnie.

Zestaw #	Kombinacja Modeli	ACC	AUC	F1	PRE	REC	SPEC
1	LR + SVM + KNN	0.869	0.914	0.852	0.798	0.922	0.835
2	LR + SVM + RF	0.870	0.917	0.853	0.801	0.918	0.838
3	LR + SVM + AdaBoost	0.870	**0.920**	0.854	0.798	**0.925**	0.835
4	LR + KNN + RF	0.861	0.912	0.838	0.807	0.878	0.850
5	LR + KNN + AdaBoost	0.868	0.913	0.850	0.800	0.913	0.839
6	LR + RF + AdaBoost	0.858	0.917	0.834	0.807	0.870	**0.870**
7	SVM + KNN + RF	0.868	0.909	0.848	0.809	0.898	0.847
8	SVM + KNN + AdaBoost	0.871	0.911	0.854	0.803	0.918	0.840
9	SVM + RF + AdaBoost	**0.876**	0.914	**0.858**	**0.815**	0.912	0.852
10	KNN + RF + AdaBoost	0.845	0.908	0.815	0.804	0.835	0.854
11	LR + SVM + KNN + RF	0.874	0.913	0.856	0.808	0.918	0.844
12	LR + SVM + KNN + AdaBoost	0.869	0.915	0.852	0.798	0.922	0.835
13	LR + SVM + RF + AdaBoost	0.870	0.918	0.853	0.801	0.918	0.838
14	LR + KNN + RF + AdaBoost	0.862	0.914	0.839	0.809	0.878	0.852
15	SVM + KNN + RF + AdaBoost	0.868	0.911	0.848	0.809	0.898	0.847
16	LR + SVM + KNN + RF + AdaBoost	0.875	0.914	**0.858**	0.808	0.921	0.844

4.4 Wybór najlepszego modelu i porównanie z pracą referencyjną

W celu wyboru optymalnego modelu zespołowego, dokonano analizy porównawczej metryk wydajności uzyskanych w niniejszym badaniu (Tabela 7) z wynikami przedstawionymi w pracy referencyjnej Tasci et al. (2022) (Tabela 8). Głównym kryterium oceny była średnia dokładność (ACC) osiągnięta w 10-krotnej walidacji krzyżowej, wspierana przez analizę wartości AUC, F1-Score oraz stabilności predykcji (niskie odchylenie standardowe).

W oryginalnym badaniu Tasci et al. (2022) dla zbioru danych TCGA, najlepszą kombinacją okazał się zespół `Ens9` (SVM + RF + AdaBoost), osiągając dokładność 0.876, AUC 0.914 oraz F1-Score 0.858. W niniejszym projekcie, odpowiadający mu zespół `Ens9-SVM-RF-ADA` uzyskał bardzo zbliżone wyniki: ACC = 0.869 (różnica 0.007), AUC = 0.919 (różnica +0.005) oraz F1-Score = 0.851 (różnica 0.007). Tak niewielkie rozbieżności potwierdzają wysoką wierność odtworzenia oryginalnej metodologii.

Analizując wyniki uzyskane w ramach tego projektu (Tabela 7), najwyższą średnią dokładność, wynoszącą **0.871**, osiągnęły dwa modele zespołowe:

- `Ens13-LR-SVM-RF-ADA` (ACC: 0.871 ± 0.038, AUC: 0.923 ± 0.022, F1: 0.856 ± 0.042)
- `Ens11-LR-SVM-KNN-RF` (ACC: 0.871 ± 0.040, AUC: 0.919 ± 0.026, F1: 0.854 ± 0.043)

Oba te modele nie tylko dorównują, ale pod względem niektórych metryk (np. AUC dla `Ens13`) nieznacznie przewyższają model `Ens9` z oryginalnej publikacji. Warto również zwrócić uwagę na model `Ens3-LR-SVM-ADA`, który w naszym badaniu uzyskał najwyższe średnie AUC (0.925 ± 0.028) przy dokładności 0.868, co jest zgodne z obserwacją z pracy Tasci et al. (2022), gdzie ten sam zestaw (Set #3) również wykazywał najwyższe AUC (0.920).

Biorąc pod uwagę najwyższą osiągniętą dokładność (0.871), bardzo wysokie AUC (0.923) oraz konkurencyjny F1-Score (0.856), jako **najlepszy model** w niniejszym badaniu wybrano zespół `Ens13-LR-SVM-RF-ADA`. Model ten charakteryzuje się również dobrą stabilnością wyników, o czym świadczy stosunkowo niskie odchylenie standardowe dla kluczowych metryk. Jego wydajność jest w pełni porównywalna z najlepszymi wynikami zgłoszonymi w pracy referencyjnej, co potwierdza skuteczność zastosowanej hierarchicznej metody selekcji cech i budowy modeli zespołowych.

5 Przykład użycia modelu

Aby zademonstrować praktyczne zastosowanie opracowanego modelu, przeprowadzono predykcję dla dwóch sztucznie stworzonych profili pacjentów przy użyciu najlepszego wybranego zespołu – `Ens13_LR_SVM_RF_ADA`, wytrenowanego na pełnym zbiorze danych z wykorzystaniem 18 cech wybranych jako najbardziej stabilne w procesie walidacji krzyżowej.

Pacjent 1 (profil sugerujący LGG):

- Wiek w momencie diagnozy: 35 lat (wartość znormalizowana: -1.0154)
- Rasa: Biała (`Race`=0)
- Status mutacji `IDH1`: Zmutowany (`IDH1`=1)
- Status mutacji `ATRX`: Zmutowany (`ATRX`=1)
- Status mutacji `CIC`: Zmutowany (`CIC`=1)
- Status mutacji `TP53`: Zmutowany (`TP53`=1)
- Status mutacji pozostałych 14 wybranych cech: Niezmutowany (wartość 0)

Predykcja dla Pacjenta 1:

- Przewidziana klasa: **0 (LGG)**
- Prawdopodobieństwo dla LGG (Grade 0): **0.7512**
- Prawdopodobieństwo dla GBM (Grade 1): **0.2488**

Wynik ten jest zgodny z oczekiwaniami, wskazując z wysoką pewnością (ok. 75%) na glejaka o niższym stopniu złośliwości.

Pacjent 2 (profil sugerujący GBM):

- Wiek w momencie diagnozy: 65 lat (wartość znormalizowana: 0.8962)
- Rasa: Biała (`Race`=0)
- Status mutacji `IDH1`: Niezmutowany (`IDH1`=0)
- Status mutacji `EGFR`: Zmutowany (`EGFR`=1)
- Status mutacji `PTEN`: Zmutowany (`PTEN`=1)
- Status mutacji pozostałych 15 wybranych cech: Niezmutowany (wartość 0)

Predykcja dla Pacjenta 2:

- Przewidziana klasa: **1 (GBM)**
- Prawdopodobieństwo dla LGG (Grade 0): **0.2563**
- Prawdopodobieństwo dla GBM (Grade 1): **0.7437**

Również w tym przypadku model z wysoką pewnością (ok. 74%) wskazał na glejaka wielopostaciowego, co jest zgodne z charakterystyką profilu.

Powyższe przykłady ilustrują potencjalną użyteczność opracowanego modelu w procesie wspomagania decyzji diagnostycznych, dostarczając nie tylko przewidywanej klasy, ale również oszacowania prawdopodobieństwa tej predykcji.

6 Podsumowanie i wnioski

Celem niniejszego projektu było odtworzenie i ocena zaawansowanej metodologii klasyfikacji stopnia złośliwości glejaka, zaprezentowanej w pracy Tasci et al. (2022), na podstawie publicznie dostępnego zbioru danych klinicznych i molekularnych. W ramach projektu zaimplementowano kompletny potok uczenia maszynowego, obejmujący wstępną analizę danych, wieloetapową selekcję cech oraz budowę i ewaluację modeli zespołowych.

Zastosowana hierarchiczna metoda selekcji cech, oparta na głosowaniu czterech niezależnych technik (IV, RFE, ważność cech z Lasów Losowych oraz LASSO), okazała się wysoce skuteczna i stabilna. Analiza wykazała, że 18 z 23 dostępnych cech było konsekwentnie wybieranych we wszystkich 10 fałdach walidacji krzyżowej, co świadczy o ich dużym znaczeniu dla problemu klasyfikacji. Do najważniejszych predyktorów, zidentyfikowanych zarówno na etapie EDA, jak i w formalnej selekcji, należą status mutacji genu `IDH1` oraz wiek pacjenta w momencie diagnozy.

W fazie modelowania oceniono 16 różnych modeli zespołowych opartych na mechanizmie miękkiego głosowania. Najlepszy z nich, `Ens13_LR_SVM_RF_ADA`, osiągnął średnią dokładność (ACC) na poziomie **0.871** przy wartości AUC równej **0.923**. Wynik ten jest w pełni porównywalny z dokładnością 0.876 uzyskaną w pracy referencyjnej, co potwierdza wysoką wierność odtworzenia metodologii i jej dużą skuteczność. Przeprowadzona demonstracja na sztucznie stworzonych profilach pacjentów dodatkowo zilustrowała praktyczną zdolność modelu do rozróżniania przypadków o niskim i wysokim stopniu złośliwości guza z dużą pewnością. Wnioskiem końcowym projektu jest potwierdzenie, że zastosowanie hierarchicznego podejścia do selekcji cech w połączeniu z modelami zespołowymi jest efektywną strategią w złożonych problemach biomedycznych. Osiągnięte rezultaty walidują wnioski z pracy Tasci et al. (2022) i pokazują potencjał tej metody do budowy solidnych i dokładnych modeli predykcyjnych. Dalsze badania mogłyby obejmować optymalizację hiperparametrów poszczególnych modeli i metod selekcji, a także testowanie tej metodologii na innych, podobnych zbiorach danych.

Literatura

[1] Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., & Krauze, A. V. (2022). *Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics.* International Journal of Molecular Sciences, 23(22), 14155.

[2] Tasci, E., Zhuge, Y., et al. (2022). *Glioma Grading Clinical and Mutation Features Data Set.* UCI Machine Learning Repository. Dostęp: https://archive.ics.uci.edu/dataset/759/glioma+grading+clinical+and+mutation+features.

[3] Glejak - Wikipedia: https://pl.wikipedia.org/wiki/Glejak.

[4] Regresja logistyczna - Wikipedia: https://pl.wikipedia.org/wiki/Regresja_logistyczna.

[5] Maszyna wektorów nośnych - Wikipedia: https://pl.wikipedia.org/wiki/Maszyna_wektor%C3%B3w_no%C5%9Bnych.

[6] Metoda k-najbliższych sąsiadów - Wikipedia: https://pl.wikipedia.org/wiki/K_najbli%C5%BCszych_s%C4%85siad%C3%B3w.

[7] Lasy losowe - Wikipedia: https://pl.wikipedia.org/wiki/Las_losowy.

[8] AdaBoost - Wikipedia: https://pl.wikipedia.org/wiki/AdaBoost.

[9] RFE: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html.

[10] Lasso - Wikipedia: https://en.wikipedia.org/wiki/Lasso_(statistics).