Katarzyna Kobylińska

Metody wyjaśnialnego uczenia maszynowego dla danych tabelarycznych z przykładami zastosowań w medycynie

Rozprawa doktorska

Promotorzy rozprawy:
prof. dr hab. inż. Przemysław Biecek
Instytut Informatyki, Uniwersytet Warszawski

prof. dr. hab. n. med. Mariusz Adamek
Wydział Nauk Medycznych, Śląski Uniwersytet Medyczny

Warszawa, 2024

Streszczenie

Rozwój metod sztucznej inteligencji doprowadził do zbudowania zaawansowanych systemów zdolnych do analizy ogromnych zbiorów danych i podejmowania decyzji w wielu dziedzinach życia. Do szerokiego zastosowania sztucznej inteligencji przyczyniła się łatwość użycia, a także zwykle większa precyzja wyników niż tradycyjnych modeli statystycznych. Zauważono także możliwości wykorzystania sztucznej inteligencji w medycynie, co daje obiecujące perspektywy w poprawie diagnostyki, personalizacji leczenia i przyczynia się do postępu w opiece zdrowotnej. Jednak szczególnie w dziedzinie obarczonej dużym ryzykiem, należy zachować ostrożność w zastosowaniu metod sztucznej inteligencji. Istotnym aspektem modelowania stała się metodologia pomagająca w zrozumieniu modeli uczenia maszynowego nazwana wyjaśnialną sztuczną inteligencją.

Celem tej pracy jest zbadanie czy użycie metod wyjaśnialnego uczenia maszynowego poprawia jakość wnioskowania na podstawie modelowania. Hipotezę badawczą popieram modelowaniem i eksperymentami na danych medycznych, sprawdzeniem czy zastosowanie wyjaśnialnej sztucznej inteligencji na różnych etapach procedury budowy modelu może wspomóc ten proces oraz zaproponowanym przeze mnie nowym algorytmem. Wkład w nauki medyczne stanowi zbudowanie modeli i zastosowanie wyjaśnialnej sztucznej inteligencji do rzeczywistych problemów medycznych, które są wykorzystywane w dużych ośrodkach medycznych w Polsce. Wkładem w uczenie maszynowe i wyjaśnialną sztuczną inteligencję jest nowa miara PDI służąca do porównania modeli na podstawie ich wyjaśnień oraz nowy algorytm `Rashomon_DETECT` służący do wyboru najbardziej różniących się modeli ze zbioru niemal optymalnych modeli. Zaproponowałam również nowe podejście do procesu budowy modeli oparte o wyjaśnialną sztuczną inteligencję i rozszerzenie do rozpatrywania zbioru modeli Rashomon.

Rozprawa powstała podczas międzywydziałowych interdyscyplinarnych studiów doktoranckich matematyczno- przyrodniczych. Łączy wiedzę i metodologię z dwóch dziedzin: informatyki stosowanej i nauk medycznych. Interdyscyplinarność tej pracy polega

na integracji metod z dziedziny informatyki z teoriami i wiedzą medyczną w celu analizy złożonych problemów medycznych.

Słowa kluczowe

wyjaśnialna sztuczna inteligencja, uczenie maszynowe, proces budowy modelu, zbiór modeli Rashomon

Dziedzina pracy

11.3 Informatyka

Klasyfikacja tematyczna

Computing methodologies → Machine learning
Applied computing → Life and medical sciences

Tytuł pracy w języku angielskim

Explainable Machine Learning methods for tabular data with applications in medicine.

Oświadczenie kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie stopnia doktora w dziedzinie nauk ścisłych i przyrodniczych w dyscyplinie informatyka.

Promotor 1: prof. dr hab. inż Przemysław Biecek

Data Podpis

Promotor 2: prof. dr hab. n. med. Mariusz Adamek

Data Podpis

Oświadczenie autora pracy

Oświadczam, że niniejsza rozprawa doktorska została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem stopnia doktora w innej jednostce. Niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Katarzyna Kobylińska

Data Podpis

Spis treści

Spis rysunków

Spis tabel

Słownik terminów

accumulated local effects profil ALE. 97

acute physiology and chronic health evaluation II skala APACHE II. 80

artificial intelligence sztuczna inteligencja. 19

bias obciążenie. 93

black box czarna skrzynka. 17

body mass index indeks masy ciała. 69

boosting metoda wzmocnienia. 35

bootstrap sample próba bootstrapowa. 34

class activation mapping mapowanie ścieżek aktywacji klas. 39

cross-industry standard process for data mining metodyka CRISP-DM. 45

decision tree split podział. 32

eXplainable Artificial Intelligence wyjaśnialna sztuczna inteligencja. 17

fairness stronniczość modeli. 37

general data protection regulation rozporządzenie o ochronie danych osobowych. 19

gradient boosting wzmocnienie gradientowe. 35

layer-wise relevance propagation metoda propagowania istotności przez warstwy. 39

learning rate współczynnik uczenia się. 36

machine learning uczenie maszynowe. 17

mean misclassification error błąd klasyfikacji. 55

model agnostic metody niezależne od rodziny modeli. 38

model class reliance metoda ważności zmiennych MCR. 92

model development process metodyka MDP. 45

model specific metody specyficzne dla rodziny modeli. 38

national early warning score narodowa skala wczesnego ostrzegania, skala NEWS. 80

OOB error błąd OOB. 34

out-of-bag-instances przykłady OOB. 34

partial dependence plot metoda cząstkowej zależności. 41

permutation feature importance permutacyjna ważność zmiennych. 41

prediction predykcja. 31

profile disparity index miara rozbieżności profili, PDI. 98

random forest las losowy. 34

Rashomon set's volume miara wielkości zbioru Rashomon. 92

regression zadanie regresji. 31

residual sum of squares błąd RSS. 91

root mean squared error błąd średniokwadratowy. 33

sepsis severity score skala SSS. 80

sequential organ failure assessment skala niewydolności narządów związanych z sepsą SOFA. 80

Shapley additive explanations wartości Shapleya, SHAP. 42

simplified acute physiology skala SAPS. 80

supervised learning uczenie nadzorowane. 31

support vector machines maszyny wektorów podpierających, SVM. 105

target attribute atrybut docelowy. 31

tasks zadania. 31

unsupervised learning uczenie nienadzorowane. 31

variable importance ważność zmiennych. 41

Rozdział 1

Wprowadzenie

W niniejszej rozprawie skupiłam się na metodach wyjaśnialnej sztucznej inteligencji (ang. *eXplainable Artificial Intelligence, XAI*) zastosowanych do problemów z dziedziny nauk medycznych. W pracy sprawdziłam przydatność metod XAI na poszczególnych etapach procesu rozwoju modelu uczenia maszynowego. Proponuję także nową metodę, algorytm służący do automatycznego porównania modeli na podstawie wyjaśnialnej sztucznej inteligencji.

1.1. Motywacja

Modele z dziedziny uczenia maszynowego (ang. *machine learning, ML*) są coraz częściej stosowane w diagnostyce, medycynie spersonalizowanej [56] czy badaniach przesiewowych [104]. Dzieje się tak, ponieważ metody te są zarówno dokładne, jak i stosunkowo łatwe do wdrożenia. Niestety, często odbywa się to kosztem interpretowalności. Modele uczenia maszynowego o skomplikowanej strukturze nazywane są modelami czarnej skrzynki (ang. *black box*).

W naukach medycznych dysponujemy ogromnymi ilościami danych klinicznych, takich jak badania diagnostyczne, obrazy medyczne, czy dane laboratoryjne pacjentów [24]. Dostępność takich baz danych umożliwia budowanie dokładnych algorytmów wykorzystujących dane w spersonalizowanych prognozach [124]. Uczenie maszynowe pozwala na analizę i wykorzystanie tak dużych baz danych w celu wykrywania wzorców, identyfikacji czynników ryzyka [5], prognozowania wyników medycznych i wspomaga-

nia procesów decyzyjnych [102]. W medycynie wiele przypadków ma złożoną strukturę, a diagnoza i leczenie wymagają analizy szeregu czynników, a także interakcji między nimi. Uczenie maszynowe może pomóc w identyfikacji skomplikowanych wzorców, trudnych do zauważenia przez człowieka [21]. Zaawansowane modele uczenia maszynowego mogą wspomagać wykrywanie chorób, rekomendować optymalne leczenie i przewidywać wyniki medyczne [4].

Uczenie maszynowe może okazać się także pomocne w optymalizacji procesów medycznych, na przykład planowania leczenia, zarządzania przepływem pracy w szpitalu [119, 127]. Poprzez analizę danych i automatyzację, uczenie maszynowe może przyczynić się do skrócenia czasu diagnostyki, poprawy wyników terapeutycznych i optymalnego wykorzystania zasobów medycznych. Modelowanie może pomóc w wykrywaniu rzadkich zdarzeń medycznych, które mogą być trudne do zidentyfikowania przez lekarzy ze względu na ich niską częstość występowania [72]. Zaawansowane modele uczenia maszynowego mogą analizować duże zbiory danych i wykrywać nietypowe wzorce, co może przyczynić się do wczesnego wykrywania rzadkich chorób.

Zautomatyzowane podejmowanie decyzji ma szereg zalet, takich jak standaryzacja, precyzja, łatwość trenowania, szybsze wdrożenie i większa efektywność. Jednak niesie ze sobą również pewne zagrożenia. Modele istnieją po to, by wspierać ludzkie decyzje. Złożoność podejścia uczenia maszynowego w medycynie może skutkować niezrozumieniem modelu przez lekarzy, a w konsekwencji prowadzić do podejmowania błędnych i potencjalnie szkodliwych decyzji. W przypadku decyzji o wysokiej stawce, zwłaszcza w naukach medycznych, kluczowe znaczenie ma zrozumienie czynników, które wpływają na decyzje modelu. Brak dobrego zrozumienia modelu stwarza poważne ryzyko w zastosowaniach. Z tej perspektywy, algorytmy czarnych skrzynek okazują się wyzwaniem [92].

Odpowiedzią na trudności w interpretacji złożonych modeli uczenia maszynowego są metody XAI [91]. Mają one pomóc w zrozumieniu działania modeli oraz w efektywnym ich wykorzystaniu. Ponadto ułatwiają znalezienie możliwych błędów modeli, a także służą do odkrywania nowych koncepcji i zależności. Wyjaśnienie modeli uczenia maszynowego może także służyć do budowania zaufania do decyzji podjętych na podstawie wyników modeli zarówno dla twórców jak i ich użytkowników, co w przypadku zastosowania ich w medycynie jest kluczowe. W ostatniej dekadzie, wraz z większą dostępnością do dużych baz danych, coraz częstszym ich wykorzystywaniem do budo-

wania systemów opartych na złożonych modelach, pojawiło się także zapotrzebowanie na metody służące do ich wyjaśnienia [109].

Organy regulacyjne również coraz częściej wzywają do przejrzystości w procesie podejmowania decyzji w zakresie sztucznej inteligencji. Prowadzi to do opracowania standardów i wytycznych XAI oraz rozwoju metodologii. W 2018 roku, Unia Europejska wprowadziła ogólne rozporządzenie o ochronie danych osobowych (ang. *general data protection regulation, GDPR*)[37]. Zgodnie z tą regulacją, osoba objęta badaniem powinna uzyskać nie tylko samą odpowiedź modelu. Rozporządzenie wymaga przejrzystości każdej pojedynczej decyzji opartej na zaawansowanych algorytmach. W ostatnich regulacjach podkreślono także potrzebę lepszej współpracy między algorytmami uczenia maszynowego (ML) a ludźmi. Komisja Europejska w dokumentach [36, 35] wskazała znaczenie nadzorczej roli człowieka nad algorytmami, która jest możliwa tylko wtedy, gdy człowiek rozumie zarówno ograniczenia, jak i zalety systemu ML. W 2024 roku, Unia Europejska wydała rozporządzenie dotyczące sztucznej inteligencji [22], które ma celu zapewnienie by rozwój i wykorzystanie sztucznej inteligencji przebiegały w sposób bezpieczny i wiarygodny. Duże agencje badawcze, takie jak DARPA [27], pracują nad technikami usprawniającymi współpracę między modelem a jego użytkownikiem. W konsekwencji, potrzeba lepszego zrozumienia algorytmów doprowadziła do znacznego rozwoju XAI i powstaniach nowych metod interpretowania wyników model.

Obecnie XAI jest aktywnym obszarem badań w celu opracowania metod które pomogą w interpretowalności modeli sztucznej inteligencji (ang. *artificial intelligence, AI*). Naukowcy badają różne techniki wyjaśniania decyzji systemów sztucznej inteligencji, takie jak generowanie wyjaśnień w języku naturalnym, wizualizacja procesów decyzyjnych i wykorzystywanie scenariuszy alternatywnych do zrozumienia czynników, które wpłynęły na decyzję. Metody te pomagają lepiej zrozumieć działanie modelu, a tym samym zwiększyć zaufanie użytkownika do jego funkcjonowania. Ułatwiają także wykrywanie błędów w przewidywaniach modelu, czyli takich, które są niezgodne z wiedzą dziedzinową [48]. Wspomagają zrozumienie modelu, autorzy w pracy [45] zbudowali model przewidujący koncentrację dwutlenku azotu w powietrzu i wyjaśnili jego działanie przy pomocy metody XAI.

Metody wyjaśnialnego uczenia maszynowego zyskują również popularność w zastosowaniach medycznych. Wyjaśnialne modele są wykorzystywane do analizy zarówno danych tabelarycznych jak i danych obrazowych zdjęć rentgenowskich, tomografii kom-

puterowej czy ultradźwięków. Sana Tonekaboni wraz ze współautorami [122] przedstawiają przegląd metod XAI zastosowanych w problemach medycznych z odniesieniami do szeregu publikacji, natomiast Ruey-Kai Sheu i Mayuresh Sunil Pardeshi [107] prezentują wnikliwą analizę dotyczącą kategoryzacji metod wyjaśnialnego uczenia maszynowego i odniesienia ich do dziedziny medycyny. Bas H.M. van der Velden [129] i Amitojdeep Singh [112] wraz ze współautorami przedstawiają dwie prace, w których wykorzystano metody XAI w analizie danych medycznych obrazowych. Chiara Zucco wraz ze współautorami [143] dokonuje przeglądu podejść dotyczących analizy sentymentu, jej wyjaśnienia i możliwości zastosowania w obszarze medycyny. Hans-Christian Thorsen-Meyer [121] demonstruje jak stosować metody wyjaśniające w analizie modeli głębokich sieci neuronowych dla elektronicznych rekordów pacjentów o wysokiej częstotliwości. Armin Thomas wraz ze współautorami [120] prezentuje platformę do analizy i modelowania danych obrazowych pochodzących z rezonansu magnetycznego mózgu, zachowując przy tym interpretowalność metod poprzez użycie XAI. Kolejnym przykładem zastosowania wyjaśnialnej sztucznej inteligencji do danych obrazowych jest praca Guannan Zhao [140]. W artykule Ziqi Tang [118] opisał możliwą do interpretacji, opartą na sieciach neuronowych metodę klasyfikacji patologii u osób z chorobą Alzheimera. Przykładem danych obrazowych w medycynie są dane dermatologiczne. Arieh Gomolin wraz ze współautorami [52] opisał potrzebę i możliwości zastosowania sztucznej inteligencji do takich baz danych. Autorzy podsumowują swoją analizę stwierdzeniem, że modele, których nie da się wyjaśnić nie są w stanie zastąpić oceny dermatologa co ogranicza możliwości zastosowania czarnych skrzynek w tej dziedzinie. Kluczowa zatem jest umiejętność wyjaśnienia takich modeli. Andreas Holzinger [58] oraz Yao Xie [136] wraz ze współautorami skupiają się z kolei na analizie danych radiologicznych wspartej metodami XAI. Stephanie Hyland [61] zastosowała złożone modele uczenia maszynowego do problemu wczesnego wykrywania niewydolności krążenia na Oddziale Intensywnej Terapii. Scott Lundberg [83] przedstawił wyjaśnienie modeli przewidywania hipoksemii podczas operacji. Simon Meyer Lauritsen [71] przedstawił rozwiązania w zakresie wykrywania ostrych uszkodzeń nerek (AKI) i ostrych urazów płuc (ALI). Christoph Jansen [62] przedstawił analizę opartą o sieci neuronowe z interpretacją modelu, która ujawnia nowe, ważne czynniki wpływające na bezsenność.

Potrzeba dokładnego zrozumienia modeli w zastosowaniach medycznych jest także podkreślona w artykule Richa Caruana [18]. Artykuł ten na przykładach modelowania

medycznego pokazuje, że nie należy stosować modeli wyłącznie na podstawie ich miary jakości. Modele, szczególnie te używane w medycynie, powinny być dokładnie zrozumiane przed ich wdrożeniem. W kontekście zastosowania modeli uczenia maszynowego w obszarze medycyny, także ważnym aspektem jest zaufanie do modelowania przez klinicystów. Sana Tonekaboni wraz ze współautorami [123] zidentyfikowali konkretne aspekty wyjaśnialności, które mogą pomóc w budowaniu zaufania do modeli uczenia maszynowego. Praca wskazuje klasy wyjaśnień, które klinicyści uznali za najbardziej istotne dla skutecznego wdrożenia modeli w praktykę lekarską.

Z drugiej strony, pojawiają się też głosy krytyki. Alex John London w swojej pracy [77] pokazał, że pewien stopień niezrozumienia może być akceptowalny. Może okazać się istotniejsze, aby uzyskać dokładne wyniki, które są empirycznie zweryfikowane, niż za bardzo skupiać się na tym, jak wyjaśnić model czarnej skrzynki. Andreas Holzinger [59] z kolei rozróżnia przyczynowość decyzji od wyjaśnienie modelu. Wskazuje by w naukach medycznych nie ograniczać się jedynie do wyjaśnialnej sztucznej inteligencji, ale wesprzeć modelowanie także o znalezienie przyczyny. Przyczynowością nazywa stopień w jaki ekspert z danej dziedziny jest w stanie zrozumieć przyczynę problemu, a nie samo działanie modelu. Arun Das [23] z kolei wskazuje jako główne wyzwanie wyjaśnialnej sztucznej inteligencji ocenę i porównanie kilku proponowanych algorytmów na rzeczywistych bazach danych.

1.2. Cel pracy

Celem rozprawy jest zbadanie metod wyjaśnialnego uczenia maszynowego w kontekście poprawy jakości wnioskowania na podstawie modelowania. Przedstawiam potencjał istniejących metod XAI w budowaniu modeli, identyfikuję luki w tym obszarze i proponuję nową metodę rozszerzającą XAI.

Podczas moich interdyscyplinarnych studiów doktoranckich skupiłam się na wykorzystaniu metod XAI stosowanych do analizy wytrenowanych wcześniej modeli. Metody te nazywać będę metodami XAI post-hoc. Sprawdziłam czy metody te mogą poprawić jakość modelowania poprzez weryfikację poprawności działania modeli uczenia maszynowego, wspomóc wybór optymalnego modelu w procesie modelowania, czy wzmocnić zaufanie do wyników modeli w czasie wdrażania modelu. W ramach moich badań zauważyłam lukę w istniejących analizach, które koncentrują się na wyborze pojedynczego

modelu. W procesie wyboru modelu zazwyczaj analizuje się modele minimalizujące zadaną funkcję straty. W trakcie moich badań pokazałam, że modele o zbliżonej jakości mogą traktować dane w inny sposób. W ten sposób zidentyfikowałam lukę w analizach, które koncentrują się na pojedynczym modelu, zamiast na szeregu niemal optymalnych modeli. Zaproponowałam technikę ułatwiającą analizę takiego zbioru przez identyfikację najbardziej zróżnicowanych modeli. Efektywność nowej metody została zilustrowana przykładami modelowania w dziedzinie medycyny oraz symulacjami.

W rozprawie pokazuję, że zastosowanie metod XAI prowadzi do usprawnienia procesu modelowania poprzez lepszą interpretację i zrozumienie działania modeli uczenia maszynowego. Ponadto zaproponowana metoda ma na celu wspomóc obecne metody XAI umożliwiając automatyczne porównanie modeli. Celem tej pracy jest również przyczynienie się do rozwinięcia dziedziny XAI w kontekście nauk medycznych, co może mieć znaczący wpływ na doskonalenie modelowania i stosowanie uczenia maszynowego w tym obszarze.

1.3. Hipotezy badawcze

W niniejszej rozprawie doktorskiej sformułowałam ogólną hipotezę, która postuluje, że **użycie metod wyjaśnialnego uczenia maszynowego post-hoc poprawia jakość wnioskowania na podstawie modelowania**. Metody XAI mogą pomóc w zrozumieniu, dlaczego model dokonuje określonych wyborów, pozwalają na identyfikację potencjalnych problemów, takich jak błędy modelu wynikające z braku zgodności z wiedzą domenową, czy z przeuczenia. Ponadto istnieje możliwość dalszego rozszerzania i udoskonalania tych metod. Wyjaśnienia wyników modeli uczenia się są również istotne w kontekście zwiększania zaufania użytkowników do decyzji podejmowanych przez modele. Szczególnie jest to istotne w zastosowaniach do obszarów o wysokim ryzyku. W ramach niniejszej pracy przeprowadziłam analizę i badania w dziedzinie nauk medycznych w celu empirycznego potwierdzenia tych tez.

Na podstawie powyższej hipotezy głównej, postawiłam również następujące szczegółowe hipotezy badawcze:

1. Metody wyjaśnialnego uczenia maszynowego post-hoc poprawiają jakość modelowania rozumianą jako dokładność predykcji lub poprawność modelu. Zastoso-

wanie tych technik pozwala na osiągnięcie lepszych wyników modeli uczenia maszynowego w medycynie. Analiza tej hipotezy jest przeprowadzona w rozdziale 3.1

2. Metody wyjaśnialnego uczenia maszynowego post-hoc identyfikują słabe strony modeli. Techniki te pozwalają na identyfikację ograniczeń i niedoskonałości modeli. Analiza tej hipotezy jest przeprowadzona w rozdziale 3.2.

3. Metody wyjaśnialnego uczenia maszynowego post-hoc wspomagają zrozumienie i interpretację wyników modelowania. Techniki te umożliwiają bardziej przejrzyste wyjaśnienie działania modeli i pomagają użytkownikom zrozumienie procesu podejmowania decyzji przez modele. Analiza tej hipotezy została przeprowadzona w rozdziale 3.3.

4. Na podstawie metod wyjaśnialnego uczenia maszynowego post-hoc opracowałam metodę automatycznego wyboru podzbioru najbardziej różnych modeli spośród wielu, dobrych modeli. Analiza tej hipotezy została przeprowadzona w rozdziale 4.

Przez przeprowadzenie analizy, badań i eksperymentów w obszarze nauk medycznych, niniejsza rozprawa doktorska ma na celu udowodnienie powyższych hipotez badawczych, co przyczyni się do rozwinięcia dziedziny wyjaśnialnego uczenia maszynowego oraz poprawy jakości modelowania w kontekście medycyny.

1.4. Publikacje

Niniejsza rozprawa doktorska opiera się na serii artykułów naukowych opublikowanych w czasopismach naukowych lub przedstawionych na konferencji i opublikowanych w materiałach pokonferencyjnych. Obecnie jeden z artykułów jest w procesie recenzji, oczekuje na publikację. Artykuły te zostały przedstawione zbiorczo w Tabeli 1.1. W dalszych częściach rozprawy przedstawiam szczegółowo artykuły naukowe, ich cele oraz sposób realizacji. Całość pracy stanowi kompleksowe spojrzenie na modelowanie przy pomocy uczenia maszynowego na przykładzie zastosowania do problemów medycznych oraz wykorzystanie metod XAI w celu lepszego zrozumienia i interpretacji predykcji

modeli. Przedstawione artykuły stanowią wkład w dziedzinę nauki, a ich wyniki mogą znaleźć praktyczne zastosowanie w obszarze medycyny.

W ramach studiów doktorskich uczestniczyłam w projektach badawczych, których wyniki zostały udokumentowane w postaci artykułów naukowych, wymienionych w Tabeli 1.2.

Należy podkreślić, że publikacje te nie stanowią integralnej części niniejszej rozprawy doktorskiej. Natomiast doświadczenia zdobyte podczas tych projektów badawczych odegrały istotną rolę w formułowaniu hipotez, które zostały poddane analizie w rozprawie doktorskiej.

W ramach współpracy naukowej z Katedrą i Kliniką Hematologii, Transplantologii i Chorób Wewnętrznych, przeprowadziłam szereg badań dotyczących pacjentów chorych na choroby hematologiczne.

Wspólnie z dr n. med. Joanną Drozd Sokołowską przeprowadziliśmy analizę dotyczącą przeżywalności pacjentów cierpiących na chłoniaka, u których dodatkowo zdiagnozowano cukrzycę. Efektem tej kooperacji jest praca naukowa [31]. Badałyśmy także metody leczenia pacjentów z białaczką, którzy jednocześnie zmagają się z COVID-19 [32].

Wraz z dr n. med. Joanną Waszczuk-Gajdą przeanalizowałam poziom graniczny kreatyniny u pacjentów dotkniętych szpiczakiem [131].

Współpracując z dr n. med. Rafałem Machowiczem badałam pacjentów cierpiących na rzadką chorobę układu krwionośnego, HLH (Hemophagocytic lymphohistiocytosis). Wyniki analizy opisowej przedstawione zostały na konferencji, a baza danych posłużyła mi do rozwoju metodologii zaproponowanej w rozdziale 4.

Współpracując z Kliniką Anestezjologii i Intensywnej Terapii Uniwersyteckiego Szpitala Klinicznego we Wrocławiu przeanalizowałam bazę danych pacjentów znajdujących się na Oddziale Intensywnej Terapii ze zdiagnozowaną SEPSą. W ramach analizy bazy danych powstał artykuł [113]. Drugie badanie [74], korzystające z tej samej bazy danych, jest integralną częścią rozprawy i zostało szczegółowo opisane w rozdziale 3.3.

W ramach współpracy z dr n. med. Tomaszem Berusem pracującym w Wojskowym Szpitalu Klinicznym we Wrocławiu, przeanalizowałam bazę danych dotyczącą pacjentów z czerniakiem oka. W ramach tego badania została opublikowana praca przedstawiająca prognostyczną rolę ekspresji PLK-1 u pacjentów z czerniakiem błony naczyniowej oka [7].

Podczas realizacji projektów badawczych mogłam uczestniczyć w licznych spotkaniach z osobami odpowiedzialnymi za zbieranie i zarządzanie bazami danych, a także z lekarzami, którzy na co dzień pracują z osobami, których dane są gromadzone w analizowanych bazach. Te interakcje pozwoliły mi zdobyć wiedzę na temat rzeczywistych potrzeb lekarzy oraz zrozumieć ich perspektyw w kontekście modelowania danych medycznych. Poznanie realnych potrzeb lekarzy i zrozumienie ich perspektywy pozwoliło mi skoncentrować się na budowaniu modeli, które odpowiadają na konkretne wyzwania związane z daną chorobą. Przyjęcie perspektywy lekarza w procesie modelowania ma kluczowe znaczenie, ponieważ ostateczne wyniki i rozwiązania muszą być akceptowalne i możliwe do wdrożenia w szpitalach, przynosząc realne korzyści i poprawiając codzienną praktykę medyczną.

Dzięki tym współpracom naukowym zyskałam również głębsze zrozumienie znaczenia interpretowalności modeli w kontekście medycyny. Lekarze muszą być w stanie zrozumieć, jak działają modele i jakie wyniki dostarczają, aby móc na tej podstawie podejmować decyzje kliniczne. Ważne jest, aby modele były transparentne i możliwe do interpretacji, co pozwala na zaufanie do wyników oraz umożliwia lekarzom dostosowanie i wdrażanie modeli w praktyce klinicznej.

W mojej pracy stawiam sobie za cel budowanie modeli, które nie tylko są naukowo poprawne i precyzyjne, ale także spełniają wymagania praktyczne i mogą być skutecznie wdrożone w realnym środowisku medycznym, przyczyniając się do walki z chorobami i poprawy opieki zdrowotnej.

Rozdział	Autorzy	Publikacja	Czasopismo	Rok
3.1	**Kobylińska, K**; Mikołajczyk, T; Adamek, M; Orłowski, T; Biecek, P	Explainable machine learning for modeling of early postoperative mortality in lung cancer [68]	Lecture Notes in Computer Science	2019
3.2	**Kobylińska, K**; Orłowski, T; Adamek, M; Biecek, P	Explainable Machine Learning for Lung Cancer Screening Models [69]	Applied Sciences	2022
3.3	Lemańska-Perek, A; Krzyżanowska-Gołąb, D; **Kobylińska, K**; Biecek, P; Skalec, T; Tyszko, M; Goździk, W; Adamik, B	Explainable Artificial Intelligence Helps in Understanding the Effect of Fibronectin on Survival of Sepsis [75]	Cells	2022
4	**Kobylińska, K**; Krzyziński, M; Machowicz, R; Adamek, M; Biecek, P	Exploration of the Rashomon Set Assists Trustworthy Explanations for Medical Data [70]	IEEE Journal of Biomedical and Health Informatics (w trakcie recenzji)	2023

Tabela 1.1: Lista publikacji, na podstawie których oparta jest niniejsza rozprawa.

Autorzy	Publikacja	Czasopismo	Rok
Berus, T; Markiewicz, A; **Kobylińska, K**; Biecek, P; et al.	Downregulation of Polo-like kinase-1 (PLK-1) expression is associated with poor clinical outcome in uveal melanoma patients	Folia Histochemica et Cytobiologica	2020
Drozd-Sokolowska, J; Zaucha, J M; Biecek, P; Giza, A; **Kobylińska, K**; et al.	Type 2 diabetes mellitus compromises the survival of diffuse large B-cell lymphoma patients treated with (R)-CHOP – the PLRG report.	Scientific Reports	2020
Waszczuk-Gajda, A; Małyszko, J; Vesole, DH; Feliksbrot-Bratosiewicz, M; Skwierawska, K; Krzanowska, K; **Kobylińska, K**; Biecek, P; et al.	Negative Impact of Borderline Creatinine Concentration and Glomerular Filtration Rate at Baseline on the Outcome of Patients With Multiple Myeloma Treated With Autologous Stem Cell Transplant	Transplantation Proceedings	2020
Skalec; T; Adamik, B; **Kobylińska, K**; Gozdzik, W	Soluble Urokinase-Type Plasminogen Activator Receptor Levels as a Predictor of Kidney Replacement Therapy in Septic Patients with Acute Kidney Injury: An Observational Study	Journal of Clinical Medcine	2022
Tyszko, M; Lipińska-Gediga, M; Lemańska-Perek, A; **Kobylińska, K**; Gozdzik, W; Adamik, B	Intestinal Fatty Acid Binding Protein (I-FABP) as a Prognostic Marker in Critically Ill COVID-19 Patients	Pathogens	2022
Drozd-Sokołowska, J; Mądry, K; Barankiewicz, J; **Kobylińska, K**; Biecek, P; et al.	SARS-CoV-2 Infection in Patients Treated with Azacitidine and Venetoclax for Acute Leukemia: A Report of a Case Series Treated in a Single Institution	Chemotherapy	2023

Tabela 1.2: Lista wybranych publikacji w ramach współpracy naukowej.

1.5. Struktura rozprawy

Rozprawa składa się z pięciu rozdziałów. W rozdziale 1 przedstawione jest wprowadzenie do tematu rozprawy. Rozdział 2 przedstawia metodologię, która jest wykorzystywana w dalszej części pracy. Rozdział 3 skupia się na badaniach, które miały na celu określenie możliwości istniejących metod wyjaśnialnego uczenia maszynowego (XAI) w zastosowaniach medycznych. W sekcji 3.1 opracowana została hipoteza badawcza 1, w sekcji 3.2 opracowana została hipoteza badawcza 2, natomiast w sekcji 3.3 opracowana została hipoteza badawcza 3. Rozdział 4 przedstawia nową metodę porównywania modeli predykcyjnych oraz nowy algorytm procesu budowania modeli. W tym rozdziale opracowana została hipoteza 4. W rozdziale 5 znajduje się podsumowanie wyników mojej pracy naukowej.

W rozprawie pisanej w języku polskim, zastosowałam konwencję tłumaczenia oryginalnych nazw metodologii i algorytmów z języka angielskiego na polski. Większość tłumaczeń jest wzorowana na pracach zbiorowych pod redakcją Mieczysława Muraszkiewicza i Roberta Nowaka [88, 46]. W niniejszej rozprawie, wszystkie nazwy pochodzące z medycznych baz danych pozostały w języku angielskim. To podejście zapewnia jednolitość w obrębie bazy danych i wykresów. W celu ułatwienia identyfikacji i zrozumienia zmiennych, w tabelach opisujących dane wprowadzone są ich tłumaczenia. Do rozprawy dołączyłam również słownik terminów.

1.6. Podziękowania

Badania przedstawione w tej rozprawie zostały wsparte finanswo przez Narodowe Centrum Badań i Rozwoju:

Grant POIR.01.01.01-00-0328/17,

Opus Grant 2016/21/B/ST6/02176.

1.7. Oświadczenie o wkładzie autora

Treść koncepcyjna rozprawy składa się z publikacji, które nie uległy znaczącym zmianom. Publikacje są wynikiem wspólnych badań, jednak we wszystkich autor odegrał aktywną rolę. Poniżej znajdują się deklaracje dotyczące roli i aktywności autora w projektach badawczych będących podstawą poszczególnych publikacji:

- *Explainable machine learning for modeling of early postoperative mortality in lung cancer* (rozdział 3.1) Byłam odpowiedzialna za analizę i interpretację wyników modelowania. Przygotowałam artykuł naukowy, w którym przedstawione zostały metody, wyniki i wnioski z badania. Wyniki badań zaprezentowałam podczas konferencji.

- *Explainable Machine Learning for Lung Cancer Screening Models* (rozdział 3.2) Byłam odpowiedzialna za analizę i opracowanie wyników porównujących modele przesiewowe. Przygotowałam artykuł naukowy.

- *Explainable Artificial Intelligence Helps in Understanding the Effect of Fibronectin on Survival of Sepsis* (rozdział 3.3) Regularnie prowadziłam rozmowy z klinicystami, aby zapewnić praktyczne uwzględnienie i ocenę naszych badań. Byłam odpowiedzialna za zdefiniowanie problemu badawczego oraz za zaproponowanie metodyki badania, analizę i opracowanie wyników. Zaprojektowałam i opracowałam aplikację, która implementuje wyniki naszego badania, umożliwiając ich praktyczne zastosowanie.

- *Exploration of the Rashomon Set Assists Trustworthy Explanations for Medical Data* (rozdział 4) Byłam odpowiedzialna za zaprojektowanie i opracowanie algorytmu, który jest głównym elementem tego badania. Opracowałam także miarę PDI, która służy do porównania profili. Uczestniczyłam w dyskusjach zespołowych, które dotyczyły identyfikacji i analizy problemu badawczego. Opracowałam modele na zbiorach medycznych służących do potwierdzenia empirycznej metodologii. Byłam odpowiedzialna za modelowanie danych dotyczących zbioru medycznego HLH. Omawiałam wyniki z klinicystami, co zapewniało ich praktyczną ocenę i uwzględnienie w dalszych etapach badania. Uczestniczyłam w pisaniu artykułu.

Rozdział 2

Metody i algorytmy

2.1. Metody i algorytmy uczenia maszynowego

Uczenie maszynowe jest to obszar sztucznej inteligencji opierający się na rozwoju algorytmów, które poprzez identyfikację wzorców w danych umożliwiają wykonywanie określonych zadań (ang. *tasks*). Można wyróżnić uczenie nadzorowane (ang. *supervised learning*), którego głównym celem jest nauczenie się wzorców na podstawie atrybutów w danych przy znajomości atrybutu docelowego (ang. *target attribute*). Takie dane nazywamy danymi etykietowanymi, które zawierają odpowiedzi, czyli pożądane wyjścia systemu uczącego się. Drugą kategorią uczenia maszynowego jest uczenie nienadzorowane (ang. *unsupervised learning*), gdzie algorytmy mają za zadnie odkrywać wzorce w zbiorze danych bez znajomości pożądanych wyjść systemu uczącego się.

W rozprawie będę skupiać się na modelach predykcyjnych z obszaru uczenia nadzorowanego. Modele predykcyjne mają za zadanie wyznaczyć atrybut docelowy bazując na wartościach innych atrybutów, czyli wektorów obserwowanych wartości, zarówno ciągłych jak i dyskretnych. Wartości atrybutu docelowego z kolei nie są znane dla dowolnego przykładu, a jedynie dla pewnego podzbioru. Jeśli atrybut docelowy jest ciągły to możemy go wyznaczyć przy użyciu zadania regresji (ang. *regression*), jeśli jest dyskretny to stosujemy zadanie klasyfikacji (ang. *classification*). Przewidywanie atrybutu docelowego nazywane jest predykcją (ang. *prediction*).

W ostatnim stuleciu powstało wiele algorytmów służących do modelowania predykcyjnego, począwszy od modeli klasycznych, takich jak modele liniowe, uogólnione

modele liniowe, drzewa decyzyjne, poprzez bardziej skomplikowane modele wykorzystujące techniki baggingu [12], boostingu, czy wreszcie sieci neuronowych. Modele te są dobrze opisane w wielu podręcznikach, na przykład [57].

W dalszej kolejnych rozdziałach tej rozprawy najczęściej stosowanymi modelami są las losowy oraz model oparty na mechanizmie wzmocnienia gradientowego, więc zdecydowałam się przedstawić poniżej opis tych metod. Modele te są komitetami modeli drzewiastych, dlatego rozpocznę ich omawianie od wprowadzenia modelu drzewa.

2.1.1. Drzewo

Drzewo to model dzielący dziedzinę na podzbiory na podstawie atrybutów [57, 88]. Struktura drzewa składa się z węzłów, gałęzi i liści. Węzły reprezentują podziały dziedziny, natomiast gałęzie prowadzą do odpowiednich węzłów na podstawie tych podziałów. Węzły bez gałęzi wychodzących, czyli liście, przechowują klasy lub prawdopodobieństwa klas potrzebne do predykcji [88].

Drzewo decyzyjne to model reprezentujący zadanie klasyfikacji. Budowa drzewa rozpoczyna się od korzenia (pierwszego węzła), który zawiera cały zbiór trenujący. Następnie podejmowana jest decyzja, czy utworzyć nowe węzły, i jaki podział w nich zastosować. Dla każdego nowego węzła stosuje się kryterium stopu. Jeśli jest ono spełnione, zapada decyzja o utworzeniu liścia, w przeciwnym przypadku tworzy się nowy węzeł. Wybór klasy dla liścia oznacza decyzję o predykcji, natomiast wybór klasy dla węzła dotyczy podziału zastosowanego w danym węźle. Podział (ang. *decision tree split*) opiera się na wartościach atrybutów opisujących przykłady i można go zdefiniować funkcją:

$$j : X \to R_j. \tag{2.1}$$

Każdemu przykładowi przypisuje się jeden z wyników ze zbioru R_j, a każdemu wynikowi $r \in R_j$ odpowiada jedna gałąź prowadząca do kolejnego węzła. Typ podziału zależy od typu atrybutów. Dla atrybutów dyskretnych można zastosować podziały wielowartościowe i binarne, natomiast dla atrybutów ciągłych stosuje się podział na podstawie przedziałów wartości atrybutu. Wybór podziału podporządkowany jest kryterium stopu, które dotyczy uzyskania bardziej jednorodnej klasy.

Złożoność drzewa mierzona jest jego głębokością. Im bardziej drzewo jest złożone,

tym lepiej jest dopasowane do danych. Aby zredukować ryzyko nadmiernego dopasowania, można tworzyć prostsze drzewa, które podejmują mniej decyzji. Zastosowanie kryterium wyboru podziału preferujące podziały z najmniejszą niejednorodnością klas pozwala na stworzenie takiego drzewa. Niejednorodność klas można zmierzyć miarami takimi jak entropia czy współczynnik Giniego [88].

Drzewa regresyjne to metoda reprezentacji zadania regresji, różniąca się od drzew decyzyjnych reprezentacją liści i ich interpretacją. W przypadku drzew regresyjnych, w liściach znajdują się predykcje wartości funkcji docelowej w postaci liczbowych wartości. Węzły i gałęzie są reprezentowane w ten sam sposób co w drzewach decyzyjnych, jednak kryterium stopu i podział ulegają modyfikacji. Kryterium stopu opiera się na minimalizacji błędu, najczęściej stosując minimalizację błędu średniokwadratowego na zbiorze trenującym (ang. *root mean squared error*, RMSE). Przy wyborze podziału preferowane są podzbiory, które najbardziej redukują rozrzut wartości funkcji docelowej, np. odchylenia standardowego.

Drzewa regresji mają pewne ograniczenia. Reprezentowane przez nie modele są przedziałami stałe, więc dla określonych regionów dziedziny wartość predykcji jest stała. Zmiana wartości atrybutów może nie spowodować zmiany liścia, co oznacza, że predykcja pozostaje niezmieniona. Jeśli zmiana atrybutów powoduje zmianę liścia, następuje skokowa zmiana predykcji. Aby ograniczyć skokowe zmiany, można wytrenować drzewa regresyjne, których liście odpowiadają wąskim regionom dziedziny. Może to jednak spowodować nadmierne dopasowanie do danych.

Formalnie można wyrazić drzewo następująco [57]:

$$T(x, \theta) = \sum_{j=1}^{J} \gamma \mathbb{1}(x_j), \tag{2.2}$$

z parametrami $\theta = \{R_j, \gamma_j\}_1^J$, gdzie γ_j to wielkość predykcji w regionie R_j.

Zaletą drzew decyzyjnych jest łatwość interpretacji wyników, jednak często odbywa się to kosztem jakości modelu. Drzewa decyzyjne wykorzystano jako komponenty do budowy komitetów modeli drzewiastych, które zostały opisane w dalszej części tej pracy.

2.1.2. Las losowy

Las losowy (ang. *random forest*) [14, 57] jest modelem zespołowym [88]. Model zespołowy łączy wiele różnych modeli bazowych dla tego samego zadania klasyfikacji lub regresji w celu uzyskania jednego, lepszego pod względem jakości modelu.

Konstrukcja lasu losowego [57] opiera się o modyfikację idei baggingu [12] tworząc duży komitet zróżnicowanych drzew a następnie je uśredniając. Przyjmijmy, że model zbudowany jest z M drzew. Pojedyncze drzewa mają ten sam rozkład stąd obciążenie średniej tych modeli, będzie równe obciążeniu pojedynczego modelu. Zatem możliwym ulepszeniem modelu zespołowego jest redukcja wariancji. Jeśli modele bazowe są ze sobą skorelowane o współczynnik ρ, a wariancja pojedynczego modelu wynosi σ^2 to wariancja średniej M modeli wynosi $\rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$. Aby zminimalizować to wyrażenie należy zminimalizować jego pierwszy człon, gdyż dla dużych M drugi człon wyrażenia znika. Stąd idea konstrukcji lasu losowego opiera się na dopasowaniu dużej liczby nisko skorelowanych modeli drzew tym samym zmniejszając wariancję docelowego modelu.

Ponadto bardziej różnorodne modele bazowe pomagają też zredukować ryzyko przeuczenia zapewniając różnorodność w predykcjach. W lesie losowym zastosowano następujące techniki mające na celu zbudowanie różnorodnych modeli zmniejszając tym samym wariancję modelu. Po pierwsze, modelami bazowymi są drzewa decyzyjne trenowane na zróżnicowanych zbiorach, tzw. próbach bootsrapowych (ang.*bootstrap sample*). Są to próby losowane z pierwotnego zbioru trenującego ze zwracaniem z rozkładem jednostajnym. Takie losowanie powoduje, że dla wystarczająco dużych zbiorów danych ok. 36% przykładów nie znajdzie się ani razu w pojedynczej próbie bootstrapowej. Zbiór przykładów trenujących, które były niewylosowane do próby bootstrapowej nazywa się przykładami OOB (ang. *out-of-bag-instances*). Mogą posłużyć do wewnętrznej estymacji błędu predykcji. Błąd OOB (ang. *OOB error*) jest estymowany na podstawie takich podzbiorów i jest dobrym estymatorem błędu rzeczywistego uzyskiwanym na etapie budowania modelu.

Po drugie, w algorytmie lasu losowego wprowadzony jest czynnik losowy w miejscu wyboru podziału. Dla każdego węzła wybierany jest losowy podzbiór atrybutów, a jego wielkość to parametr algorytmu. Podczas budowania drzew ustala się kryterium stopu, które preferuje głębokie drzewo. Pojedyncze drzewo jest mocno dopasowanie do danych, co wiąże się z większą liczbą podziałów zapewniając ich większe zróżnicowanie. Opisane

techniki zapewniają, że modele bazowe powinny mieć niską korelację co przyczynia się do osiągania lepszej skuteczności przez model zespołowy.

Predykcja lasu losowego dla pojedynczej obserwacji jest obliczana na podstawie predykcji wszystkich drzew w lesie. Są one agregowane przy użyciu głosowania dla zadania klasyfikacji lub uśredniania dla zadania regresji. Modelowanie przy pomocy lasu losowego często osiąga bardzo dobre wyniki pod względem jakości predykcji, choć odbywa się to kosztem interpretowalności.

2.1.3. Wzmocnienie gradientowe

Wzmocnienie gradientowe (ang. *gradient boosting*) [44, 20] jest modelem zespołowym wykorzystującym modele bazowe, przeważnie drzewa w celu poprawy jakości predykcji. Wykorzystuje metodę wzmocnienia (ang. *boosting*) czyli poprawy błędu w każdym kolejnym kroku. Działa na zasadzie iteracyjnego dodawania kolejnych słabych modeli do modelu zespołowego w celu minimalizacji funkcji straty. Opiera się na koncepcji gradientu, który wskazuje kierunek najszybszego wzrostu funkcji straty. Budowa kolejnego drzewa ma na celu redukcję błędów poprzedniego drzewa.

Model zespołowy drzewiasty można zapisać jako sumę drzew:

$$f_M(x) = \sum_{m=1}^{M} T(x, \theta_m), \tag{2.3}$$

gdzie $\theta_m = \{R_j, \gamma_j\}_1^J$ to parametr dla m-tego drzewa.

Niech L będzie funkcją straty. W każdym kroku algorytmu należy rozwiązać:

$$\hat{\theta}_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m)) \tag{2.4}$$

przez minimalizację:

$$\hat{\gamma}_{jm} = \arg\min \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_{jm}). \tag{2.5}$$

W algorytmie opisanym przez Grega Ridgewaya [98], podczas m-tej iteracji rozwiązaniem jest drzewo $T(x; \theta_m)$, którego predykcje są najbliższe ujemnemu gradientowi.

Algorytm rozpoczyna się od znalezienia początkowego modelu $f_0(x)$, który minimalizuje funkcję straty. Następnie w kolejnych krokach od 1 do M, liczone są ujemne gradienty funkcji straty:

$$r_{im} = -\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} | f(x) = f_{m-1}(x), \quad dla \;\; i = 1, 2, \ldots, n. \tag{2.6}$$

Dopasowywane są drzewa, których atrybutem docelowym jest wyliczony w ten sposób gradient. Model zespołowy jest aktualizowany o model zawierający wyliczone drzewo w następujący sposób:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} 1(x \in R_{jm}), \tag{2.7}$$

gdzie j to liczba liści w drzewie, a γ_{jm} są parametrami drzewa optymalizowanymi dla każdego z regionów drzewa. Wynikiem całej procedury jest model zespołowy $f_M(x)$.

Dla zadania regresji popularnymi funkcjami straty są błąd średniokwadratowy, czy miara Hubera, a dla zadania klasyfikacji dewiancja [88].

Hiperparametrami tego modelu są liczba iteracji M, czyli liczba modeli bazowych. Procedura może być również wzbogacona o współczynnik uczenia się η (ang. *learning rate*). W każdym kolejnym kroku, algorytm stara się poprawić predykcję poprzedniego modelu poprzez dopasowanie nowego modelu do gradientu. Reguła aktualizacji w algorytmie ma następującą postać:

$$f_m(x) = f_{m-1}(x) + \eta \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{1}(x \in R_{jm}), \tag{2.8}$$

gdzie $0 < \eta \leq 1$.

Pojedyncze drzewo jest dobrze interpretowalne, zatem cały model można przedstawić przy pomocy grafu prostego. Niestety model zespołowy, który jest liniową kombinacją wielu drzew traci tą własność. Zarówno w przypadku modelu lasu losowego jak i techniki wzmocnienia gradientowego należy zastosować inne metody interpretowalności.

2.2. Interpretowalne uczenie maszynowe

Wiele modeli sztucznej inteligencji jest bardzo złożonych ze względu na ilość parametrów czy zaawansowane formuły matematyczne. Zrozumienie jakie cechy mają wpływ na decyzję modelu, czy w jaki sposób działa model staje się trudne, a czasami niemożliwe. Tim Miller [86] zaproponował definicję interpretowlności jako stopień, w jakim człowiek może zrozumieć przyczynę decyzji. Proste modele, takie jak regresja liniowa czy drzewa decyzyjne są zazwyczaj bardziej interpretowalne, niestety kosztem dokładności w porównaniu do bardziej zaawansowanych modeli.

Istotnym aspektem modeli uczenia maszynowego jest ich transferowalność [79]. Zdarza się, że model dokonuje prawidłowych predykcji na podstawie szumu zawartego w danych treningowych, a nie na podstawie prawdziwego sygnału. Przeniesienie takiego modelu na nowe dane może skutkować błędnymi predykcjami. Przykładem takiego modelu jest sieć neuronowa, która miała rozróżnić zdjęcia wilków od psów husky i została zaprezentowana przez Marco Ribeiro w artykule dotyczącym metody LIME [97]. Jakość liczona funkcją straty na zbiorze testowym była wysoka. Natomiast model ten nie był transferowalny na inne zbiory danych. Zastosowanie metod XAI pomogło zrozumieć, że model rozpoznawał psy na podstawie otoczenia. Jeśli zdjęcie miało śnieżne tło model rozpoznawał psa, a w przeciwnym przypadku wilka. Przykład ten pokazuje, że zrozumienie, które zmienne spośród wielu analizowanych wpływają najmocniej na predykcję jest potrzebne przy ocenie jakości modelu. Metody XAI mogą pomóc w sprawdzeniu, które cechy danych model uznał za istotne przyczyniając się do wychwycenia błędu modelu czy jego poprawy. Oczekuje się, że metody wyjaśnialnej sztucznej inteligencji pomogą zbudować zaufanie do modelu przez osoby, które mają z niego korzystać [73]. Przedstawienie wyjaśnienia modelu może przyczynić się do zwiększenia zaufania do predykcji poprzez zrozumienie przyczyny danej decyzji. Kolejnym aspektem wyjaśnialności modeli jest ich informatywność [84]. Metody XAI mogą wspomóc wyjaśnienie i wnieść dodatkowe informacje dotyczące predykcji. Szczególnie istotne jest to w przypadku decyzji we wrażliwych dziedzinach, np. w medycynie. Istotna jest także kwestia stronniczości modeli predykcyjnych (ang. *fairness*) [85]. Metody XAI mogą wspomóc sztuczną inteligencję w aspektach etycznych [128]. Brak wglądu w modele czarnych skrzynek wiąże się z możliwością zachowania przez modele dyskryminująco ze względu na rasę czy płeć. Przykładem modelu, który dyskryminował ze względu na płeć jest

system rekrutujący firmy Amazon [6]. Model miał służyć do analizy zgłoszeń kandydatów do pracy. Okazało się, że faworyzował mężczyzn wyszukując w dokumentach słowa, które wskazywały, że pisała je kobieta. Firma szybko wycofała się z tego systemu.

Potrzeba stosowania interpretowalnych modeli wiąże się z problemem, który analizujemy. Im bardziej decyzje modelu wpływają na ludzkie życie tym większa potrzeba dokładnego, ale też interpretowalnego modelu. Dobrym przykładem są zastosowania w naukach medycznych, gdzie zarówno wysoka precyzja jak i interpretowalność są istotne. Interpretowalność w tym przypadku wpływa na zaufanie i akceptację społeczną. Finale Doshi-Velez i Been Kim [29] wskazują, że problem interpretowalności wynika z niedoprecyzowania w formalizacji problemu badawczego. Istotne są odpowiedzi nie tylko na pytanie "jaka"jest predykcja, ale także "dlaczego"jest taka predykcja. W wielu przypadkach analiza "dlaczego"model podejmuje taką decyzję może doprowadzić do wyciągnięcia wniosków o badanym problemie, ale także do znalezienia słabych punktów modelowania.

2.2.1. Taksonomia

Wyjaśnialna sztuczna inteligencja stanowi zbiór metod i modeli mających wspomóc zrozumienie działania modeli sztucznej inteligencji. Metody te mają pomóc w interpretowaniu wyników nawet najbardziej skomplikowanych modeli. Istnieją różne klasyfikacje metod XAI [33, 145, 54].

Na Rysunku 2.1 znajduje się ilustracja podziału metod służących do eksploracji modeli predykcyjnych. Można podzielić na te dotyczące modeli z intepretowalną strukturą modelu i przybliżone wyjaśnienia modelu. Interpetowalną strukturę mają na przykład modele liniowe, uogólnione modele addytywne czy drzewa. Są to modele, dla których można opisać wyjaśnienia dokładne dzięki strukturze takiego modelu. Dla modeli liniowych istotność atrybutów i ich wpływ na predykcję można oprzeć o analizę współczynników modelu. Dla drzew decyzyjnych wyjaśnienie predykcji to reguła, która prowadzi do decyzji modelu. Ponadto, drzewa można prezentować w formie graficznej, co także ułatwia ich interpretowalność. Warto jednak zauważyć że bardzo głębokie drzewo lub model liniowy z ogromną liczbą zmiennych mogą stać się słabiej wyjaśnialne. Przybliżone wyjaśnienia modelu z kolei można podzielić na wyjaśnienia specyficzne dla rodziny modeli (ang. *model specific*) oraz niezależne od rodziny modeli (ang. *model agnostic*).

Wyjaśnienia specyficzne dla rodziny modeli są to metody zaprojektowane w oparciu o budowę danej rodziny modeli. Na przykład, dla sieci neuronowych, opartych o strukturę warstwową zaproponowano metodę propagowania istotności przez warstwy (ang. *layer-wise relevance propagation, LRP*) [11] czy metody typu mapowanie ścieżek aktywacji klas (ang. *class activation mapping, CAM*) [141]. Dla modeli zespołowych można zastosować metodę wyliczenia wartości Shapley w czasie wielomianowym (TreeShap [89]). Metody niezależne od rodziny modeli z kolei można zastosować do dowolnej struktury.

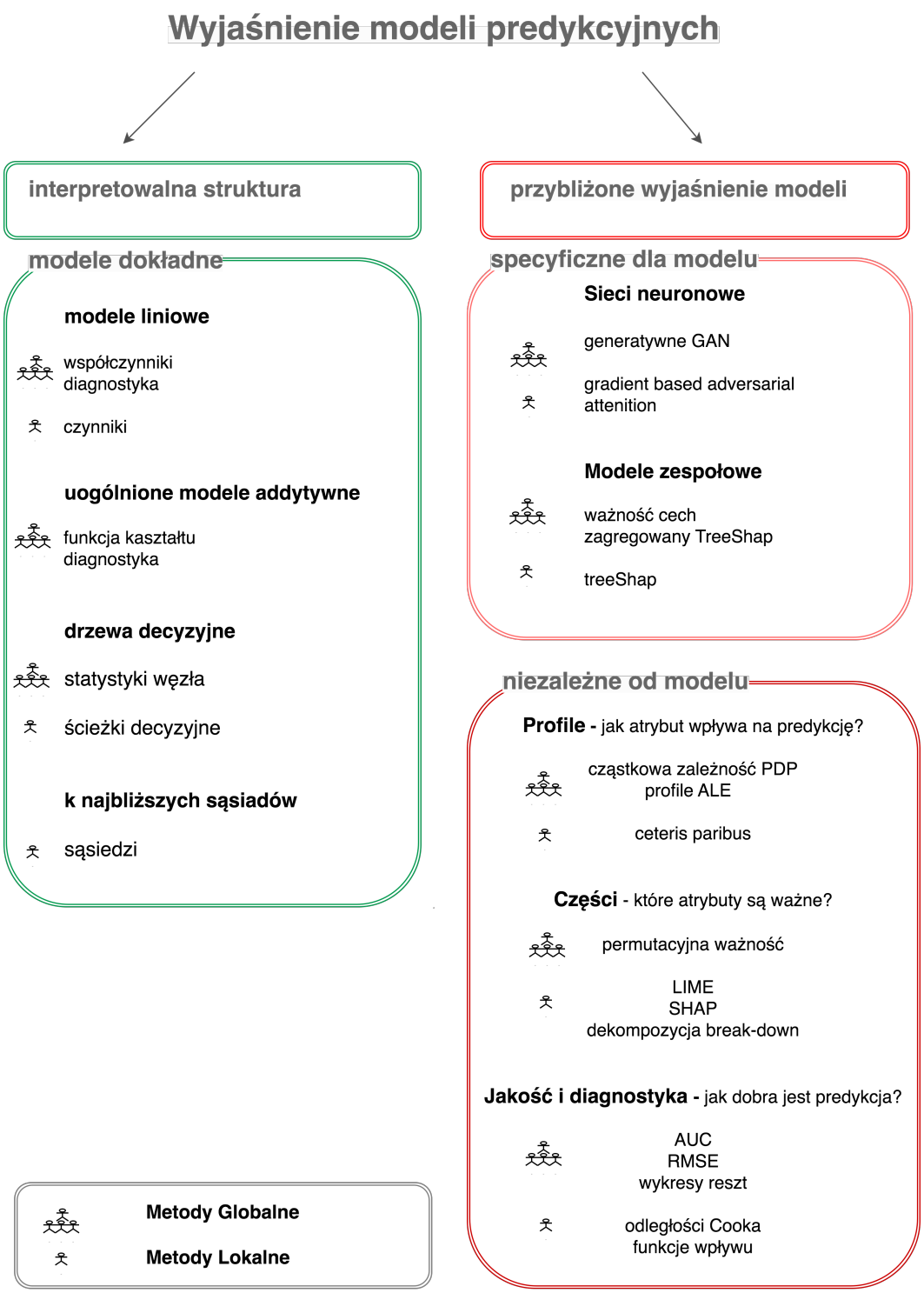
W niniejszej rozprawie wykorzystuję metody niezależne od rodziny modeli, które są powszechnie stosowane w analizie modeli uczenia maszynowego. Zaletą tych metod jest możliwość porównania wyjaśnień dla różnych klas modeli.

Metody wyjaśniania modeli predykcyjnych można podzielić też na dwie kategorie: metody wyjaśniające działanie całego modelu (globalne) i metody wyjaśniające predykcję dla pojedynczej obserwacji (lokalne). Metody lokalne stosuje się w celu analizy predykcji dla pojedynczej jednostki. Pomagają zrozumieć i uzasadnić pojedynczą modelu. Globalne metody z kolei dotyczą średniego zachowania modelu dla całego zbioru danych lub podzbioru danych. Ten podział posłuży mi do klasyfikacji metod XAI w dalszych częściach rozprawy.

2.3. Globalne wyjaśnienia modeli predykcyjnych

Globalne metody wyjaśnień służą do zrozumienia średniego zachowania modelu [46]. Metody te można wykorzystać na etapie trenowania i walidacji modelu. Pozwalają na wybór najlepszego modelu spośród wielu dobrych, dostrzeżenie różnic i znalezienie ewentualnych błędów. Mogą służyć do znalezienia atrybutów najbardziej wpływających na badane zjawisko, czy poznać średni profil odpowiedzi.

W celu formalnej definicji metod służących do wyjaśniania modeli predykcyjnych niech $\mathbf{M} = [\mathbf{X}\ \mathbf{Y}]$ będzie macierzą składającą się z n obserwacji i p niezależnych zmiennych, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$, gdzie $(\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}] \in \mathcal{X})$, oraz zmiennej zależnej w postaci wektora: $\mathbf{Y} \in \mathcal{Y}$. Przestrzeń modeli predykcyjnych f jest postaci: $\mathcal{F} = \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$, gdzie f to modele predykcyjne. Niech przestrzeń $\mathcal{L} : (\mathcal{F} \times \mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ zawiera funkcje straty L użyte do ewaluacji jakości modeli predykcyjnych.



Rysunek 2.1: Taksonomia metody służących do wyjaśnienia modeli predykcyjnych z przykładowymi metodami.

2.3.1. Metoda cząstkowej zależności

Metoda cząstkowej zależności (ang. *partial dependence plot, PDP*) wyznacza średni profil odpowiedzi, czyli relację pomiędzy wartościami j-tej zmiennej a wartościami predykcji modelu. Po raz pierwszy zaproponowana została dla metody wzmocnienia gradientowego przez Jerome Friedmana [44]. Później została zastosowana dla innych klas modeli. Pokazuje jak zmienia się odpowiedź modelu w ramach zmiany pojedynczej zmiennej. Przedstawia krańcowy wpływ zmiennej na przewidywane wartości [87]. Jest to wartość oczekiwana predykcji modelu dla ustalonej wartości j-tej zmiennej w punkcie z przy rozkładzie łącznym $\mathcal{X}^{(-j)}$. Metodę cząstkowej zależności można zapisać jako:

$$PD^{j}_{f,PDP}(z) = \mathbb{E}_{\mathcal{X}^{(-j)}}[f(X^{(1)}, \ldots, X^{(j-1)}, z, X^{(j+1)}, \ldots, X^{(p)})]. \tag{2.9}$$

Wykorzystując empiryczny rozkład danych, estymatorem jest następująca średnia:

$$\widehat{PD^{j}_{f,PDP}}(z) = \frac{1}{n}\sum_{i=1}^{n} f(x_i^{(1)}, \ldots, x_i^{(j-1)}, z, x_i^{(j+1)}, \ldots, x_i^{(p)}). \tag{2.10}$$

Porównanie krzywych PDP dla kilku modeli może potwierdzić czy modele nauczyły się dobrych zależności, czy wręcz przeciwnie popełniają błędy.

2.3.2. Permutacyjna ważność zmiennych

Permutacyjna ważność zmiennych (ang. *permutation feature importance, variable importance*) to metoda oszacowania istotności zmiennych na podstawie wzrostu błędu predykcji po spermutowaniu wybranej zmiennej. Zmienna uznana jest za ważną, gdy błąd modelu rośnie po usunięciu efektu tej zmiennej. Metoda ta wprowadzona przez Leo Breimana [14] dla lasów losowych, została później rozszerzona o wersję, którą można użyć dla dowolnego modelu [41, 40].

Permutacyjna ważność zmiennej VI dla zmiennej j to różnica między funkcją straty dla modelu zawierającego wszystkie zmienne oraz dla modelu z wyłączonym efektem analizowanej zmiennej:

$$VI(j) = L(f, X, y) - L(f, X^{j}, y), \tag{2.11}$$

gdzie X^j to dane ze spermutowaną j-tą zmienną.

Powstało wiele modyfikacji tej metody. Oszacowanie błędu modelu można także otrzymać na podstawie profilu PDP. Wtedy ważność zmiennej to średnia odległość pomiędzy PDP dla danej zmiennej a średnią odpowiedzią modelu.

Niech $VI_{PD}(f, j)$ będzie ważnością zmiennej na podstawie cząstkowej zależności PDP dla funkcji f i zmiennej j. Wtedy:

$$VI_{PD}(f, j) = \frac{1}{n} \sum_{i=1}^{n} \mid PD(f, j, z_i) - \overline{PD}(f, j) \mid, \tag{2.12}$$

gdzie:
$\overline{PD}(f, j)$ jest średnią wartością PDP policzoną po wszystkich zmiennych.

Metoda wyliczenia ważności cech jest bardzo przydatna w analizach. W zastosowaniach medycznych może służyć do walidacji modeli z wiedzą domenową. Może być także wykorzystana do znalezienia nowych, istotnych cech które są odpowiedzialne za dany mechanizm biologiczny.

2.4. Lokalne wyjaśnienia modeli predykcyjnych

Lokalne wyjaśnienia to takie, które dotyczą analizy predykcji dla pojedynczej jednostki [46]. W przypadku zastosowania do zbiorów medycznych, takie spojrzenie na lokalne zachowanie modelu jest niezwykle ważne. Pomaga w interpretacji wyniku predykcji dla konkretnej osoby wspomagając spersonalizowaną medycynę. Lokalne wyjaśnienia mogą pomóc w budowaniu zaufania do predykcji zarówno przez klinicystów jak i osób których predykcja dotyczy. W moich badaniach skupiłam się na metodach lokalnych, które mogłam zastosować do każdej rodziny modeli i porównać wyniki pomiędzy różnymi modelami.

2.4.1. SHapley Additive eXplanations

Metoda SHAP (ang. *Shapley additive explanations*) jest metodą wyjaśnień złożonych modeli. Opiera się o analizę zmiennych, które odpowiadają za przesunięcie predykcji od średniej predykcji modelu dla pojedynczej jednostki. Jest to metoda przeniesiona z teorii gier kooperacyjnych dotycząca sposobu podziału zysku pomiędzy graczy za-

proponowanej przez Lloyda Shapleya [106]. Pierwszy raz idea ta wdrożona została do uczenia maszynowego w pracach [125, 130]. W 2017 roku, Lundberg zaproponował [80] sposób wyznaczania wartości Shapleya w sposób efektywny.

Metoda SHAP przedstawia dekompozycję predykcji modelu na wkład wszystkich zmiennych dla danej jednostki. Można interpretować wynik metody SHAP jako wkład w predykcję poszczególnych zmiennych. Wartości SHAP mierzą każdej zmiennej jej średni, oczekiwany wkład w wynik modelu. Predykcję modelu f w wybranym punkcie x^* można wyrazić jako:

$$f(x^*) = baseline + \sum_{j=1}^{p} v_j(f, x^*), \tag{2.13}$$

gdzie $v_j(f, x^*)$ to wkład dla obserwacji x^* i dla j-tej zmiennej.

Warto zauważyć, że gdy model jest nieliniowy lub cechy wejściowe nie są niezależne, kolejność dodawania cech ma znaczenie, a wartości SHAP wynikają z uśrednienia wartości v we wszystkich możliwych uporządkowaniach.

Niech S oznacza podzbiór zmiennych, bez zmiennej j. W celu obliczenia wkładu dowolnej cechy j, obliczamy predykcję modelu, używając wszystkich cech w podzbiorze S i odejmujemy ją od wartości predykcji podzbioru, w którym ta cecha jest nadal obecna. Wielkość tą ważymy całkowitą liczbą permutacji zmiennych. Następnie sumujemy po wszystkich możliwych różnicach predykcji modelu, uzyskując średni wkład cechy do predykcji wyszkolonego modelu:

$$v_j(f, x^*) = \sum_{S \subseteq \{1,2,\ldots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)}{p!} (f_{S \cup j}(x^{(S \cup j)})) - f(S)). \tag{2.14}$$

Metoda ta rozkłada predykcję na czynniki, co pozwala na zrozumienie, które zmienne i w jaki sposób wpływają na wynik dla pojedynczej obserwacji. Wartości Shapleya mają przydatne własności, m.in. addytywność i symetrię. Suma dla wszystkich zmiennych wartości Shapleya daje predykcję modelu. Własność symetrii oznacza, że jeśli dwa atrybuty mają taki sam wpływ na predykcję to ich wartości Shapleya są sobie równe. Ponadto, wartość Shapleya wynosi 0 dla atrybutu, który nie wpływa na predykcję modelu.

2.4.2. Dekompozycja modelu break-down

Modyfikacją metody SHAP jest metoda dekompozycji predykcji break-down [53, 10]. Dekompozycję tą można zapisać również zgodnie z równaniem 2.13. Metody liczenia wkładu w tych dwóch metodach różnią się. Intuicyjnie, w metodzie SHAP liczona jest uśredniona wartość wkładu po wszystkich uporządkowaniach zmiennych. Metoda dekompozycji break-down z kolei zależy od tylko jednego uporządkowania zmiennych, a szczegółowy algorytm znajduje się w pracy [115].

Dla wektora losowego $\hat{X}$, wzór 2.13 można wyrazić następująco:

$$E[f(\hat{X})|X_1 = x_1^*, \ldots, X_p = x_p^*] = E[f(\hat{X})] + \sum_{j=1}^{p} v_j(f, x^*). \tag{2.15}$$

Wtedy wkład zmiennej j można wyznaczyć wzorem:

$$v_j(f, x^*) = E[f(\hat{X})|X_1 = x_1^*, \ldots, X_j = x_j^*] - E[f(\hat{X})|X_1 = x_1^*, \ldots, X_{j-1} = x_{j-1}^*]. \tag{2.16}$$

2.4.3. Lokalny profil

Lokalny profil ceteris paribus [51, 132, 8, 50] służy do pokazania lokalnej odpowiedzi modelu gdy jedną analizowaną cechę zmienimy, a resztę pozostawiamy bez zmian. Jest to szczególny przypadek profilu PDP, jednak w tym przypadku liczymy odpowiedzi modelu tylko dla jednej obserwacji. Pozwala na zrozumienie w jaki sposób wartości zmiennej wpływają na predykcję modelu dla wybranej obserwacji.

Dla pojedynczej zmiennej x^* i dla zmiennej j wartość profilu:

$$CP^{f,j,x^*}(z) = f(x^*|j = z) = f(x_1^*, \ldots, x_{j-1}^*, z, x_{j+1}^*, \ldots, x_p^*), \tag{2.17}$$

gdzie $f(x^*|j = z)$ oznacza wartość predykcji modelu jeśli dla zmiennej j ustalimy wartość z, a wszystkie pozostałe zmienne pozostaną niezmienione.

W naukach medycznych takie spojrzenie na predykcję pozwala przeanalizować, które zmienne i w jaki sposób pacjent powinien zmienić w celu poprawy predykcji.

2.5. Proces budowy modelu

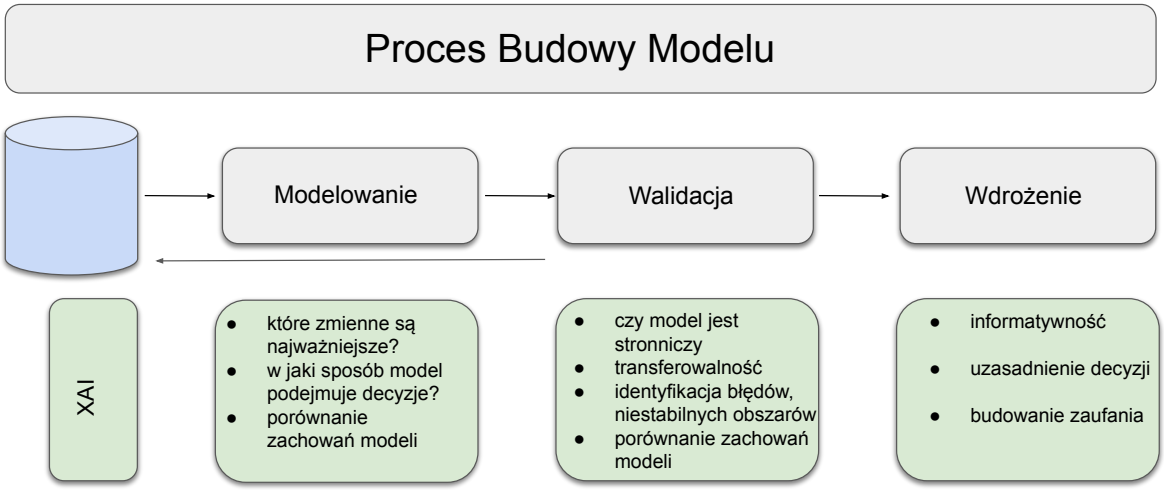
Jednym z uniwersalnych procesów służących do analizy danych jest procedura CRISP-DM (ang. *cross-industry standard process for data mining*) [133, 19]. Został opracowany jako standard postępowania w projektach dotyczących analizy danych. Proces ten składa się z następujących etapów: zrozumienie problemu, eksploracja danych, przygotowanie danych, modelowanie, ewaluacja i wdrożenie. Dzięki zastosowaniu takiego podejścia możliwe jest identyfikowanie i rozwiązywanie problemów związanych z modelem, takich jak nadmierna generalizacja, niedopasowanie danych czy wykrywanie błędów. Wykorzystanie procesu umożliwia skalowanie projektu w przyszłości, a także pozwala na powtarzanie i modyfikację procesu w zależności od potrzeb, co ułatwia wprowadzanie ulepszeń do modelu i dostosowywanie go do zmieniających się wymagań. W oparciu o procedurę CRISP-DM, Przemysław Biecek [9] zaproponował metodykę MDP (ang.*model development process*, która pokazuje jak działa proces CRISP-DM w czasie. Kolejne iteracje procesu różnią się chociażby ze względu na pogłębioną wiedzę na temat problemu.

Oczekiwanym rezultatem zastosowania tych procedur jest wybór najlepszego modelu opisującego badane zjawisko, jednak cały proces wymaga rzetelnej analizy. Metody wyjaśnialnego uczenia maszynowego mogą być zastosowane na różnych etapach takiego procesu. W rozprawie skupiam się na zastosowaniu metod XAI do następujących etapów procesu: modelowanie, walidacja i wdrożenie modelu.

Etap modelowania nazywam także etapem gromadzenia modeli, gdyż polega na tworzeniu i testowaniu modeli, z różnymi zestawami parametrów czy różnych klas modeli, w celu wybrania najlepszego, który najtrafniej opisuje badane zjawisko. Na tym etapie, wyjaśnialna sztuczna inteligencja może wspomóc wybór lepszego modelu. Istotnymi mogą okazać się odpowiedzi na pytania o czynniki które przyczyniają się do predykcji, które atrybuty są najbardziej istotne. Ponadto metody XAI na tym etapie mogą wspomóc porównanie różnych klas modeli.

Podczas etapu walidacji modelu, metody XAI mogą dostarczyć istotnych informacji o wyjaśnieniu predykcji, wspomóc zidentyfikować błędy, niestabilne obszary, czy upewnić się, że model będzie transferowalny na inne zbiory danych.

Podczas etapu wdrożenia modelu, metody XAI mogą ułatwić budowanie zaufania do modelu, wspomóc uzasadnienie decyzji oraz wnieść dodatkowe informacje do opisu



Rysunek 2.2: Proces budowy modelu. Wszystkie jego etapy można wspomóc o wyjaśnialną sztuczną inteligencję.

badanego zjawiska.

Na Rysunku 2.2 znajduje się podsumowanie procesu, który posłużył mi do sprawdzenia jak metody XAI pomagają w wyborze modelu uczenia maszynowego. Zaznaczone są przykładowe pytania badawcze, przy których pomóc może wyjaśnienie modeli uczenia maszynowego. Przy wszystkich analizach rzeczywistych zbiorów medycznych polegałam na tym procesie. Metody XAI można wykorzystać na różnych jego etapach. W sekcji 3 przedstawię empiryczne sprawdzenie w jaki sposób metody XAI mogą pomóc w przeprowadzeniu procesu budowy modelu przy analizach danych z obszaru medycyny. Na podstawie zebranego doświadczenia w sekcji 4 przedstawiam moją modyfikację tego procesu o wybór nie jednego modelu, ale całego zbioru niemal optymalnych modeli. Proponuję by analizując dane nie skupiać się tylko na wyborze jednego optymalnego modelu lecz sprawdzić szereg modeli. W tym celu przedstawię nową metodę XAI potrzebną do odpowiedniego porównania modeli.

Rozdział 3

Wyjaśnialne uczenie maszynowe w obszarze medycyny

W niniejszym rozdziale przedstawiam przykłady, które pokazują skuteczność i potencjał metod XAI w kontekście rzeczywistych problemów medycznych. Metody te pozwalają na głębsze zrozumienie procesów podejmowania decyzji przez modele sztucznej inteligencji. Wykorzystanie metod XAI nie ogranicza się tylko do analizy gotowych modeli. Te metody mogą być użyte na wszystkich etapach procesu budowania modeli predykcyjnych - począwszy od gromadzenia modeli, przez ich walidację, aż do wdrożenia modelu. Dzięki nim możliwe jest pozyskanie dodatkowej wiedzy na temat funkcjonowania modeli oraz potencjalnych błędów, co przekłada się na zwiększenie jakości i wiarygodności modelowania predykcyjnego.

W sekcji 3.1 przedstawiam przykład zastosowania metod XAI na etapie gromadzenia modeli predykcyjnych, gdzie wyjaśnialność modeli pomaga w zrozumieniu działania, a następnie wyboru lepszego bądź poprawy istniejącego modelu.

W sekcji 3.2 prezentuję przykład zastosowania metod XAI w celu porównania istniejących modeli i wyboru jednego spośród wielu. Przykład ten wskazuje, jak metody XAI mogą pomóc w walidacji modeli oraz wyborze modelu, który lepiej opisuje badane zjawisko.

W sekcji 3.3 omawiam przykład zastosowania metod XAI do wyjaśnienia wyników modelowania na etapie wdrożenia, gdzie interpretowalność modelu jest istotna dla budowy zaufania do wyników przez lekarzy.

Przedstawione przykłady ilustrują, jak metody XAI mogą być efektywnie wykorzystane przy analizie problemów medycznych, umożliwiając zrozumienie decyzji podejmowanych przez modele predykcyjne oraz dostarczając informacji i wyjaśnień dla personelu medycznego.

3.1. Metody XAI na etapie gromadzenia modeli

Rozdział ten powstał na podstawie wyników opublikowanych w publikacji [68]:

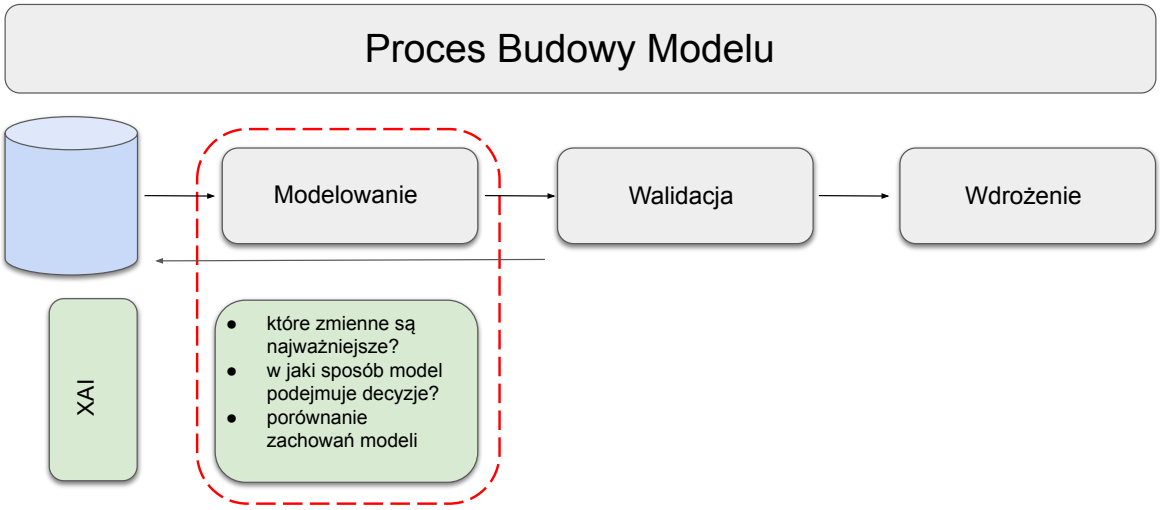
Katarzyna Kobylińska, Tomasz Mikołajczyk, Mariusz Adamek, Tadeusz Orłowski, Przemysław Biecek, "Explainable machine learning for modeling of early postoperative mortality in lung cancer"przyjęty do publikacji w materiałach Międzynarodowej Konferencji: Artificial Intelligence in Medicine, Lecture Notes in Computer Science, 2019, Volume 11979, pages 161–174..

Stanowi odpowiedź na hipotezę 1 z Sekcji 1.3, która brzmi: "Metody wyjaśnialnego uczenia maszynowego post-hoc poprawiają jakość modelowania rozumianą jako dokładność predykcji lub poprawność modelu. Zastosowanie tych metod pozwala na osiągnięcie lepszych wyników modeli uczenia maszynowego w medycynie".

3.1.1. Wprowadzenie

W niniejszym rozdziale skupiam się na omówieniu, w jaki sposób metody XAI mogą być zastosowane na etapie modelowania w trakcie procesu budowy modelu. Wyjaśnialne uczenie maszynowe stanowi narzędzie w procesie budowania modeli, odpowiadając na pytania dotyczące sposobu podejmowania decyzji przez modele oraz identyfikując potencjalne obszary do poprawy. Prezentowane metody XAI pełnią istotną rolę w procesie budowy modeli na etapie ich gromadzenia, umożliwiając badaczom odpowiedź na kluczowe pytania podczas dążenia do zbudowania optymalnego modelu, co przedstawia Rysunek 3.1.

W tym rozdziale dokonuję analizy wykorzystania metod XAI w kontekście poprawy jakości modelowania na przykładzie rzeczywistego problemu medycznego. Po pierwsze, analizuję potencjał metod XAI jako wsparcie w procesie budowy modeli w dziedzinie medycyny. Pokazuję, w jaki sposób można wykorzystać metody XAI do sprawdzenia zgodności modelowania z wiedzą domenową. Badam, czy wyjaśnienia dostarczane



Rysunek 3.1: Metody XAI użyte na etapie gromadzenia modeli w procesie budowy modelu.

przez metody XAI posiadają wartość informacyjną dla osób odpowiedzialnych za budowanie tych modeli. Czy mogą pomóc w ocenie poprawności i jakości działania model, identyfikacji istotnych zmiennych, zwiększeniu interpretowalności modeli i zapewnieniu pewności co do ich działania w kontekście medycznym. Badam w jaki sposób metody XAI mogą być wykorzystane do poprawy tradycyjnych modeli. Przedstawiam podejścia, w których metody XAI są używane jako narzędzie identyfikacji słabych stron modeli mających na celu poprawę jakości predykcji. Sprawdzam czy metody XAI mogą dostarczyć istotnych wskazówek dotyczących usprawnień tradycyjnych modeli.

3.1.2. Problem medyczny

Prezentowane badanie dotyczy przeżywalności pooperacyjnej u osób chorujących na raka płuca. Rak płuca to drugi najczęściej diagnozowany nowotwór na świecie. W Polsce stanowi główną przyczynę śmierci spowodowaną nowotworem wśród mężczyzn, drugą wśród kobiet. Chirurgiczne wycięcie raka płuca jest metodą radykalną i przez to najskuteczniejszym sposobem leczenia. Z drugiej strony zabiegi chirurgiczne w obrębie

klatki piersiowej wiążą się z dużym ryzykiem zachorowalności pooperacyjnej lub nawet śmierci. Decyzja, czy pacjent będzie zakwalifikowany do zabiegu zależy od stanu pacjenta i postępu choroby. Czynnikiem związanym z wyższym pooperacyjnym ryzykiem zgonu jest wiek. Kolejnym ryzykiem są zmiany somatyczne i zwiększona częstość występowania chorób współistniejących u osób starszych [49]. Mimo, że podeszły wiek nie jest przeciwwskazaniem do operacji, należy dokładnie zbadać inne cechy pacjenta, aby ocenić szanse na pozytywny wynik operacji [144]. Proces podejmowania takiej decyzji mógłby być wspierany i usprawniany przez algorytmy uczenia maszynowego. Obecnie w Polsce decyzja o zakwalifikowaniu pacjenta do operacji opiera się tylko na wiedzy, intuicji i doświadczeniu lekarzy. W tym badaniu przedstawiam model uczenia maszynowego, który przewiduje prawdopodobieństwo 3-miesięcznego przeżycia po operacji raka płuca na podstawie cech pacjenta dostępnych przed operacją. Taki model może wesprzeć decyzję chirurga, ale także dzięki zastosowaniu metod XAI wskazać czynniki ryzyka, czy cechy, które należy zmienić przed zabiegiem w celu zwiększenia prawdopodobieństwa przeżycia pacjenta. Jednocześnie głównym celem tego badania jest sprawdzenie czy metody XAI dostarczają cennych informacji dla badacza w procesie budowy modelu.

3.1.3. Analizowane dane

Badanie oparte jest na danych pochodzących z Polskiego Rejestru Nowotworów Płuca, który zawiera wywiad chorobowy 32698 chorych na raka płuca. To wszystkie polskie przypadki resekcyjnego raka płuca, zebrane w ciągu 14 lat (2002-2016). Dane w badaniu obejmują historię palenia tytoniu, cechy radiologiczne i patologiczne guzków w płucach, wyniki badań, czynniki ryzyka raka, objawy i choroby współistniejące. Szczegółowe informacje dotyczące danych użytych w badaniu podane są w tabelach 3.3, 3.2, 3.1.

W tabeli 3.1 znajdują się dane dotyczące palenia papierosów: liczba lat palenia, liczba papierosów wypalanych dziennie, czy informacja o rzuceniu palenia. Zmienna paczkolata mierzy ekspozycję pacjenta na tytoń. Dla każdego pacjenta obliczana jest jako liczba lat palenia papierosów pomnożona przez liczbę paczek papierosów wypalanych dziennie. W tabeli 3.2 znajdują się dane dotyczące podstawowych cech pacjenta i towarzyszących symptomów choroby. Skala Zubroda określa stan ogólny i jakość życia pacjenta, gdzie 0 oznacza prawidłowy stan, 1 oznacza istnienie objawów choroby,

ale dobry stan pacjenta, 2 oznacza zdolność do wykonywania czynności osobistych ale niezdolność do pracy, 3 oznacza ograniczoną zdolność do wykonywania czynności osobistych, 4 oznacza konieczność spędzenia całego czasu w łóżku i konieczność stałej opieki nad chorym, a 5 oznacza zgon pacjenta. PCO2 to pomiar gazometrii podstawowej, a APTT to czas częściowej tromboplastyny aktywowanej odpowiedzialny za oznaczenie poziomu krzepnięcia krwi. W tabeli 3.3 znajdują się dane dotyczące cech radiologicznych i patologicznych. Są to wstępne oznaczenia zmiennych, które można wykonać jeszcze przed operacją: stadium choroby, rozmiar guza, czy wstępne rozpoznanie histopatologiczne.

Wczesna, 3-miesięczna śmiertelność pooperacyjna wyniosła 7%. Pacjenci, którzy nie przeżyli pierwszych 3 miesięcy po operacji, mają wyższą medianę lat palenia, wypalonych paczek papierosów dziennie oraz wieku niż osoby, które przeżyły ten okres.

Cecha	Nazwa	Zmarł	Przeżył
Płeć	Sex		
Mężczyźni		1577 (70.4%)	19398 (63.7%)
Kobiety		664 (29.6%)	11059 (36.3%)
Wiek	Age		
Średnia (SD)		65.4 (8,23)	63 (8,59)
Mediana [Min, Max]		66 [18, 88.0]	63 [15, 90]
Palenie papierosów	Smoking		
Nie		695 (31%)	10091 (33,1%)
Tak		1546 (69%)	20366 (66,9%)
Liczba lata palenia papierosów	Smoking_years		
Średnia (SD)		24,2 (19,3)	21,4 (19)
Mediana [Min, Max]		30 [0, 80]	25 [0, 80]
Liczba papierosów	Cigarettes_count		
Średnia (SD)		14 (11,4)	12,9 (11,9)
Mediana [Min, Max]		20 [0, 60]	20 [0, 61]
Rzucenie palenia (miesiące)	Quitted_smoking		
Średnia (SD)		19.2 (60,9)	19.4 (60,6)
Mediana [Min, Max]		0 [0, 480]	0 [0, 480]
Paczkolata	Packyears		
Średnia (SD)		25,5 (23.4)	22,4 (23)
Mediana [Min, Max]		30 [0, 165]	20 [0, 189]
APTT	APTT		
Średnia (SD)		32,2 (14,7)	30.3 (7,56)
Mediana [Min, Max]		30.1 [3,02, 200]	29.7 [0,027, 200]
Braki danych		1714 (76,5%)	17396 (57,1%)
PCO2	PCO2		
Średnia (SD)		38,7 (7,06)	38.6 (10,9)
Mediana [Min, Max]		38 [2,5, 111]	38 [2,42, 639]
Braki		1462 (65,2%)	20720 (68%)

Tabela 3.1: Statystyki opisowe zmiennych dotyczących pacjentów chorych na operowalnego raka płuca, część 1. Źródło: publikacja autora [68]

Cecha	Nazwa	Zmarł	Przeżył
Obciążenie nowotworem	`Cancer_Burden`		
Nie		2020 (90,1%)	27202 (89,3%)
Tak		221 (9,9%)	3255 (10,7%)
Zewnętrzne czynniki ryzyka	`External_Risk_Factors`		
Nie		1847 (82,4%)	25132 (82,5%)
Tak		394 (17,6%)	5325 (17,5%)
Choroby układu oddechowego	`Respiratory_System_Diseases`		
Nie		1981 (88,4%)	27809 (91,3%)
Tak		260 (11,6%)	2648 (8,7%)
Plwocina	`Sputum`		
Nie		2204 (98,3%)	29890 (98,1%)
Tak		37 (1,7%)	567 (1,9%)
Zapalenie oskrzeli	`Bronchial_Lavage`		
Nie		2149 (95,9%)	29135 (95,7%)
Tak		92 (4,1%)	1322 (4,3%)
Utrata wagi	`Weight_Loss_kg`		
Średnia (SD)		0,55 (2,36)	0.44 (2,07)
Mediana [Min, Max]		0 [0, 30]	0 [0, 30]
Skala Zubroda	`Zubrod_Performance`		
0		1168 (52,1%)	16705 (54,8%)
1		866 (38,6%)	11562 (38%)
2		193 (8,6%)	2139 (7%)
3		14 (0,6%)	51 (0,2%)
Przerzuty	`Metastesis`		
Nie		2133 (95,2%)	29416 (96,6%)
Tak		108 (4,8%)	1041 (3,4%)
Powiększone węzły chłonne	`Enlarged_Nodes`		
Nie		1663 (74,2%)	24150 (79,3%)
Tak		578 (25,8%)	6307 (20,7%)
Symptomy	`Symptoms`		
Nie		929 (41,5%)	13689 (44,9%)
Tak		1312 (58,5%)	16768 (55,1%)

Tabela 3.2: Statystyki opisowe zmiennych dotyczących pacjentów chorych na operowalnego raka płuca, część 2. Źródło: publikacja autora [68]

Cecha	Nazwa	Zmarł	Przeżył
Płuco	`Lung`		
Lewe		930 (41,5%)	13342 (43,8%)
Prawe		1311 (58,5%)	17115 (56,2%)
Stadium	`Stadium`		
Średnia (SD)		1,85 (0,828)	1,74 (0,781)
Mediana [Min, Max]		1,5 [1, 4]	1.5 [1, 4]
Rozmiar guza	`Tumor_Dimension`		
poniżej 1		36 (1,6%)	541 (1,8%)
[1-2)		307 (13,7%)	4934 (16,2%)
[2-3)		572 (25,5%)	8270 (27,2%)
[3-5)		1104 (49,3%)	14891 (48,9%)
[5-7)		146 (6,5%)	1220 (4%)
[7-10)		62 (2,8%)	512 (1,7%)
ponad 10		14 (0,6%)	89 (0,3%)
Rozpoznanie	`Histopatological_Diagnosis`		
Adenocarcinoma		189 (8,4%)	2682 (8,8%)
Adenosquamous carcinoma		3 (0,1%)	43 (0,1%)
Atypical carcinoma		3 (0,1%)	83 (0,3%)
Bronchoalveolar carcinoma		3 (0,1%)	45 (0,1%)
Non-small cell carcinoma		568 (25,3%)	7369 (24,2%)
Large cell lung carcinoma		11 (0,5%)	131 (0,4%)
Neuroendocrine carcinoma		4 (0,2%)	81 (0,3%)
Pleomorphic carcinoma		2 (0,1%)	18 (0,1%)
Small cell lung carcinoma		6 (0,3%)	105 (0,3%)
Squamous carcinoma		428 (19,1%)	5384 (17,7%)
Typical carcinoma		16 (0,7%)	397 (1,3%)
Inne		10 (0,4%)	174 (0,6%)
Niezdefiniowane		998 (44,5%)	13945 (45,8%)

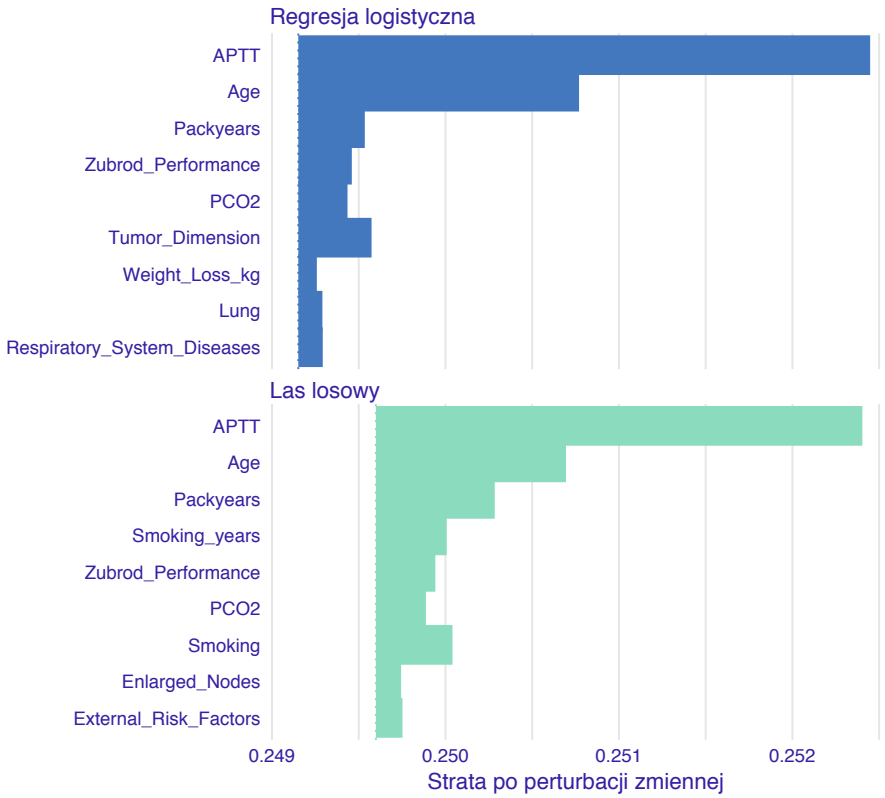
Tabela 3.3: Statystyki opisowe zmiennych dotyczących pacjentów chorych na operowalnego raka płuca, część 3. Źródło: publikacja autora [68]

3.1.4. Modelowanie

W celu oceny czy pacjent przeżyje pierwsze 3 miesiące po operacji zostały użyte dwa rodzaje modeli klasyfikacyjnych: regresja logistyczna i las losowy. Regresja logistyczna jest częstym wyborem dla binarnych problemów klasyfikacji. Model jest łatwy do interpretacji. Niestety często dobra intepretowalność uzyskana jest kosztem dokładności tego modelu. Drugi model użyty w analizie to model lasu losowego, który jest przykładem czarnej skrzynki, trudniejszym do interpretacji, ale często osiągającym lepsze wyniki niż modele tradycyjne. W zastosowaniach medycznych modele regresji wciąż są bardzo często stosowane, na przykład w badaniu [63] zastosowano regresję logistyczną w celu predykcji śmiertelności u osób zainfekowanych COVID-19. Coraz częściej jednak obserwuje się badania pokazujące zarówno klasyczne modele jak i te bardziej złożone, modele uczenia maszynowego. W badaniu [103] przewidywano śmiertelność pooperacyjną na podstawie regresji logistycznej oraz sieci ANN. W artykule [142] z kolei modelowano problem śmiertelności pacjentów poddawanych operacji zastawek serca w oparciu o regresję logistyczną LASSO i szereg modeli uczenia maszynowego. W badaniu [110] przedstawiono metody wyjaśnialnej sztucznej inteligencji zastosowane do problemu modelowania przeżywalności pooperacyjnej.

Zbiór danych został podzielony na dwa podzbiory, uczący, który zawiera 85% przypadków i testowy składający się z pozostałych 15%. Problem modelowania śmiertelności pooperacyjnej jest skomplikowany ze względu na niezbalansowaną zmienną objaśnianą (mały odsetek osób, które umierają w przeciągu 3 miesięcy po operacji). Śmiertelność pooperacyjna wynosiła tylko 7%. Dlatego w modelach klasyfikacyjnych, ustalony został próg odcięcia na poziomie 93%. Pomimo pewnych braków danych w badanym zbiorze, do modelowania wykorzystano wszystkie zmienne objaśniające przedstawione w podrozdziale 3.1.3. Ze względu na współliniowość niektórych zmiennych ograniczyłam zbiór zmiennych objaśniających w modelu regresji logistycznej.

W oparciu o zmienne kliniczne powstały modele przewidujące 3-miesięczne przeżycie po operacji. Skuteczność modeli została potwierdzona przy pomocy walidacji krzyżowej i obliczona za pomocą dwóch miar dla każdego modelu: pola pod krzywą ROC (AUC, ang.) i średniego błędu klasyfikacji (MMCE, ang. *mean misclassification error*). Obie miary regresji logistycznej były nieco niższe na zbiorze testowym w porównaniu z wartościami dla lasu losowego: wartość AUC wyniosła $0,65$ dla regresji logistycznej



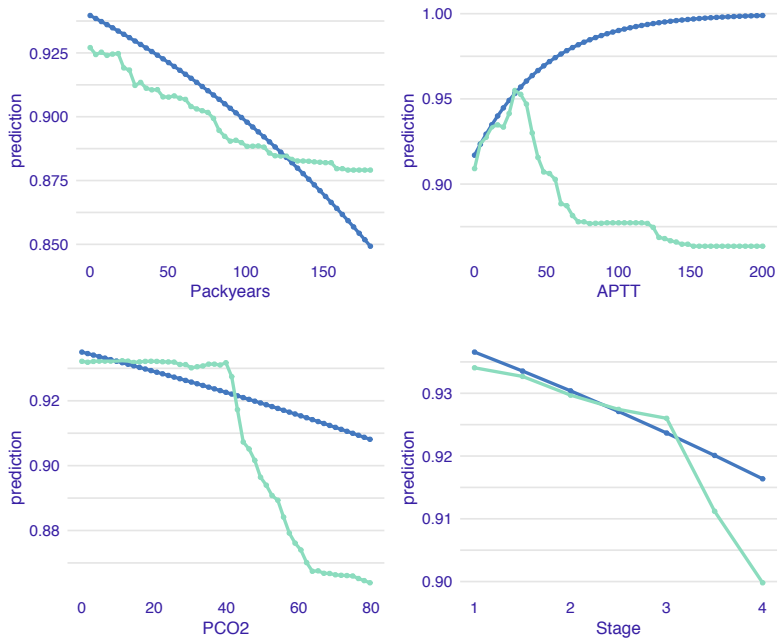
Rysunek 3.2: Ważność zmiennych dla dwóch analizowanych modeli: regresji logistycznej (górny wykres) i lasu losowego (dolny wykres). Źródło: publikacja autora [68]

i $0,67$ dla lasu losowego; Wartości MMCE wynosiły odpowiednio $0,39$ i $0,4$. Zastosowałam metody XAI, by sprawdzić czy mogą one pomóc w wyborze optymalnego modelu. Poniżej przedstawiam kolejno zastosowane metody.

Globalne metody wyjaśniania

W analizie przedstawionej na Rysunku 3.2, pokazana została ważność poszczególnych zmiennych dla zgromadzonych modeli na całym zbiorze danych. Im dłuższy słupek, tym większe znaczenie przypisywane jest danej zmiennej. Zgodnie z przedstawionym wykresem, dla obu modeli najważniejszą zmienną jest `APTT`. Dodatkowo, zarówno dla modelu lasu losowego, jak i regresji logistycznej, istotnymi cechami są `Age`, `Packyears`, `Zubrod_Performance` i `PCO2`.

Model lasu losowego wskazuje również na ważność zmiennych `Smoking_years`, `Smoking`, `Enlarged_Nodes` i `External_Risk_Factors`. Natomiast regresja logistyczna



Rysunek 3.3: Wykresy PDP dla najbardziej istotnych zmiennych. Największe różnice pomiędzy modelami są widoczne dla `APTT` i `PCO2`. Źródło: publikacja autora [68]

wskazuje na ważność zmiennych `Respiratory_System_Diseases`, `Tumor_Dimension`, `Weight_Loss_kg` oraz `Lung`.

Na wykresie została również oznaczona przerywaną pionową linia, a zarazem początkiem słupków, wartość błędu średniokwadratowego. Wyniki analizy wskazują, że zarówno regresja logistyczna, jak i model lasu losowego osiągają bardzo zbliżone wartości tego błędu.

Przedstawione wyniki mogą mieć istotne znaczenie dla interpretacji i wyboru modeli w kontekście badanej dziedziny. Wskazują one na istotność określonych zmiennych oraz podobieństwo bądź różnice wyników osiąganych przez różne modele. Przedstawione obserwacje mogą przyczynić się do wyboru modelu, który budzi większe zaufanie eksperta z danej dziedziny. Mogą też stanowić podstawę dla dalszych badań i udoskonaleń modeli w tym obszarze. W przypadku przedstawionych modeli, oba modele wskazują na te same dwie najistotniejsze zmienne. Zauważalną różnicą jest podejście do zmiennych związanych z paleniem papierosów. Za jedyną istotną zmienną model lasu losowego uwzględnił paczkolata `Packyears`. Natomiast regresja logistyczna oprócz paczkolat,

wskazuje zarówno na lata palenia (`Smoking_years`), jak i na fakt, czy pacjent w ogóle palił papierosy (`Smoking`).

Drugą metodą analizy i porównania modeli są wykresy cząstkowej zależności PDP. Stanowią one istotne narzędzie do analizy relacji, jakie model wykrył między zmiennymi a predykcjami. Przedstawiają zmienność przewidywanej średniej predykcji w zależności od wartości analizowanej zmiennej. W przypadku niniejszego badania przedstawione są wykresy PDP dla zmiennych `APTT`, `Age`, `Packyears` i `PCO2`, ze względu na ich ważność dla obu modeli.

Na wykresach PDP dla zmiennej `Packyears` oba modele wykazują podobną zależność. Obserwuje się intuicyjną relację, gdzie większa wartość paczkolat (liczba lat palenia papierosów pomnożona przez liczbę paczek wypalanych dziennie), prowadzi do gorszych prognoz dla danego pacjenta. Oznacza to, że większa ekspozycja na palenie papierosów negatywnie wpływa na wyniki modelu.

Analiza wykresów PDP dla zmiennych `APTT` i `PCO2` w przypadku modelu lasu losowego wykazuje nieliniowe zależności. Zmienna `APTT` przedstawia krzywoliniowy wzorzec, co sugeruje, że wpływ tej zmiennej na odpowiedź modelu nie jest liniowy. Natomiast zmienna `PCO2` wykazuje pewną nieregularność w zależnościach, co może wskazywać na bardziej skomplikowany wpływ tej zmiennej na prognozy lasu losowego.

Wykresy PDP dla zmiennej `Stadium` w obu modelach ukazują podobne zależności. Jednak różnica występuje w przypadku pacjentów w stadium 4, gdzie średnia odpowiedź modelu lasu losowego jest znacznie niższa niż dla regresji logistycznej.

Przedstawione analizy przy użyciu wykresów PDP mają kluczowe znaczenie dla walidacji modeli. Umożliwiają zrozumienie zależności, które modele wykazują między zmiennymi a odpowiedziami. W przypadku badanych zmiennych, wykresy PDP dla lasu losowego potwierdzają wiedzę domenową oraz poprawność zależności wskazywanych przez modele. W przypadku regresji logistycznej, nie wszystkie zależności są zgodne z wiedzą i intuicją.

Przeprowadzając porównanie wydajności modeli za pomocą miar AUC i MMCE, można stwierdzić, że oba modele działają bardzo podobnie. Regresja logistyczna, będąca tradycyjnym modelem, posiada prostszą interpretację co skłaniałoby do wyboru tego modelu. Niestety, analiza wykresów PDP wskazuje, że ten model może nieprawidłowo wykorzystywać zmienne pokazując tylko zależności liniowe. Z kolei wykresy PDP dla modelu lasu losowego wskazują na nieliniowe zależności. Dzięki zastosowaniu

PDP dla obu modeli, istnieje możliwość lepszego wykorzystania zmiennych poprzez ich ponowne zakodowanie w celu ulepszenia modelu regresji.

Na podstawie wykresów PDP można zidentyfikować dwie zmienne, tj. `PCO2` i `APTT`, które wykazują nieliniowe zależności. W celu wykorzystania tych zmiennych w modelu regresji, konieczne jest ich odpowiednie przekształcenie, na przykład poprzez kodowanie ich jako cechy binarne. W przypadku zmiennej `PCO2`, wartość nowo utworzonej zmiennej wynosi 1, jeśli poziom `PCO2` przekracza 40, w przeciwnym razie przyjmuje wartość 0. Analogicznie jest przekształcana zmienna `APTT`, gdzie wartość cechy wynosi 1, gdy mieści się w przedziale (5, 44), a w przeciwnym przypadku wynosi 0.

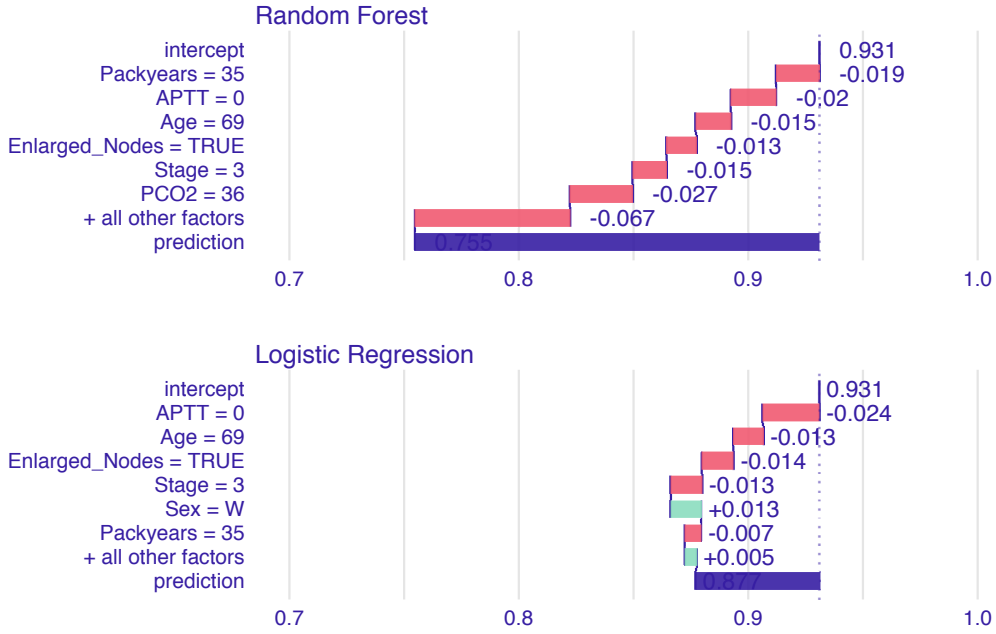
Po dokonaniu takiej transformacji, średnia wartość AUC dla regresji logistycznej, oszacowana za pomocą 5-krotnej walidacji krzyżowej, wynosi 0,66. Ta wartość AUC jest tylko o 0,01 większa w porównaniu z wynikami uzyskanymi na nieprzekształconym zbiorze danych.

Globalne metody XAI, zastosowane w tym przypadku, przyczyniają się do uchwycenia prawidłowych relacji między zmiennymi a odpowiedzią modelu. Dodatkowo, mogą one pomóc w dokładnym przekształceniu zmiennych w celu poprawy jakości modelu. Analiza miar dokładności modeli jedynie wskazuje na bardzo zbliżone wyniki, co mogłoby skłaniać do wyboru modelu interpretowalnego, takiego jak regresja logistyczna. Jednakże, model ten nieprawidłowo przedstawia zależności między zmiennymi a predykcją, co zostało zauważone dzięki zastosowaniu technik XAI.

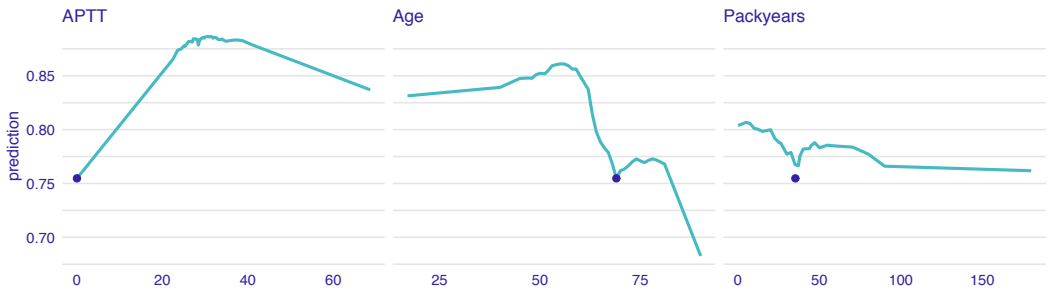
Wykresy PDP odgrywają istotną rolę w procesie odpowiedniego przekształcania zmiennych i pozwalają uniknąć błędnych wskazań modelu liniowego, a także wpływają na poprawę jakości modelu. Dzięki metodom XAI, możliwe jest wykrycie nieprawidłowości i skuteczne wykorzystanie wykresów PDP do dokładniejszej transformacji zmiennych, co przyczynia się do poprawy jakości i usprawnienia modelu.

Lokalne metody wyjaśniania

Lokalne metody XAI, takie jak profil ceteris paribus i dekompozycja modelu breakdown, są niezwykle interesujące, ponieważ umożliwiają dogłębne zrozumienie predykcji dla konkretnego pacjenta. Zwłaszcza w kontekście podejmowania decyzji dotyczących kwalifikacji do operacji, istotne jest zrozumienie działania modelu dla jednostki. Wyjaśnienia te mogą stanowić wsparcie dla lekarza, umożliwiając analizę udziału poszczególnych cech w ostateczną prognozę. Mogą także wspomóc zaufanie pacjenta do decyzji kwalifikacyjnej. Co więcej, profile ceteris paribus mogą wskazać, którą cechę należy



Rysunek 3.4: Dekompozycja break-down prezentuje wkłady poszczególnych zmiennych w predykcję dla wybranego pacjenta. Zielone słupki odpowiadają za dodatni wpływ, a czerwone za ujemny wpływ na prognozę. Wszystkie wartości odpowiadające zmiennym sumują się do ostatecznej predykcji (fioletowy słupek). Przerywana, pionowa linia odpowiada za średnią odpowiedź modelu. Źródło: publikacja autora [68]



Rysunek 3.5: Profile ceteris paribus dla wybranego pacjenta i trzech zmiennych. Lewy wykres wskazuje, że prognoza rośnie wraz ze wzrostem zmiennej APTT do wartości 35, następnie maleje. Wykres środkowy wskazuje na wyższą predykcję dla młodszych. Prawy wykres wskazuje niewielkie oscylacje predykcji. Źródło: publikacja autora [68]

Zmienna	**Wartość**	**Zmienna**	**Wartość**
Płeć	Kobieta	Choroby układu oddechowego	Nie
Wiek	69	Zewnętrzne czynniki ryzyka	Tak
Palenie papierosów	Tak	Wstępne rozpoznanie hist.	Not Defined
Liczba lat palenia	44	Płuco	Prawe
Liczba papierosów	16	Plwocina	Nie
Rzucenie palenia (miesiące)	0	Zapalenie oskrzeli	Nie
Paczkolata	35	Total lung capacity norm	0
Cancer burden	No	PCO2	35,6
Powiększone węzły	Tak	Fibrynogen	0
Utrata wagi	0	PTTS	0
Symptomy	Yes	APTT	0
Skala Zubroda	0	Stadium	3
Wielkość guza	3-5	Przerzuty	Nie
Czy przeżył 3 miesiące:		**Nie**	

Tabela 3.4: Wartości zmiennych dla wybranego pacjenta, dla którego wykonana jest lokalna analiza. Źródło: publikacja autora [68]

poprawić, aby zwiększyć szanse pacjenta na przeżycie operacji.

Na rysunku 3.9 przedstawiona jest dekompozycja break-down, która prezentuje atrybucje zmiennych dla modeli lasu losowego i regresji logistycznej. Szczegółowe informacje na temat wybranego pacjenta zawarte są w tabeli 3.4. Przerywana, pionowa linia na wykresach reprezentuje średnie odpowiedzi modeli (dla obu modeli wynosi 0,93). Czerwone słupki wskazują cechy, które mają negatywny wkład i pogarszają predykcję w porównaniu do średniej predykcji modelu, podczas gdy zielone słupki wskazują cechy, które poprawiają predykcję. Cechy o niewielkim wpływie są agregowane. Ostateczna prognoza modelu jest reprezentowana przez fioletowy słupek. Oba modele wskazują, że pacjent nie przeżyje operacji (predykcje są niższe niż ustalony próg). Jednak decyzje zostały podjęte nieco inaczej. W przypadku regresji logistycznej, bycie kobietą ma pozytywny wpływ na predykcję. Natomiast las losowy nie wskazuje na dodatni wkład żadnej zmiennej. Oba modele wskazują, że wartości zmiennych `APTT`, `Packyears`, `Age`, `Enlarged_Nodes`, `Stadium` mają negatywny wpływ na predykcję. Las losowy dodatkowo wskazuje na negatywny wpływ zmiennej `PCO2`.

Analizę profili ceteris paribus przedstawia Rysunek 3.5, na którym przedstawione są zmienne o największym negatywnym wpływie na predykcje (zgodnie z dekompozycją

break-down). Na każdym wykresie można zobaczyć, jaka byłaby odpowiedź modelu lasu losowego, gdyby zmienna `APTT` (na lewym wykresie), `Age` (na środkowym wykresie) i `Packyears` (na prawym wykresie) uległa zmianie, podczas gdy inne zmienne wejściowe pozostałyby niezmienione.

Na podstawie lokalnej analizy modelu nie możemy wnioskować o poprawności i jakości modelu, ale możemy dokładnie zbadać predykcje dla poszczególnych obserwacji. W dziedzinie medycyny tego rodzaju analiza ma szczególne zastosowanie, zwłaszcza gdy istnieje możliwość zmiany wartości danej zmiennej, na przykład sposobu leczenia w celu poprawy rokowania.

3.1.5. Podsumowanie

Metoda PDP odegrała istotną rolę w tym badaniu. Na etapie modelowania, metoda ta umożliwiła porównanie zależności między zmiennymi a odpowiedziami modeli uczenia maszynowego. Pozwoliło to ocenić, czy modele traktują daną zmienną w podobny sposób. W przypadku, gdy różne modele traktują tę samą zmienną w podobny sposób, można potwierdzić ich poprawność. Z kolei, jeśli modele różnie traktują daną zmiennej, można wyciągnąć wnioski dotyczące złożoności i niejednoznaczności tej zależności bądź istnienia błędów w modelowaniu. Należy wtedy skupić szczególną uwagę na tej zmiennej i przeprowadzić dogłębną analizę mającą na celu zidentyfikowanie możliwości poprawy modelowania. Możliwe kroki mogą obejmować modyfikację procesu uczenia maszynowego, zmianę hiperparametrów modelu lub dostosowanie danych w celu lepszego uwzględnienia tej zmiennej. W przypadku wykrycia takich nieprawidłowości konieczne jest przeprowadzenie analizy samej zmiennej oraz kontekstu, w którym jest ona stosowana. W tym procesie należy skupić się na ocenie, czy istnieje możliwość dostarczenia lepszych danych, a także czy istnieje potrzeba zastosowania innych technik modelowania.

Porównanie zależności między zmiennymi a odpowiedziami modeli pozwala na identyfikację podobieństw i różnic między modelami, co ma kluczowe znaczenie w kontekście badanego obszaru nauk medycznych. Wnioski wyprowadzone z analizy PDP stanowią istotny wkład w dyskusję na temat jakości modelowania i interpretacji wyników w dziedzinie medycyny.

Otrzymane wyniki wskazują, że metody wyjaśnialnego uczenia maszynowego (XAI)

są nie tylko przydatne do poprawy jakości modelu, ale także mogą pomóc w dokonaniu wyboru lepszego modelu. Analiza przy pomocy wykresów PDP umożliwia identyfikację różnic w traktowaniu zmiennych przez dwa modele. Mimo niewielkich różnic pod względem jakości modeli, udało się zauważyć duże różnice w zależnościach między zmiennymi a predykcjami modeli. Porównanie uzyskanych wyników z wiedzą domenową oraz wykorzystanie metod XAI umożliwia identyfikację błędów modelu. Ponadto wykorzystanie PDP pozwala na usprawnienie modelu, który nieprawidłowo uchwytuje zależności między zmiennymi. W rezultacie potwierdza się hipoteza 1 przedstawiona w sekcji 1.3.

3.2. Metody XAI na etapie walidacji modeli

Rozdział ten powstał na podstawie wyników opublikowanych w publikacji [69]:

Katarzyna Kobylińska, Tadeusz Orłowski, Mariusz Adamek, Przemysław Biecek, "Explainable Machine Learning for Lung Cancer Screening Models"Applied Sciences 2022, 12(4).

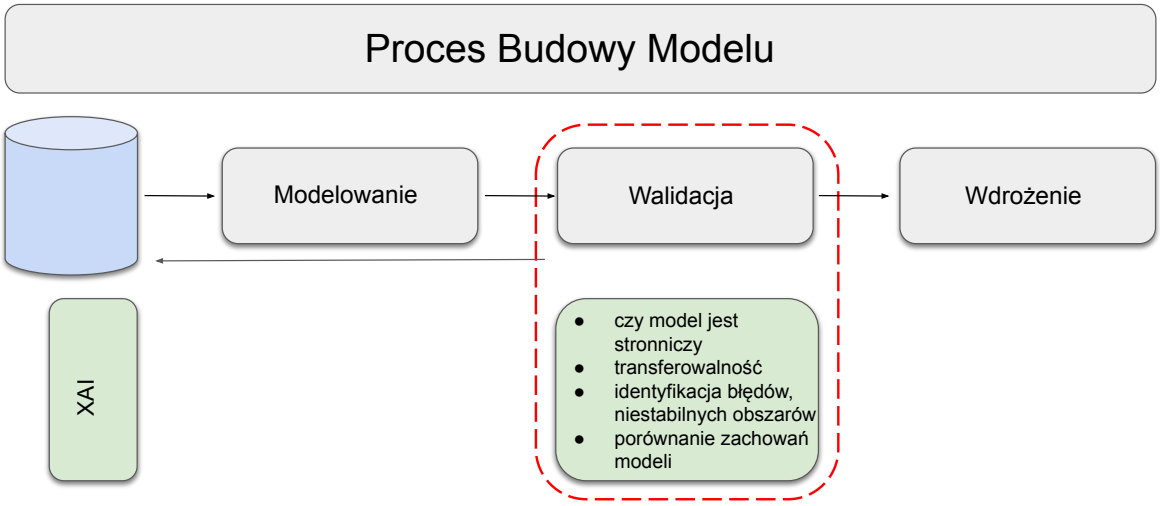
Stanowi odpowiedź na hipotezę 2 z Sekcji 1.3, która brzmi: "Metody wyjaśnialnego uczenia maszynowego post-hoc identyfikują słabe strony modeli. Metody te pozwalają na identyfikację ograniczeń i niedoskonałości modeli."W tym rozdziale przedstawiam dowody na prawdziwość tej hipotezy.

3.2.1. Wprowadzenie

W niniejszym rozdziale badam, czy metody XAI mogą wspomóc proces walidacji modeli. Walidacja modeli jest etapem, który odbywa się po zgromadzeniu różnych modeli i stanowi część procesu wyboru optymalnego modelu. Jej celem jest porównanie i ocena jakości oraz poprawności poszczególnych modeli by wyłonić ten, który najlepiej odzwierciedla badane zjawisko.

W celu potwierdzenia postawionej hipotezy przedstawiam przykład wykorzystania metod XAI do walidacji już istniejących modeli, mając na celu wybór najbardziej odpowiedniego spośród nich. Badanie to skupiło się na zastosowaniu metod XAI do modeli przesiewowych dla raka płuca. W tym przypadku, nie przeprowadzałam pełnego procesu wyboru modelu, lecz sam etap walidacji i porównania modeli, co zostało zilustrowane na Rysunku 3.6. Wykorzystując metody XAI, analizuję zachowanie tych modeli na wybranej populacji pacjentów. Poprzez porównanie uzyskanych wyników z wiedzą domenową, identyfikuję potencjalne błędy i niezgodności w działaniu modeli, co pozwala na dokonanie wyboru modelu.

W sekcji 3.1 pokazałam, że metody XAI mogą poprawić jakość modelowania na etapie budowy nowych modeli. W tym badaniu interesuje mnie czy metody XAI mogą doprowadzić do wyboru jednego, najlepszego modelu lub przeciwnie, do odrzucenia modelu, nawet jeśli dokładność modeli liczona miarą jakości jest bardzo zbliżona, a modele są już wdrożone i powszechnie stosowane.



Rysunek 3.6: Metody XAI użyte na etapie walidacji modeli w procesie budowy modelu.

3.2.2. Problem medyczny

Wykrycie raka płuca na wczesnym etapie rozwoju zwiększa szanse na przeżycie [94]. Jednak na takim etapie nowotwór ten zwykle nie daje charakterystycznych objawów. Jednym ze sposobów walki są badania przesiewowe. Pokazano, żę niskodawkowa tomografia komputerowa (LDCT) jest w stanie wykryć wczesnego raka płuc. Wyniki dwóch randomizowanych badań klinicznych (NLST i NELSON) jednoznacznie wykazały zmniejszenie śmiertelności z powodu raka płuca w grupie interwencyjnej, tj. u osób, które zaplanowały podstawowy i coroczny LDCT, więcej szczegółów dostępnych jest w publikacjach [30] i [25]. Dodatkową zaletą tej procedury diagnostycznej jest uzyskanie znacznej ilości informacji o innych patologiach lub chorobach współistniejących, co stanowi kolejny wymiar działań profilaktycznych z wykorzystaniem techniki obrazowania. W badaniu [116] pokazano, że przesiewowa tomografia komputerowa może zmniejszyć śmiertelność z powodu raka płuc nawet o 20% u pacjentów z grupy wysokiego ryzyka. Modele statystyczne pomagają precyzyjniej rejestrować pacjentów, którzy mogą odnieść największe korzyści z badań, zabezpieczając jednocześnie efektywnie alokowane dostępnych środków finansowych. Opracowano wiele modeli przewidujących

	LCRAT	**BACH**	**PLCO$_{m2012}$**
Wiek	+	+	+
Płeć	+	+	
Rasa	+		+
Edukacja	+		+
BMI	+		+
Palenie papierosów			+
Rzucenie papierosów (w latach)	+	+	+
Liczba lat palenia	+	+	+
Liczba papierosów wypalanych dziennie	+	+	+
Paczkolata	+		
Wcześniejszy nowotwór			+
Choroby płuc	+		+
Ekspozycja na azbest		+	
Rak płuca w rodzinie	+		+
Liczba krewnych z rakiem płuca	+		
Liczba zmiennych	12	6	11
Ryzyko zachorowania w przeciągu	5 lat	10 lat	6 lat

Tabela 3.5: Podsumowanie zmiennych użytych w modelach LCRAT, BACH, PLCO$_{m2012}$. Znak + oznacza, że zmienna jest użyta w modelu.

ryzyko zachorowania na raka płuca lub ryzyko zgonu z powodu raka płuca. Modele te są stosowane w gabinetach lekarskich w celu oceny prawdopodobieństwa zachorowania na nowotwór pacjenta i klasyfikacji go do dalszych badań przesiewowych.

Autorzy [65] opracowali porównanie dziewięciu modeli przesiewowych. W badaniu wykazali, że cztery z dziewięciu wiodących modeli ryzyka mają zbliżoną skuteczność i są najdokładniejsze. Na podstawie danych z National Health Interview Survey z lat 2010-2012, następujące modele najdokładniej przewidują ryzyko raka płuc: BACH [2], PLCO$_{m2012}$ [117], LCRAT i LCDRAT [64]. W mojej analizie przedstawiam i porównuję tylko modele oceniające ryzyko raka płuca: model BACH, model PLCO$_{m2012}$ oraz model LCRAT. Model LCDRAT przewiduje ryzyko zgonu z powodu raka płuca, a nie ryzyko wystąpienia raka płuca, dlatego wyłączyłam go z dalszych analiz.

Hormuzd Katki wraz ze współautorami [65] porównali modele na dwóch kohortach (AARP i CPS-II). Wartości pola pod krzywą ROC (AUC) dla wyżej opisanych trzech modeli wahają się od 0.75 do 0.77 i są wyższe niż dla pozostałych modeli przesiewowych.

Gdy porównamy tylko AUC analizowanych modeli, model LCRAT osiąga najwyższą wartość równą 0.77, natomiast model BACH najniższą wartość równą 0.75. Należy zauważyć, że różnice między AUC tych trzech modeli są niewielkie.

Modele przesiewowe raka płuca są opracowywane przy użyciu różnych technik statystycznych. Modele BACH i LCRAT wykorzystują modele regresji Coxa dla danych dotyczących przeżycia. Model $PLCO_{m2012}$ oparty jest na regresji logistycznej. Modele opierają się na innych zestawach zmiennych, które zostały wybrane przez autorów modeli. Dla każdego modelu ryzyko raka płuc jest przewidywane w różnych okresach czasu (5, 6, 10 lat). W tabeli 3.5 znajduje się szczegółowe podsumowanie. Chociaż modele te oparte są na klasycznych, interpretowalnych metodach statystycznych, to same modele przesiewowe są złożone i trudne do wyjaśnienia i porównania między sobą.

3.2.3. Model przesiewowy raka płuca - model BACH

Równanie 3.1 przedstawia funkcję ryzyka w modelu BACH, który opiera się na wielowymiarowej regresji proporcjonalnego hazardu Coxa.

$$risk = \sum_{i=0}^{n}(1 - S_0^{exp(\beta X_i)})(S_1^{exp(\beta X_i)}) \prod_{j<i}(S_0^{exp(\beta X_j)})(S_1^{exp(\beta X_j)}), \tag{3.1}$$

gdzie:

- S_0 odpowiada wyjściowemu przeżyciu bez rozpoznania raka płuca dłużej niż 1 rok, jego oszacowanie wynosi 0.996229,
- S_1 odpowiada wyjściowemu całkowitemu przeżyciu powyżej 1 roku, jego oszacowanie wynosi 0.9917663,
- X jest macierzą wartości zmiennych dla danej osoby,
- X_i to wartości $i - tej$ zmiennej,
- β są współczynnikami wyestymowanymi w modelu, szczegóły w Tabeli 3.6.

Model BACH szacuje prawdopodobieństwo zachorowania na raka płuca w przeciągu 10 lat. Efekty nieliniowe zostały dostarczone przez splajny sześcienne, dopasowane do predyktorów ciągłych (na przykład wieku). Nawet model z małą liczbą zmiennych

Zmienna	Wyrażenie	Współczynnik
$Intercept$		$-9.796\,057\,1$
$wiek$		$0.070\,322\,812$
$wiek_2$	$(wiek - 53.459001)^3 * I(age > 53))$	$-0.000\,093\,821\,22$
$wiek_3$	$(wiek - 61.954825)^3 * I(age > 61)$	$0.000\,182\,826\,61$
$wiek_4$	$(wiek - 70.910335)^3 * I(age > 70)$	$-0.000\,089\,005\,389$
$female$		$-0.058\,272\,61$
$qtyears$		$-0.085\,684\,793$
$qtyears_2$	$(qtyears)^3$	$0.006\,549\,969\,3$
$qtyears_3$	$(qtyears - 0.50513347)^3 * I(qtyears > 0)$	$-0.006\,830\,584\,5$
$qtyears_4$	$(qtyears - 12.295688)^3 * I(qtyears > 12)$	$0.000\,280\,615\,19$
$smkyears$		$0.114\,252\,97$
$smkyears$	$(smkyears - 27.6577)^3 * I(smkyears > 27)$	$-0.000\,080\,091\,477$
$smkyears_3$	$(smkyears - 40)^3 * I(smkyears > 40)$	$0.000\,170\,694\,83$
$smkyears_4$	$(smkyears - 50.910335)^3 * I(smkyears > 50)$	$-0.000\,090\,603\,358$
cpd		$0.060\,818\,386$
cpd_2	$(cpd - 15)^3 * I(cpd > 15)$	$-0.000\,146\,522\,16$
cpd_3	$(cpd - 20.185718)^3 * I(cpd > 20)$	$0.000\,184\,869\,38$
cpd_4	$(cpd - 40)^3 * I(cpd > 40)$	$-0.000\,038\,347\,226$
asb		$0.215\,393\,6$

Tabela 3.6: Wyrażenia i współczynniki β w modelu BACH. Źródło: publikacja autora [69].

objaśniających, oparty na dobrze znanym wzorze matematycznym, może być trudny do zrozumienia.

Pomimo tego, że jest to model powszechnie stosowany, nadal konieczna jest walidacja, jak model zachowuje się na nowych danych lub innych populacjach pacjentów. Z tego powodu niezbędna jest analiza takich modeli na nowych zbiorach danych. Możemy wtedy empirycznie sprawdzić, czy działają poprawnie na określonej populacji.

3.2.4. Analizowane dane

Modele przesiewowe zostały poddane analizie na zbiorze danych dotyczących chorych na operowalnego raka płuca w Polsce, opisanych szczegółowo w rozdziale 3.1.3. Dane

Cecha	Nazwa	Wartość
Wiek	`age`	
Średnia (SD)		63.02 (8.66)
Liczba lat palenia	`smkyears`	
Średnia (SD)		21.59 (18.9)
Rzucenie palenia (miesiące	`qtyears`	
Średnia (SD)		1.7 (5.03)
Liczba papierosów (dziennie)	`cpd`	
Średnia (SD)		13.02 (11.84)
Płeć	`female`	
Mężczyzna (0)		22219 (64.6%)
Kobieta (1)		12174 (35.4%)
Rasa	`race`	
Biała (0)		34393 (100%)
Rozedma płuc	`emp`	
Nie (0)		34284 (99.7%)
Tak (1)		109 (0.3%)
Liczba krewnych z rakiem płuca	`fam.lung.trend`	
0 (0)		33332 (96.9%)
1 lub więcej (0)		1061 (3.1%)
Azbest	`asb`	
Nie (0)		34223 (99.5%)
Tak (1)		170 (0.5%)

Tabela 3.7: Statystyki opisowe zmiennych wykorzystanych w badaniu. Zmienne ciągłe są przedstawione za pomocą średniej i odchylenia standardowego, a zmienne dyskretne za pomocą częstości. Źródło: publikacja autora[69]

te zawierają większość zmiennych niezbędnych do obliczenia prognozy ryzyka na podstawie modeli przesiewowych. Dwie zmienne: BMI (ang. *body mass index*) i poziom wykształcenia nie są dostępne w tym zbiorze danych. Te dwie zmienne zostały zatem zasymulowane na potrzeby tego badania. Statystyki opisowe zmiennych wykorzystywanych w badaniu są pokazane w tabeli 3.7.

3.2.5. Walidacja modeli przesiewowych

Porównanie modeli opiera się na porównaniu poprawności ich predykcji, a nie porównaniu miar jakości. Kohorta składa się wyłącznie z pacjentów z wykrytym rakiem płuca, więc miary jakości nie miałyby charakteru informacyjnego. Co więcej, takie porówna-

nie zostało dokładnie zbadane w innych dostępnych analizach, na przykład w artykule [65]. W tym rozdziale przedstawiam porównanie predykcji modeli przy pomocy metod XAI. W przeciwieństwie do porównywania modeli tylko przy pomocy wybranej miary jakości (np. AUC czy ACC), porównanie metodami XAI pozwala na ocenę różnic i podobieństw modeli z różnych perspektyw.

Globalne metody wyjaśnienia

Pierwszą zastosowaną metodą XAI służącą porównaniu modeli jest wykres przedstawiający ważności zmiennych zaprezentowany na Rysunku 3.7. Modele wykorzystują inne zbiory zmiennych stąd niektóre zmienne ważne dla jednego modelu, nie są w ogóle uwzględnione w przypadku innego modelu. Taka sytuacja ma miejsce w przypadku zmiennej opisującej narażenie na azbest. Zmienna ta jest ważna dla modelu BACH, a nie jest w ogóle uwzględniona w dwóch pozostałych modelach. Z kolei zmienne opisujące występowanie raka płuca w rodzinie, rasę i rozedmę płuc, które są ważne dla modeli LCRAT i PLCO_{m2012}, nie są uwzględnione w modelu BACH. Ciekawa jest analiza części wspólnej zbioru zmiennych. Takie cechy, które są istotne dla przynajmniej jednego modelu to `age`,`smkyears`, `cpd` oraz `qtyears`. Wiek pacjenta jest najważniejszą zmienną dla modeli LCRAT i PLCO_{m2012}, ale w modelu BACH zajmuje dopiero czwartą pozycję. Podobnie różnica dotyczy zmiennej opisującej lata palenia papierosów `smkyears`. Jest to najważniejsza zmienna dla modelu BACH, natomiast w modelu PLCO_{m2012} zajmuje czwartą pozycję. Według modelu LCRAT liczba lat palenia jest uważana za drugą po wieku, najważniejszą cechę pacjenta. Warto też zauważyć że względna ważność tych zmiennych jest różna w zależności od modelu.

Drugą metodą XAI użytą w badaniu jest cząstkowa zależność PDP zaprezentowana na Rysunku 3.8. Przedstawiam tą metodę dla czterech zmiennych, które zostały wybrane na podstawie ważności zmiennych. Wykresy te podsumowują jak zmienia się predykcja modelu w zależności od zmiany wybranej zmiennej. W zestawieniu wykresów PDP dla modeli przesiewowych, największe różnice widać dla dwóch zmiennych: `age` i `qtyears`. Warto zauważyć, że rozważane modele przewidują ryzyko wystąpienia raka płuca w różnych okresach czasu, dlatego porównujemy kształty krzywych, a nie wartości bezwzględne prognoz. Predykcja modelu PLCO_{m2012} dotycząca wieku znacznie różni się od innych modeli. Model BACH dla najstarszych osób wskazuje na zależność sprzeczną z intuicją, podczas gdy przewidywania modelu PLCO_{m2012} bardzo rosną dla osób w wieku powyżej 75 lat. W przypadku modelu LCRAT prognoza wzrasta wraz

z wiekiem, ale tylko do wieku 77 lat. Dla osób najstarszych, wykres wypłaszcza się a prognozy pozostają niezmienione.

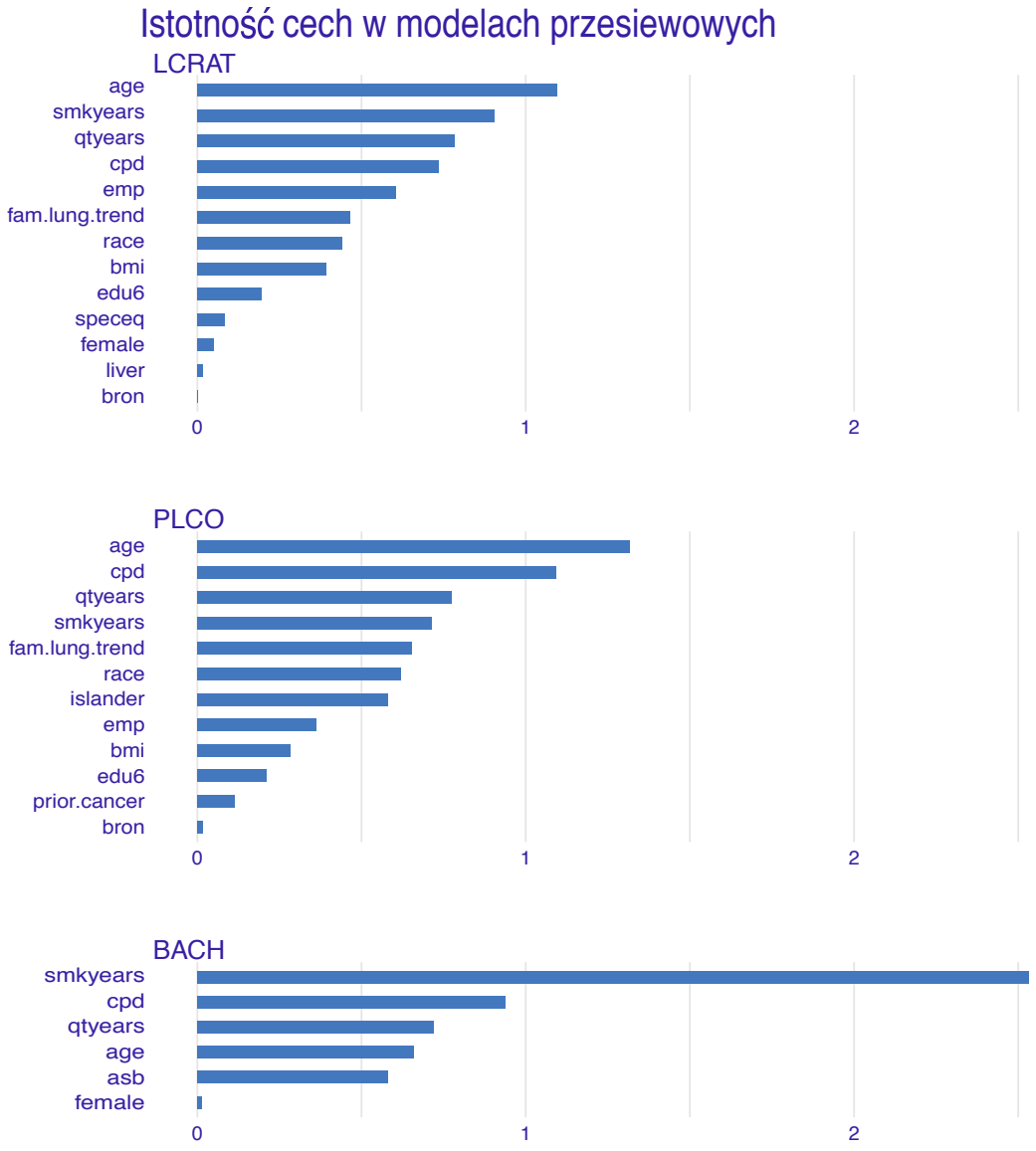
Wykresy wskazują na podobne zależności między latami palenia (`smkyears`) a predykcjami dla 2 modeli: LCRAT i PLCO_{m2012}. W przypadku czasu rzucenia palenia (`qtyears`) model BACH ponownie pokazuje sprzeczną z intuicją zależność dla najwyższych wartości. Bardzo podobną krzywą pokazano również na Wykresie 1 w pracy [2], w której po raz pierwszy przedstawiano ten model. Najbardziej podobne wykresy zależności cząstkowych wykazuje zmienna dotycząca dziennej liczby wypalanych papierosów (`cpd`). Prognoza modelu BACH wzrasta dla tych, którzy palili ponad 20 lat w bardziej radykalny sposób niż inne modele. Opisane różnice pomiędzy modelami wskazują, że wykorzystują one poszczególne zmienne w inny sposób. Porównując opisane zależności między sobą oraz z wiedzą domenową, można zauważyć, że niektóre zależności wydają się być nieintuicyjne, a nawet błędne, co sugeruje konieczność korekty tych modeli.

Lokalne metody wyjaśnienia

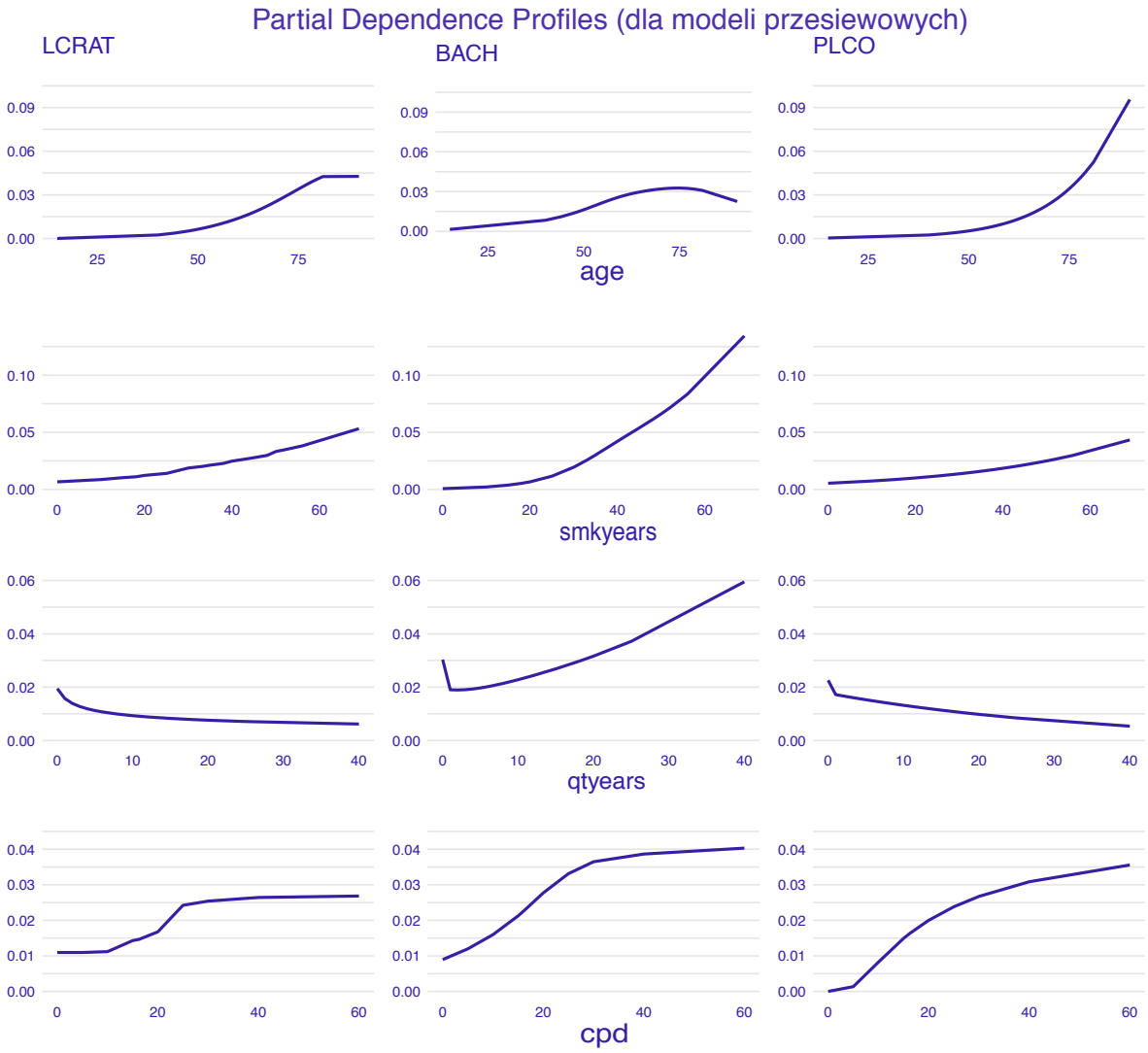
Wyjaśnienia na poziomie indywidualnym pomagają zrozumieć zachowanie modelu w kontekście pojedynczego pacjenta. Ta perspektywa jest znacznie bardziej interesująca dla pacjenta, którego mniej interesują przeciętne zachowania, a najczęściej interesuje jego własna sytuacja. Lokalne wyjaśnienia mogą wspomóc odpowiedzieć na dwa kluczowe pytania to: 1) które zmienne są najważniejsze dla konkretnego pacjenta 2) jak prognoza ryzyka dla tego pacjenta zmienia się wraz ze zmianą badanej cechy.

W celu zaprezentowania takiej lokalnej analizy, wybrałam pacjenta z największymi rozbieżnościami wśród przewidywań modeli. Analizę przeprowadziłam dla osoby z wysokim ryzykiem raka płuca według modelu PLCO_{m2012} (powyżej 10%) i niskim ryzykiem według modelu BACH (poniżej 2%, czyli poniżej progu kwalifikującego do badań przesiewowych). Model LCRAT z kolei wskazuje ryzyko tylko nieznacznie powyżej progu. Celem było przedstawienie głównych lokalnych różnic między modelami i wskazanie ich przyczyn.

Rysunek 3.9 pokazuje, które zmienne są najważniejsze dla wybranego pacjenta na podstawie dwóch metod: A.) dekompozycji break-down i B.) wartości Shapleya. Porównanie tych dwóch metodologii pokazuje, że wyniki rozkładu atrybucji metodą dekompozycji break-down i wartości Shapleya są bardzo podobne. Jedyna różnica dotyczy udziału wieku równego 90 lat dla modelu BACH. Wykres rozkładu predykcji metodą break-down wskazuje, że jego wkład jest dodatni, podczas gdy metoda wartości Sha-



Rysunek 3.7: Ważność zmiennych dla trzech analizowanych modeli. Długość słupka odpowiada za istotność zmiennej. Źródło: publikacja autora: [69].



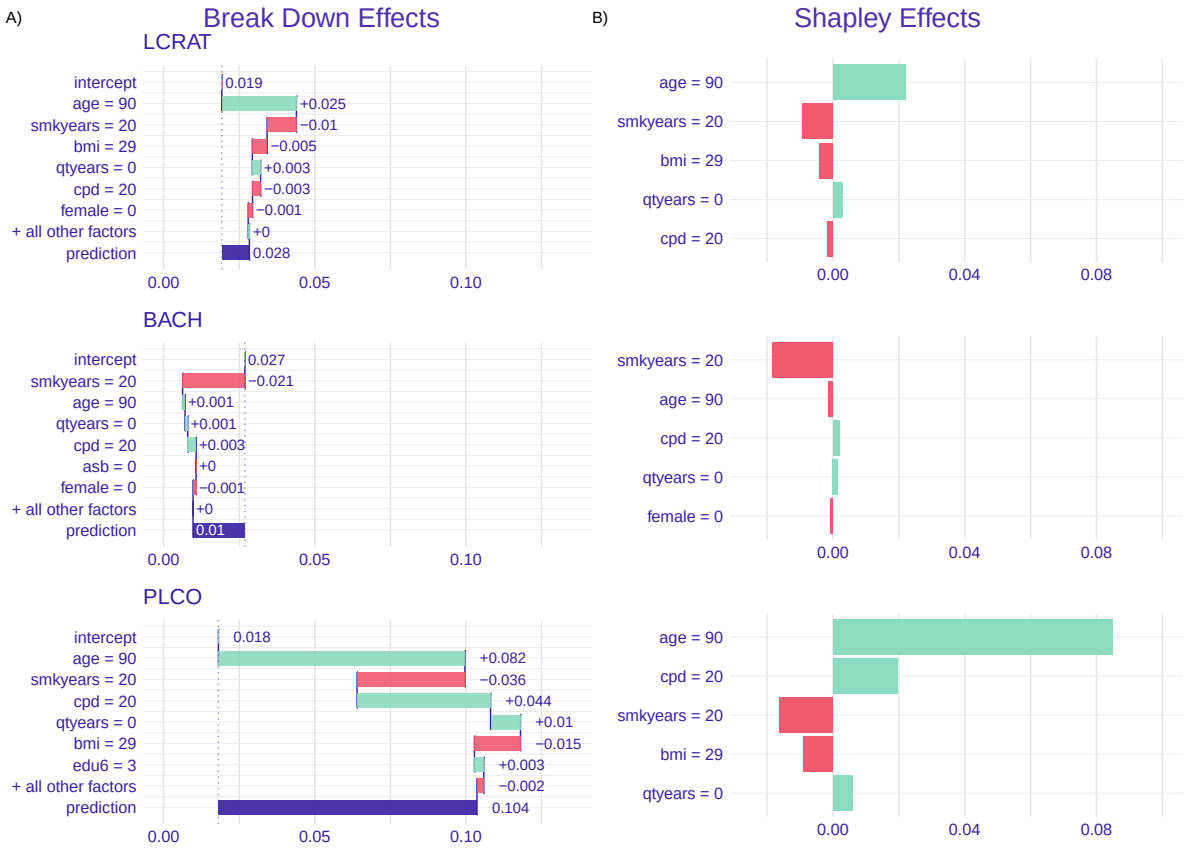
Rysunek 3.8: Wykresy PDP dla wybranych zmiennych. Na osi X, pokazane są zakresy poszczególnych zmiennych, a na osi Y wartości średniej predykcji. Źródło: publikacja autora: [69].

pleya pokazuje, że jest to wkład ujemny. Interesujące jest też porównanie trzech modeli. Jeśli chodzi o dekompozycję break-down, modele LCRAT i PLCO_{m2012} wskazują, że pacjent ma wysokie ryzyko raka płuc, podczas gdy model BACH przypisuje bardzo niskie przewidywanie (niższe niż próg 2%). Najważniejszą różnicą między modelami jest udział czasu rzucenia palenia. Tylko model PLCO_{m2012} sugeruje, że bycie aktywnym palaczem (wartość rzucenia palenia równa zeru) zwiększa ryzyko zachorowania na nowotwór. Drugą różnicą między modelami jest wpływ wieku. Wiek równy 90 lat zwiększa predykcję dla modeli PLCO_{m2012} i LCRAT, podczas gdy według modelu BACH wiek ma wkład bliski zeru. Zgodnie z modelami PLCO_{m2012} i LCRAT, wiek jest najważniejszą zmienną dla tego pacjenta. Wypalanie dziennie 20 papierosów ma przeciwne skutki: model LCRAT uznaje ją za zmienną o ujemnym wpływie, podczas gdy modele PLCO_{m2012} i BACH sugerują jej pozytywny wpływ. Warto zauważyć, że miara `BMI` ma dość istotny wpływ na rokowanie dla modeli LCRAT i PLCO_{m2012}, podczas gdy nie jest uwzględniana w modelu BACH.

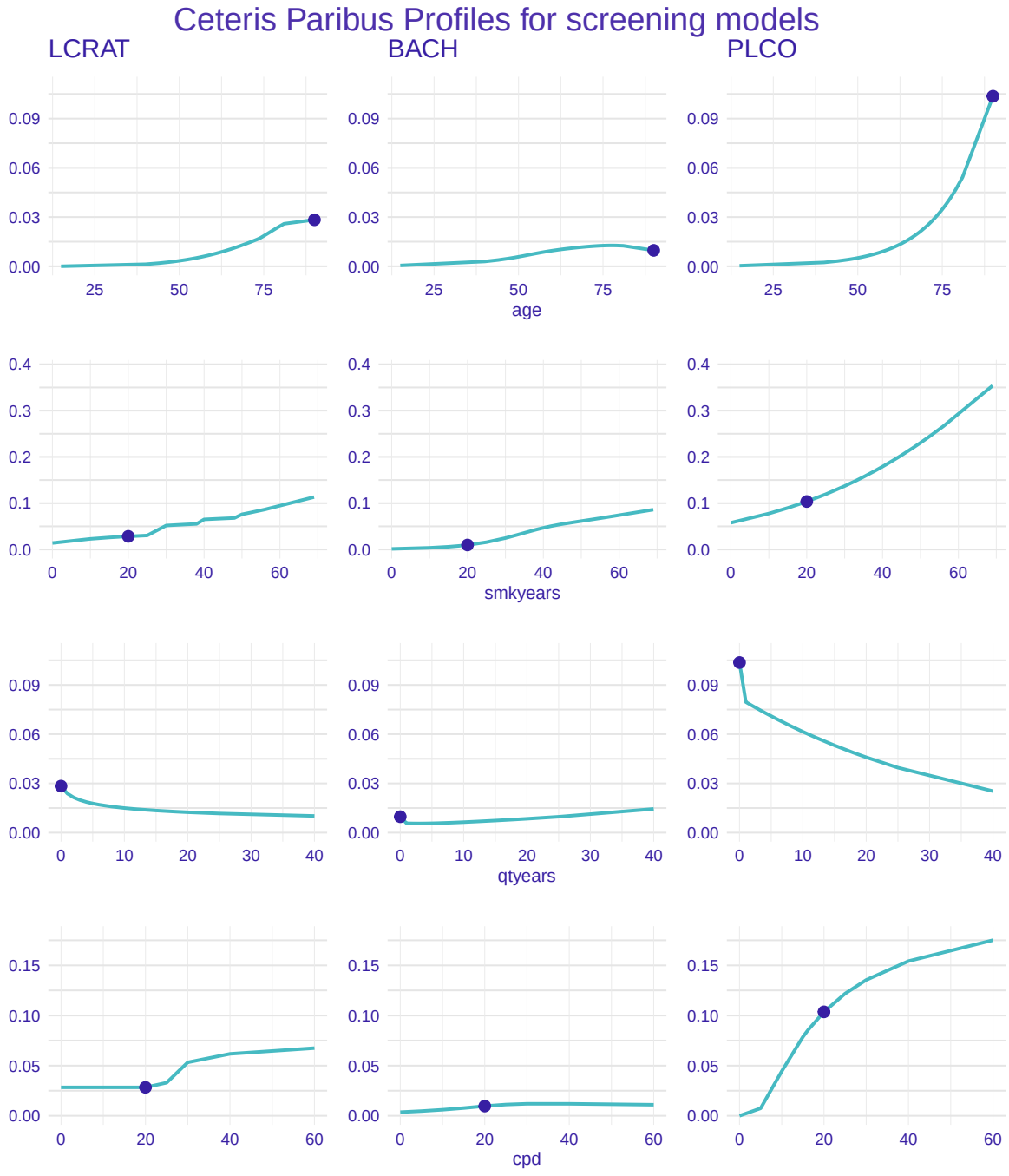
Rysunek 3.10 ilustruje, w jaki sposób prognoza ryzyka dla wybranego pacjenta zmienia się wraz ze zmianą cechy. Wykresy ceteris paribus przedstawiają zmianę predykcji w ramach zmiany zmiennej dla wybranego pacjenta. Można zaobserwować, że zmienność wieku, lat palenia, czasu rzucania palenia i ilości papierosów dziennie jest największa dla modelu PLCO_{m2012}. Na przykład, analizując zmienną wiek, gdyby pacjent był młodszy, przewidywane ryzyko zachorowania na raka byłoby znacznie niższe niż dla starszej osoby. Modele LCRAT i BACH wykazują znacznie mniejszą zmienność prognoz. Dla wybranego pacjenta model BACH nie zmieniłby bardzo prognozy, nawet gdyby pacjent rzucił palenie lub palił znacznie mniej papierosów dziennie. Należy jednak zauważyć, że lokalne metody przedstawiają zachowania modeli dla wybranej osoby, a nie dla całej kohorty. W związku z tym te wyniki można porównywać tylko między modelami dla wybranego pacjenta, a nie pomiędzy pacjentami.

3.2.6. Podsumowanie

Modele szacujące ryzyko zachorowania na raka płuca ułatwiają włączenie populacji ryzyka do procesu przesiewowego, a tym samym pośrednio zmniejszają śmiertelność z powodu jednego z najczęstszych i najbardziej agresywnych nowotworów. W tym celu coraz częściej wykorzystuje się złożone statystyczne modele ryzyka. Istnieją badania po-



Rysunek 3.9: Dekompozycja break-down (A) i wartości Shapley (B) dla wybranego pacjenta i trzech modeli przesiewowych. Wykres pokazuje wkłady poszczególnych zmiennych w końcową predykcję. A) Fioletowy słupek odpowiada za końcową predykcję modelu, zielone słupki wskazują pozytywne wkłady, czyli wartości zmiennej powodują wzrost predykcji. Analogicznie, czerwone słupki obrazują negatywne wkłady. B) Zielone i czerwone słupki wskazują wartości Shapleya prezentując odpowiednio pozytywny i negatywny wkład w końcową predykcję modelu. Źródło: publikacja autora [69].



Rysunek 3.10: Profile ceteris paribus dla trzech modeli przesiewowych. Granatowe punkty oznaczają faktyczną wartość zmiennej wybranego pacjenta, a zielone krzywe pokazują zmianę w wartościach predykcji, gdyby zmieniła się analizowana zmienna. Źródło: publikacja autora [69].

równujące modele pod kątem dokładności na poszczególnych zbiorach danych, np. [65]. W tym badaniu wypełniłam lukę związaną z interpretowalnością tych modeli. Przedstawiłam, w jaki sposób można zwalidować modele, które są powszechnie uznawane i stosowane przy pomocy metod wyjaśnialnego uczenia maszynowego.

W trakcie walidacji trzech modeli szacujących ryzyko zachorowania na raka płuca, zastosowałam zarówno globalne, jak i lokalne metody wyjaśnienia. W przypadku tego badania, to właśnie metody globalne ujawniły problemy występujące w modelach. Porównując wyniki PDP oraz weryfikując je z wiedzą medyczną, ujawnione zostały istotne błędy w procesie prognozowania. Pomimo podobnych wartości miar AUC dla wszystkich trzech modeli, dokładna analiza sposobu podejmowania decyzji przez modele pozwoliła zidentyfikować błędy w ich implementacji.

Przedstawiony przykład jest empirycznym dowodem na to, że metody XAI mogą być efektywnie wykorzystane do analizy i porównywania modeli medycznych. Metody XAI pomagają w wyszukaniu słabych stron modeli uczenia maszynowego, a także w wyborze właściwego modelu. Dzięki nim możliwe jest lepsze zrozumienie działania modeli w kontekście medycznym. Hipoteza 2 z sekcji 1.3 została zatem potwierdzona.

3.3. Metody XAI na etapie wdrożenia modelu.

Rozdział powstał na podstawie wyników opublikowanych w publikacji [75]:

Anna Lemańska-Perek, Dorota Krzyżanowska-Gołąb, Katarzyna Kobylińska, Przemysław Biecek, Tomasz Skalec, Maciej Tyszko, Waldemar Gozdzik, Barbara Adamik, "Explainable Artificial Intelligence Helps in Understanding the Effect of Fibronectin on Survival of Sepsis", Cells, Aug 5, 11(15).

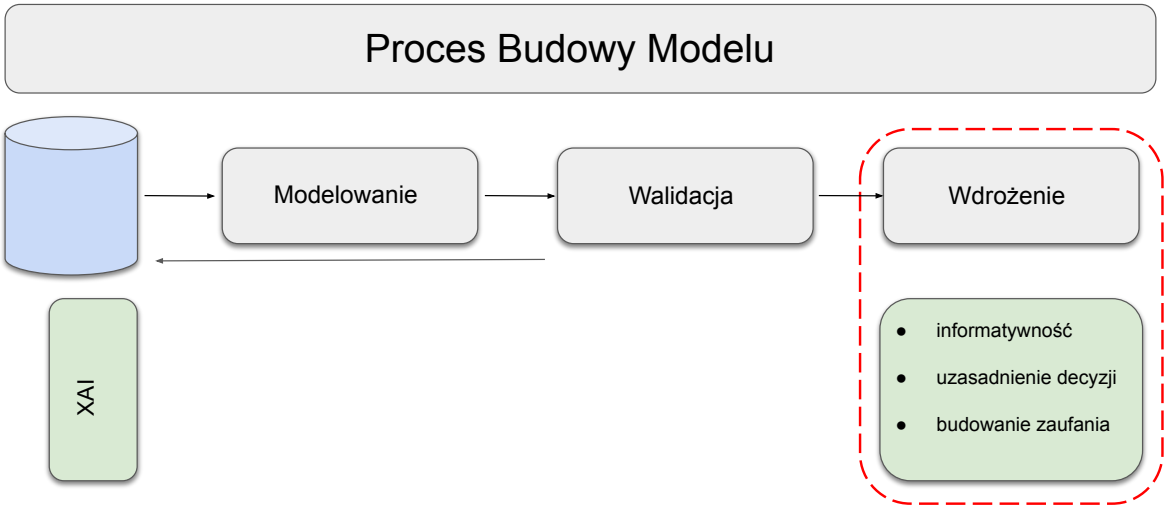
Stanowi odpowiedź na hipotezę 3 z Sekcji 1.3, która brzmi: "Metody wyjaśnialnego uczenia maszynowego post-hoc wspomagają zrozumienie i interpretację wyników modelowania. Metody te umożliwiają bardziej przejrzyste wyjaśnienie działania modeli i pomagają użytkownikom zrozumienie procesu podejmowania decyzji przez modele". W tym rozdziale przedstawiam dowody na prawdziwość tej hipotezy.

3.3.1. Wprowadzenie

Metody wyjaśnialnego uczenia maszynowego mają za zadanie wspomóc zrozumienie działania modelu. W procesie budowy modelu, można je wykorzystać także w na końcowym etapie, przy wdrożeniu modelu. Wspomogą budowanie zaufania do modelu przez osoby, które na co dzień zajmują się badanym zjawiskiem, a wyniki modelu wykorzystują w swojej pracy. W badaniu, które przeprowadziłam we współpracy z Katedrą i Kliniką Anestezjologii i Intensywnej Terapii Uniwersytetu Medycznego im. Piastów Śląskich we Wrocławiu, nowo powstały model uczenia maszynowego wsparłam o metody XAI w celu poprawy zaufania do modelu przez lekarzy.

Na przykładzie tego badania prezentuję, w jaki sposób zrozumienie modeli może przyczynić się do budowania zaufania. Ponadto pokazuję, w jaki sposób metody XAI mogą być wykorzystane do wspomagania wyjaśnialności i zrozumienia modeli dla osób związanych z medycyną, które są odpowiedzialne za korzystanie z wyników tych modeli.

W pracy pokazałam skuteczność modelu uczenia maszynowego opartego na algorytmie lasu losowego, opracowanego w celu przewidywania prawdopodobieństwa przeżycia pacjenta z sepsą przy przyjęciu na Oddział Intensywnej Terapii (OIT). Wyjaśnialne uczenie maszynowe zostało użyte na etapie wdrożenia modelu w celu poprawy zrozumienia działania modeli oraz budowania zaufania do odpowiedzi modelu. Użyta metodologia wspomaga w odpowiedzi na zaprezentowane na Rysunku 3.11 pytania. W tym



Rysunek 3.11: Metody XAI użyte na etapie wdrożenia modeli w procesie budowy modelu.

badaniu również istotna okazała się możliwość personalizacji przewidywań dla konkretnych pacjentów. Aby ułatwić wykorzystanie metod XAI dla praktyków, opracowałam i udostępniłam internetowy kalkulator ryzyka do badania indywidualnego ryzyka pacjenta. Aplikacja internetowa może być stale aktualizowana o nowe dane w celu dalszego ulepszania modelu.

3.3.2. Problem medyczny

Przewidywanie ryzyka zgonu u pacjentów OIT ma wiele zastosowań. Jest praktyczne przy planowaniu alokacji zasobów i ocenie wydajności oddziałów intensywnej terapii. Ocena ryzyka śmiertelności jest wykorzystywana w badaniach klinicznych do charakteryzowania i porównywania grup pacjentów, a także jest ważną częścią oceny jakości. Dokładna identyfikacja pacjentów z sepsą, którzy są bardziej narażeni na śmierć i którzy mogą odnieść największe korzyści z dodatkowego monitorowania lub leczenia, pozostaje wyzwaniem. Proponowanym rozwiązaniem w publikacji [93] jest dodatkowe wykorzystanie biomarkerów do identyfikacji takich pacjentów. Jednak ze względu na

heterogeniczność i złożoną patofizjologię sepsy pojedynczy biomarker często dostarcza niewystarczających informacji i nie można go wiarygodnie zakwalifikować jako czynnika prognostycznego u pacjentów z sepsą.

Modele predykcyjne oparte na sztucznej inteligencji, które okazały się przydatne w diagnostyce i prognozowaniu w innych dziedzinach medycyny [139, 82], mogą wnieść dużą wartość dodaną w tych obszarach dla pacjentów z sepsą. Powstały badania wykazujące skuteczność modeli uczenia maszynowego w wykrywaniu sepsy, wykazując poprawę we wczesnej identyfikacji pacjentów z grupy ryzyka [16, 15]. W badaniu [108] pokazano, że zastosowanie modelowania poprawiło wyniki pacjentów z sepsą, statystycznie istotne różnice stwierdzono w skróceniu długości pobytu i śmiertelności podczas pobytu w szpitalu.

Sepsa jest stanem zagrażającym życiu, spowodowanym niezrównoważoną reakcją organizmu na infekcję. Może szybko doprowadzić do niewydolności narządów i śmierci. Jest główną przyczyną zgonów na oddziałach OIT, a rokowanie pacjentów z sepsą jest często złe. Ciężkość stanu pacjenta przy przyjęciu na OIT można określić przy pomocy skal klinicznych, takich jak skala APACHE II (ang. *acute physiology and chronic health evaluation II*), NEWS (ang. *national early warning score*) oraz stopnia dysfunkcji narządu, a wyniki można codziennie oceniać przy pomocy skali SOFA (ang. *sequential organ failure assessment*), skali SAPS (ang. *simplified acute physiology*) oraz skali SSS (ang. *sepsis severity score*) [38, 111].

Fibronektyna (FN) jest białkiem występującym w tkankach organizmów i jednym z podstawowych składników macierzy pozakomórkowej, czyli struktury otaczającej komórki. Jej obecność i funkcje są istotne w wielu dziedzinach medycyny i może być wykorzystywana jako marker diagnostyczny w różnych chorobach. Odgrywa na przykład istotną rolę w odpowiedzi gospodarza na infekcję. We wcześniejszych badaniach obserwowano spadek poziomu FN w sepsie, ale nie wyjaśniono jednoznacznie, jak ten parametr wpływa na przeżycie chorego. Badanie [74] wykazało, że poziomy fibronektyny są związane z klinicznymi wskaźnikami ciężkości sepsy.

Celem modelowania w tym badaniu było znalezienie związku między FN a przeżyciem pacjentów z sepsą. Stworzono w tym celu modele uczenia maszynowego przewidujące czy pacjent przeżyje 28 dni od momentu przyjęcia na OIT i uwzględniające poziomy markerów pobieranych u pacjentów przy przyjęciu na oddział OIT.

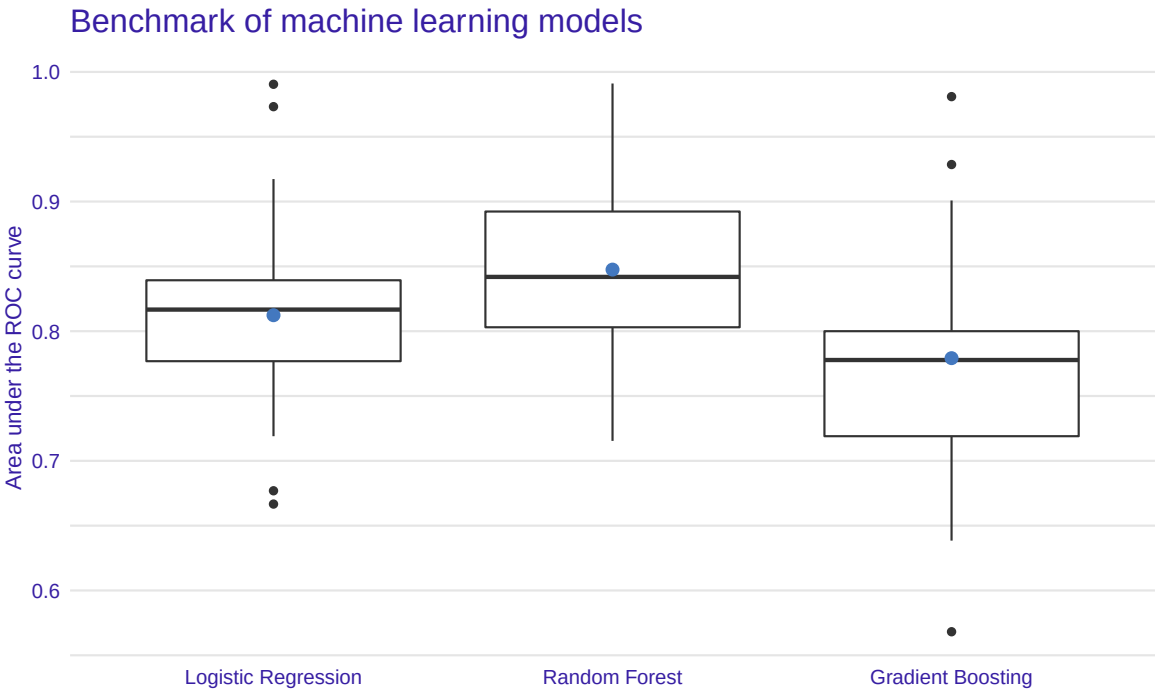
Cecha (Zmienna)	**Zmarł**	**Przeżył**	**p-wartość**
Wiek (`Age`)	71 (65-79)	64 (56-74)	**0,001**
Płeć Kobieta/Mężczyzna	27/27	31/37	0,627
skala APACHE II (`APACHE`)	28 (22-33)	20 (15-26)	**<0,001**
skala SOFA (`SOFA`)	11,5 (10-15)	9 (7-11)	**<0,001**
Procalcitonin (`PCT`)	14,57 (3,9-34,2)	4,47 (0,8-15,47)	**<0.001**
C-reactive protein, CRP	197,3 (123-307,9)	186,7 (100,4-302,5)	0,726
INR (`INR`)	1,49 (1,32-1,8)	1,2 (1,12-1,43)	**<0,001**
Liczba trombocytów (`PLT`)	138,5 (74-243)	209,5 (155-335)	**0,001**
D-dimery (`D-dimers`)	5,7 (3,97-15,59)	5,54 (3,37-11,47)	0,294
Luekocyty	15,9 (9,7-22,5)	15 (11,2-22,9)	0,66

Tabela 3.8: Statystyki opisowe zmiennych dla osób przyjętych na OIT. W pierwszej kolumnie w nawiasach podane są nazwy zmiennych użytych w modelu. Źródło: publikacja autora [75]

3.3.3. Analizowane dane

Badanie obejmowało pacjentów z sepsą przyjętych na Oddział Intensywnej Terapii w Szpitalu Uniwersyteckim. Kryteriami włączenia do badania były: ukończone 18 lat oraz rozpoznanie sepsy/wstrząsu septycznego. Natomiast kryteriami wykluczenia były: wcześniejsze leczenie na OIT, ciąża lub nieuleczalna choroba bez szans na znaczące wyzdrowienie lub przewidywany czas pobytu na OIT wynoszący 24 godziny lub mniej. Do badania włączono wszystkich pacjentów przyjętych na OIT, którzy spełnili powyższe kryteria. Zebrano z dokumentacji medycznej pacjentów dane demograficzne, laboratoryjne i kliniczne. Stan kliniczny chorego przy przyjęciu na OIT określono przy pomocy skali `APACHE II` [67]. Stopień dysfunkcji narządowej pacjentów przyjętych na OIT oceniano przy pomocą skali `SOFA`, służącej do monitorowania ciężkości sepsy. Zarówno skala `APACHE II`, jak i `SOFA` są narzędziami rutynowo używanymi. Rejestrowano również dane demograficzne i parametry laboratoryjne, takie jak liczba białych krwinek (`WBC`), poziom białka C-reaktywnego (`CRP`), poziom prokalcytoniny (`PCT`) oraz parametry krzepnięcia (`D-dimers`; `INR`). Oprócz powszechnie stosowanych wskaźników stanu klinicznego pacjentów z sepsą w modelowaniu uwzględniono stężenie fibronektyny (`pFN`) odnotowane w dniu przyjęcia na OIT.

W Tabeli 3.8 podsumowane są zmienne ciągłe przy pomocy mediany i kwartyli (0,25, 0,75), a zmienne kategoryczne przedstawione są w postaci częstości występo-



Rysunek 3.12: Porównanie wyników miary AUC po 5-krotnej walidacji krzyżowej dla trzech typów modeli. Źródło: publikacja autora [75].

wania. Porównanie zmiennych ciągłych między dwiema niezależnymi grupami (osoby, które nie przeżyły do osób, które przeżyły) przeprowadzone jest przy pomocy testu Manna-Whitneya, a zmienne kategoryczne przy pomocy testu chi-kwadrat.

3.3.4. Modelowanie

Opracowałam trzy typy modeli: model regresji logistycznej oraz dwa złożone modele oparte na drzewach: las losowy i wzmocnienie gradientowe. Jakość modeli została oceniona na podstawie 5-krotnej walidacji krzyżowej. Średnie testowe AUC wyniosło $0,85$ dla modelu lasów losowych, $0,78$ dla modelu gradient boosting i $0,81$ dla modelu regresji logistycznej. Wyniki testowych AUC dla modeli przedstawione są na Rysunku 3.12.

Najlepsze wyniki uzyskał model lasu losowego i model ten został wdrożony jako końcowy model przewidujący predykcję dla pacjentów z sepsą. Zmienne objaśniające,

które wprowadzono do modelu, wybrano na podstawie testów istotności przedstawionych w tabeli 3.8 oraz na podstawie wcześniejszych wyników uzyskanych dla fibronektyny [74]. Dodatkowo do analizy włączono zmienną `D-dimers`, wskazaną przez lekarzy jako istotny parametr wskazujący na degradację fibryny. Ostatecznie model uczenia maszynowego został opracowany z uwzględnieniem wejściowych cech stężenia fibronektyny w osoczu, wartości INR, wyniku SOFA, wieku pacjenta, wyniku APACHE II, poziomu prokalcytoniny, liczby płytek krwi oraz poziomu d-dimerów. Przeprowadzono 10-krotną walidację krzyżową w celu optymalizacji parametrów modelu lasu losowego i uniknięcia przetrenowania modelu. Średnia wartość AUC 10-krotnej walidacji krzyżowej obliczona dla zestawów danych testowych wynosiła $0,82$. Ostateczny model przygotowano na zbiorze danych treningowych. Analiza krzywej ROC modelu lasu losowego wykazała, że wskaźnik udanej klasyfikacji pacjentów za pomocą modelu wyniósł $0,92$ (AUC obliczone dla całego zbioru danych), z czułością $0,92$, wartością precyzji równą $0,76$ i dokładnością równą $0,79$.

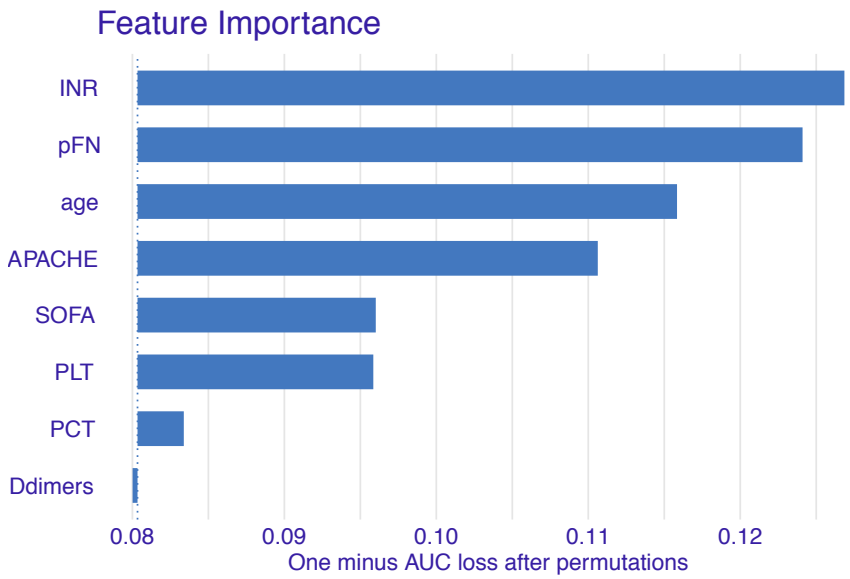
3.3.5. Wdrożenie modelu uczenia maszynowego

Przy podejmowaniu decyzji obarczonych dużym ryzykiem kluczowe jest zaufanie do modelu. Zrozumienie struktury modelu i zawartych w nim zależności oraz porównanie z wiedzą ekspercką może wspomóc budowanie zaufania. W celu wyjaśnienia działania modeli, wykorzystałam metody XAI na poziomie globalnym i metody XAI na poziomie lokalnym.

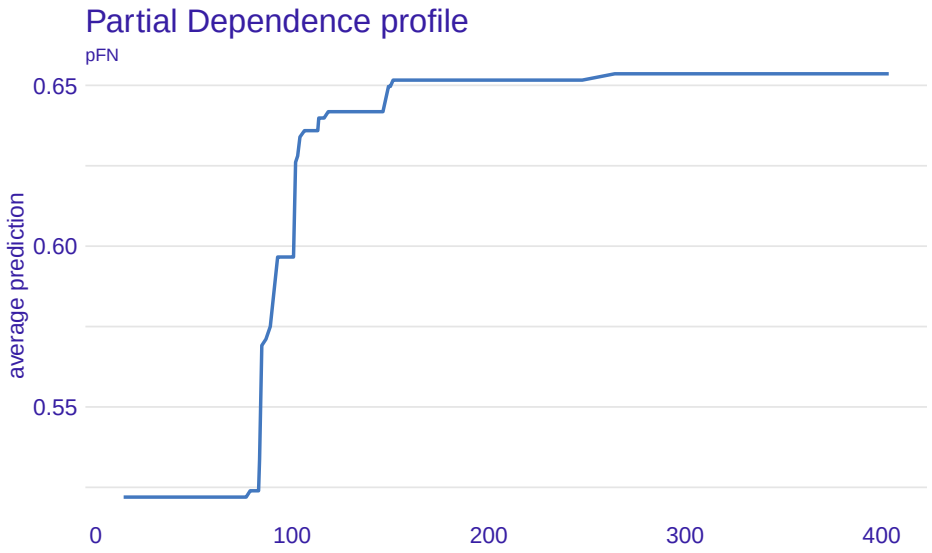
Globalne metody wyjaśnienia

Ważność zmiennych w modelu lasu losowego przedstawiona jest na Rysunku 3.13. Zgodnie z wykresem, wartość `INR` i stężenie `pFN` były najważniejszymi zmiennymi w modelu lasu losowego. Kolejne zmienne to `Age` i `APACHE II`, które również są uznane przez model za istotne. Wydaje się, że poziom zmiennej `D-dimers` nie miał wpływu na przewidywania modelu. Potwierdziły się przewidywania lekarzy, że `pFN` odgrywa istotną rolę w rokowaniu pacjentów z sepsą.

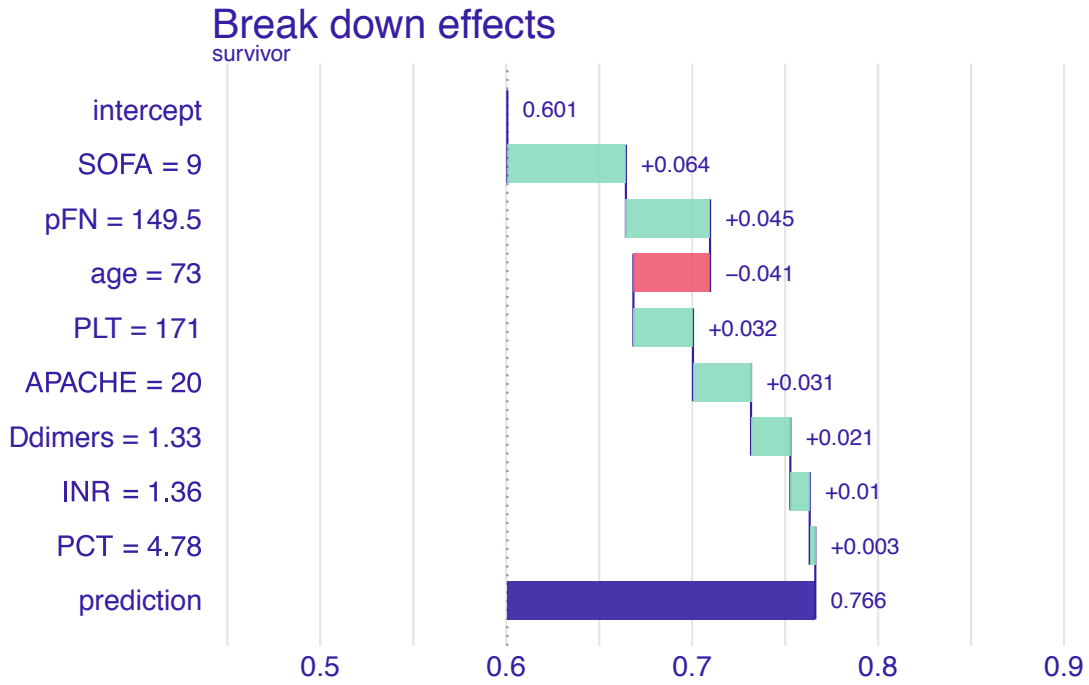
Wykres 3.14 pokazuje metodę PDP dla analizowanej zmiennej. Pokazuje punkt odcięcia (`pFN` w okolicy 105), dla którego wartości predykcji rosną. Jest to potwierdzenie wcześniejszych badań i obserwacji na oddziałach OIT, że pacjenci z wyższą wartością tego białka mają lepsze rokowania.



Rysunek 3.13: Ważność zmiennych w modelu lasu losowego. Źródło: publikacja autora [75].



Rysunek 3.14: Wykres PDP dla zmiennej pFN w modelu lasu losowego. Wykres wskazuje na próg zmiennej powyżej którego predykcja mocno wzrasta. Źródło: publikacja autora [75].



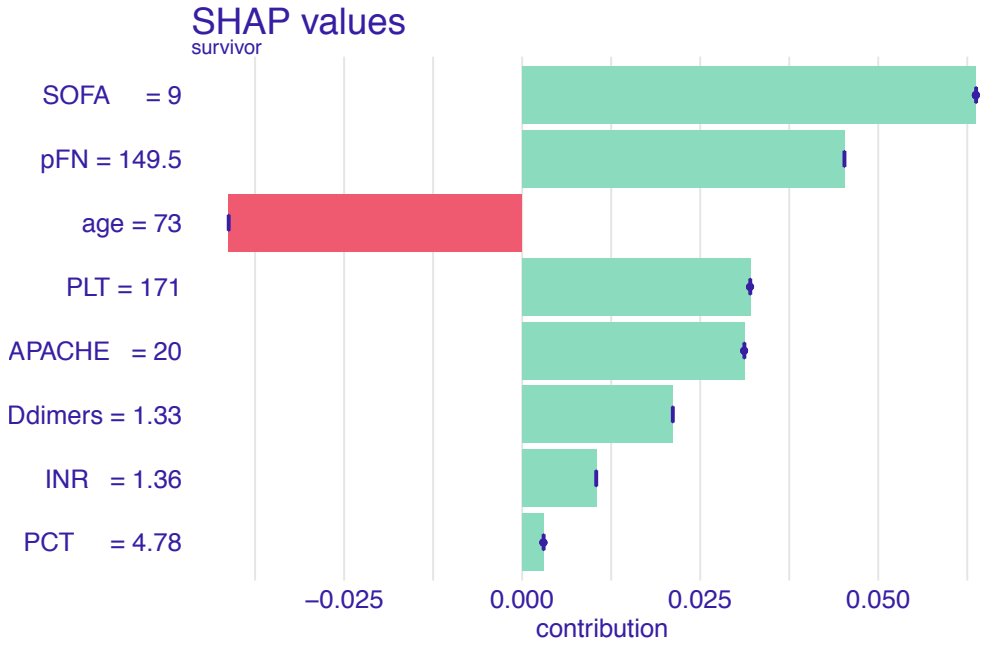
Rysunek 3.15: Dekompozycja break-down pokazuje wpływ poszczególnych zmiennych na ostateczną predykcję dla wybranego pacjenta. Źródło: publikacja autora [75].

Lokalne metody wyjaśnienia

Lokalna perspektywa wyjaśniania modelu ma na celu wsparcie w podejmowaniu decyzji medycznych dla indywidualnego pacjenta. Metody te są pomocne w zrozumieniu, które zmienne są najważniejsze dla wybranego pacjenta i jak wpływają na wynik modelu. Wyniki potwierdziły intuicję medyczną i wiedzę domenową lekarzy. W celu łatwiejszego dostępu do metod XAI i wyników modelu, stworzyłam aplikację dostępną online, która zawiera wszystkie wyżej wymienione metody wyjaśnień dla opracowanego modelu. Aplikacja ta ułatwia zapoznanie się z metodami i stosowanie ich przez lekarzy, stwarza łatwą dostępność do modelu i jego wyjaśnień. Umożliwia stosowanie lokalnych wyjaśnień dla nowych pacjentów w codziennej pracy. Aplikacja jest dostępna pod adresem https://stats4med.shinyapps.io/xai2shiny/, a przedstawiona poniżej analiza jest przykładem wykorzystania modelu dla wybranego pacjenta.

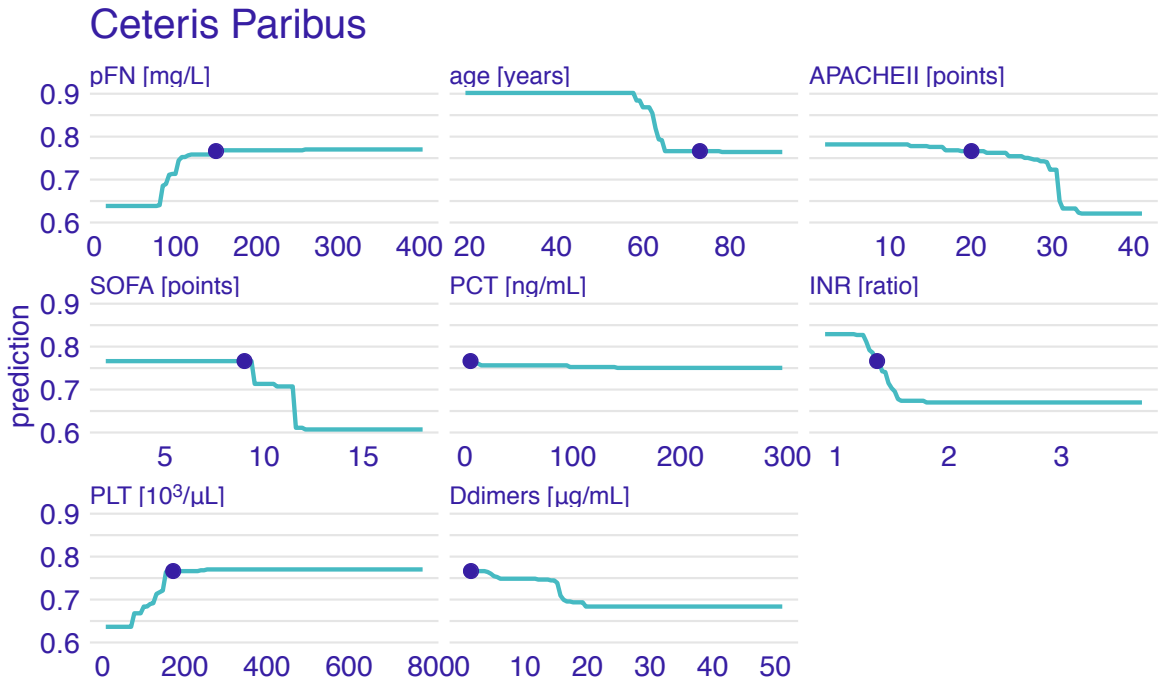
Analiza przypadku

Osoba chora została przyjęta na OIT bezpośrednio z bloku operacyjnego po re-



Rysunek 3.16: Metoda wartości Shapleya pokazuje wpływ poszczególnych zmiennych na ostateczną predykcję dla wybranego pacjenta. Źródło: publikacja autora [75].

laparotomii z powodu perforacji jelita ślepego i zapalenia otrzewnej. Przy przyjęciu rozpoznano wstrząs septyczny. Z powodu niewydolności oddechowej chorego wentylowano mechanicznie z zastosowaniem wysokiego ciśnienia i wspomagania tlenowego. Wdrożono antybiotykoterapię oraz wprowadzono żywienie pozajelitowe. Stan kliniczny pacjenta oceniano w skali `APACHE II` na 20 pkt. oraz w skali `SOFA` na 9 pkt. Pozostałe parametry zastosowane w modelu predykcyjnym to poziom fibronektyny w osoczu `pFN` równy 149,51mg/L, poziom `PCT` równy 4,78 ng/L, poziom `D-dimers` równy 1,33 mg/L, poziom `INR` równy 1,36. Zgodnie z przedstawionym modelem prognoza 28-dniowego przeżycia obliczona dla tej pacjentki w chwili przyjęcia na OIT wyniosła 0,764 i była wyższa od średniej prognozy modelu. W oparciu o dekompozycję break-down najważniejszą zmienną w modelu, która zwiększała predykcję, był wynik SOFA z udziałem +0,064 oraz poziom (`pFN`) z udziałem +0,45 3.15. Pozostałe zmienne miały mniejsze znaczenie, a jedyną zmienną, która miała negatywny wpływ na predykcję był wiek. Bardzo podobne wyniki uzyskano dla wartości SHAP 3.16. Co więcej, biorąc pod uwagę profile ceteris paribus, możemy zauważyć, że obniżenie poziomu `pFN` lub nawet niewielki wzrost poziomu SOFA skutkowałby gorszą prognozą dla tego pacjenta 3.17. Zgodnie



Rysunek 3.17: Profile ceteris paribus prezentują zależności między zmiennymi i predykcjami dla wybranego pacjenta. Źródło: publikacja autora [75].

z dokumentacją szpitalną pacjent żył w 28. dobie leczenia.

3.3.6. Podsumowanie

W badaniu wykorzystany został model uczenia maszynowego z fibronektyną jako zmienną objaśniającą, nowym potencjalnym biomarkerem sepsy, w połączeniu z rutynowo mierzonymi wskaźnikami życiowymi. Wyniki wykazały, że najważniejszymi wskaźnikami przewidywania przeżycia były `INR` i poziom fibronektyny (`pFN`) w osoczu, a następnie wiek (`Age`) i skala `APACHE II`. Aby lepiej zrozumieć związek między `pFN` a przeżyciem pacjentów z sepsą, wykorzystałam wyjaśnialne uczenie maszynowe. Wyjaśnienia lokalne pomogły zrozumieć wkład `pFN` w przewidywanie przeżycia na poziomie pojedynczego pacjenta. Metodologia daje nowe możliwości personalizacji przewidywań dla pacjentów. Na podstawie danych z analizowanej kohorty opracowałam internetową aplikację do przewidywania przeżycia poszczególnych pacjentów. Aplikacja ułatwiła lekarzom za-

poznanie się z modelem i jego wyjaśnieniami, a także korzystanie z modelu w codziennej pracy.

Fibronektyna była już we wcześniejszych badaniach proponowana jako biomarker sepsy, kiedy zaobserwowano, że niskie stężenia FN w osoczu pacjentów z podejrzeniem sepsy były zgodne z ostatecznym rozpoznaniem sepsy [76]. Obniżone stężenie pFN w osoczu obserwowano również w ostrych stanach zapalnych, urazach chirurgicznych i rozsianym wykrzepianiu wewnątrznaczyniowym [101, 96]. Wyjaśnienia modelu lasu losowego potwierdziły intuicję i wiedzą domenową, co pozwoliło zaufać predykcjom modelu, a także stosować model w praktyce lekarskiej na oddziale OIT.

Metody XAI pomogły w zrozumieniu dostarczonego modelu, ułatwiły współpracę między lekarzami a modelem, umożliwiając lekarzom lepsze zrozumienie wniosków płynących z modelu oraz wykorzystanie ich w praktyce medycznej. Dzięki metodom XAI model ten może być stosowany na Oddziale Szpitala Klinicznego. Hipoteza 3 z sekcji 1.3 została zatem empirycznie potwierdzona.

Rozdział 4

Rozszerzenie procesu budowy modelu przy użyciu metod XAI

Rozdział powstał na podstawie publikacji [70]:

Katarzyna Kobylińska, Mateusz Krzyziński, Rafał Machowicz, Mariusz Adamek, Przemysław Biecek, "Exploration of the Rashomon Set Assists Trustworthy Explanations for Medical Data", arXiv:2308.11446.

Stanowi odpowiedź na hipotezę 4 z Sekcji 1.3, która brzmi: "Na podstawie metod wyjaśnialnego uczenia maszynowego post-hoc opracowałam metodę automatycznego wyboru podzbioru najbardziej różnych modeli spośród wielu, dobrych modeli". W tym rozdziale przedstawiam dowody na prawdziwość tej hipotezy.

4.1. Wprowadzenie

W celu poparcia przedstawionej hipotezy, prezentuję nowy algorytm, którego zadaniem jest wybór najbardziej różniących się modeli spośród zbioru niemal optymalnych modeli. Ten algorytm jest rozszerzeniem standardowego procesu budowy modeli uczenia maszynowego. Pozwala na weryfikację i porównanie różnych modeli, które są zbliżone pod względem jakości, a mogą być odległe pod względem wyjaśnienia predykcji.

Wprowadzam także nową miarę, która służy do porównywania modeli, opartą o metody XAI. Ta miara umożliwia porównanie różnic między modelami na podstawie tego w jaki sposób modele traktują zależności między zmiennymi a predykcją. Przedstawiam

szczegółowy opis miary, wraz z możliwościami zastosowania w medycynie. Porównuję ją także z innymi miarami służącymi do pokazania odległości na danych symulacyjnych oraz na danych rzeczywistych.

Prezentowana miara oraz algorytm mają na celu dostarczenie wsparcia badaczom i praktykom zajmującym się budowaniem modeli w dziedzinie medycyny. Ich wprowadzenie umożliwia lepsze zrozumienie i ocenę wielu modeli pod kątem ich interpretowalności oraz wybór najlepszego modelu spośród dostępnych. Przedstawione rozwiązania mają znaczenie dla poprawy jakości modelowania medycznego i umożliwiają bardziej świadome podejmowanie decyzji opartych na wynikach tych modeli.

Wnioski wyprowadzone z prezentowanej miary i algorytmu stanowią istotny wkład w dziedzinę modelowania medycznego i otwierają nowe perspektywy badawcze. Poparcie przedstawionej hipotezy za pomocą nowej miary oraz algorytmu stanowi ważny krok w rozwoju i doskonaleniu modelowania medycznego opartego na metodach XAI. Przedstawione rozwiązania przyczyniają się do poszerzenia wiedzy na temat interpretowalności modeli, co ma istotne implikacje dla praktyki klinicznej i podejmowania decyzji opartych na wynikach modeli predykcyjnych.

W sekcji 3.1 pokazałam w jaki sposób metody XAI są pomocne w trakcie modelowania problemu medycznego na podstawie dużego, rzeczywistego zbioru danych medycznych. Zastosowałam w tym celu standardowy proces wyboru modelu predykcyjnego, a na podstawie modelu czarnej skrzynki i porównania profili PDP, ulepszyłam interpretowalny model regresji logistycznej. Przykład ten pokazał, że mogą istnieć modele, które pomimo zbliżonej dokładności posiadają zupełnie inne wyjaśnienie predykcji i traktują poszczególne zmienne w inny sposób, co prezentuje wykres 3.3.

W sekcji 3.2 pokazałam, że metody XAI mogą służyć do porównania modeli, które są już wdrożone. Okazało się, że pomimo tego, że modele mają podobną dokładność (mierzoną miarą AUC) oraz są stosowane w różnych ośrodkach medycznych, metody XAI mogą służyć do ich walidacji, wyszczególnienia bądź zauważenia błędów i nieścisłości w wielu aspektach modelowania, co pokazane jest na wykresie 3.10. W obu badaniach szczególnie pomocnymi metodami okazały się profile przedstawiające zależności między poszczególnymi zmiennymi ciągłymi a odpowiedziami modelu, takie jak PDP czy ceteris paribus.

W sekcji 3.3 pokazałam, że zaawansowane modele uczenia maszynowego z użyciem metod XAI, a także aplikacji umożliwiającej korzystanie z modelu przez praktyków

mogą z powodzeniem być używane w dużym ośrodku medycznym. Aplikacja i metody XAI ułatwiły zrozumienie decyzji modelu co pomogło w budowaniu zaufania do wyników. Metoda przedstawienia profili PDP okazała się najbardziej przydatną techniką opisującą wyniki modelowania wskazując lekarzom punkt odcięcia zmiennej, przy którym rokowania pacjenta są lepsze.

Te trzy studia przypadku wyłoniły kluczową rolę stosowania profili przedstawiających związki między zmiennymi a predykcją. Ponadto analizy te ilustrują istnienie bardzo podobnych pod względem dokładności modeli (niemal optymalnych), których analiza okazuje się być niezmiernie istotna i przynosząca znaczący wkład. Te rozważania skłoniły mnie do zainteresowania się modyfikacją procedury poszukiwania optymalnego modelu predykcyjnego oraz wykorzystaniem metod XAI do jego doskonalenia. W przypadku każdego z przedstawionych w poprzednich rozdziałach problemów medycznych, udawało mi się znaleźć modele różnych klas, które pod względem jakości charakteryzowały się bardzo zbliżonymi miarami. Analiza przy pomocy metod XAI pozwoliła na wyłonienie tych, które najlepiej opisywały badane zjawisko. Postawiłam sobie zatem dwa pomocnicze pytania badawcze. Po pierwsze, w jaki sposób opisać różnice między modelami, które pod względem jakości są do siebie bardzo zbliżone. Po drugie, czy proces oceny różnic między modelami przy pomocy metod XAI można zautomatyzować.

4.1.1. Przegląd literatury

Procedurę budowy modelu proponuję wzmocnić o badanie wielu modeli ze zbioru niemal optymalnych, czyli zbliżonych pod względem jakości modeli. Koncepcja takiego zbioru znana jest jako zbiór Rashomon i została po raz pierwszy wprowadzona przez Leo Breimana [13]. Zademonstrował ten problem na przykładzie modeli regresji wykorzystujących 5 zmiennych ze zbioru 30 dostępnych. Kilka modeli miało podobną wartość miary służącej do mierzenia błędu RSS (ang. *residual sum of squares*), a w każdym z nich stosowano inny ich podzbiór. Breiman wskazał technikę baggingu jako sposób radzenia sobie z różnorodnością modeli w zbiorze Rashomona. Znalezienie zbioru wielu niemal optymalnych modeli może poprawić wydajność predykcyjną co zostało poźniej pokazane w pracy opisującej modele zespołowe [17].

Celem modelowania, oprócz znalezienia modelu o wysokiej jakości, może być też chęć wyciągania wniosków na temat modelownego zjawiska. Dlatego korzystne może

być przeanalizowanie całego zbioru Rashomona. Daje to pełniejszy obraz badanego zjawiska, zwłaszcza gdy modele w inny sposób opisują badane zjawisko.

Idea zbioru Rashomon, była analizowana przez wielu badaczy z różnych perspektyw. Podejście oparte na statystycznej teorii uczenia się prezentuje artykuł opublikowany przez Lesię Semenovą wraz ze współautorami [105]. Autorzy badają kwestię istnienia prostych, ale dokładnych modeli, prezentując miarę wielkości tego zbioru (ang. *Rashomon set's volume*). Jest to proporcja liczby modeli w zbiorze do wielkości przestrzeni hipotez i może być wykorzystana do sprawdzenia, jak złożony jest badany problem. Daniel Nevo [90] rozważa minimalną klasę modeli o podobnej dokładności przewidywania, ale niekoniecznie podobną pod względem struktury. Theja Tulabandhula i Cynthia Rudin[126] używają zbiorów Rashomona do przygotowywania solidnych problemów optymalizacyjnych wspomagających podejmowanie decyzji na złożonych danych. Hsiang Hsu i Flavio Calmon [60] wprowadzają metrykę mierzącą sprzeczne predykcje dla poszczególnych obserwacji w zadaniu klasyfikacji. Masaki Hamamoto [55] wdraża metodologię wykorzystania efektu Rashomona w statystykach w celu skorygowania wyjaśnień. Nicholas Kissel i Lucas Mentch [66] proponują wybór ścieżki modelu, efektywną metodę opartą na selekcji w przód w celu znalezienia dokładnych modeli ze zbioru Rashomona.

W kilku badaniach bada się również przecięcie zbiorów Rashomona z metodami wyjaśnialnej sztucznej inteligencji. Przede wszystkim preferowane są metody niezależne od modelu, mające zastosowanie globalne, ponieważ zapewniają wgląd w cały proces wnioskowania modelu bez zakładania czegokolwiek na temat jego struktury. Jedną z takich metod jest permutacyjna ważność zmiennych, początkowo zaproponowana dla lasów losowych [14], a następnie zastosowana przez Fisher i Rudin [42] w badaniu dotyczącym istotności zmiennych w zbiorze Rashomon (ang. *model class reliance, MCR*). Autorzy obliczyli zakres ocen ważności zmiennych dla modeli ze zbioru Rashomona. MCR może dać pełniejszy opis istotności zmiennej niż PFI, gdyż wykorzystuje wiele dobrych model a nie tylko jeden model.

Powstały dalsze rozszerzenia koncepcji MCR. Jiayun Dong i Cynthia Rudin [28] wprowadzili idee ważności zmiennych dla wielu dobrych modeli z tej samej klasy. Gavin Smith [114] z kolei zaproponował w jaki sposób obliczać MCR dla klasy modeli lasu losowego. Riu Xin [137] zbadał zbiór rzadkich drzew decyzyjnych Rashomona, zaproponował technikę wyliczania takiego zbioru i zbadał ważności zmiennych przy pomocy

MCR. Jednak pomimo tak wielu rozszerzeń, MCR nie został wykorzystany do wyznaczenia podzbioru najbardziej zróżnicowanych modeli. Podkreśla to wartość badania alternatywnych perspektyw w celu charakteryzowania modeli ze zbioru Rashomona.

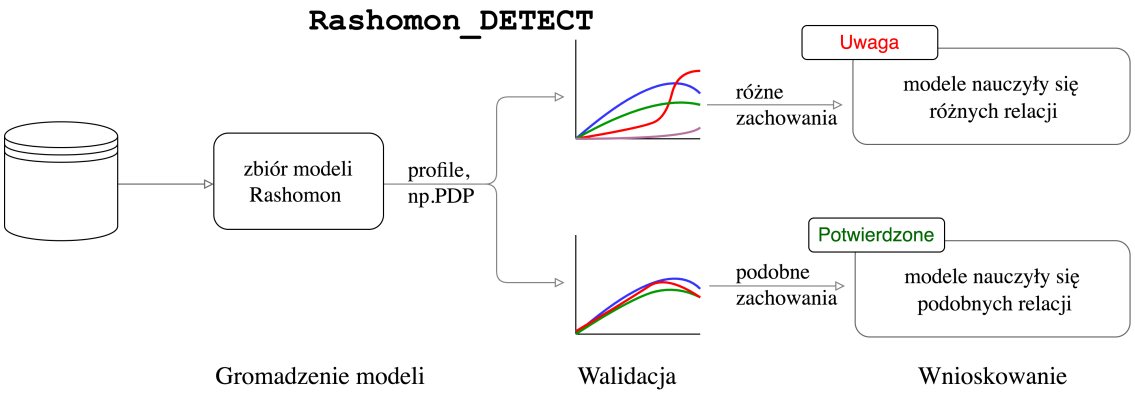
Odpowiedź na pytanie o ważność zmiennej prowadzi do zbadania, w jaki sposób model wykorzystuje poszczególne zmienne [3]. W tym celu można zastosować techniki opisujące zależność pomiędzy przewidywaniami modelu a wartościami zmiennych (profilami). Najbardziej odpowiednimi przykładami są wykresy częściowej zależności (PDP) [43] i skumulowane efekty lokalne (ALE) [1]. Pomimo ich popularności metody te nie zostały jeszcze zbadane w kontekście zbiorów Rashomona.

Charakterystyka zbioru Rashomona pozostaje jednak nadal otwartym problemem [99], podobnie jak towarzyszące mu pytanie: „Jaki model wybrać ze zbioru Rashomona?.

Proponuję, by w badaniach nad modelami predykcyjnymi rozważać nie jeden, najlepszy model, lecz cały zbiór modeli. Zbiór Rashomon może być bardzo licznym zbiorem [105], co może utrudnić dokładną analizę wszystkich modeli. Istotnym wyzwaniem dotyczącym zbioru Rashomona jest określenie najodpowiedniejszego modelu ze zbioru dokładnych modeli o różnych właściwościach. Dlatego proponuję metodę na automatyczne porównywanie modeli ze zbioru Rashomon w celu wyboru podzbioru najbardziej różnych modeli. Wybór odległych modeli ma na celu ułatwienie ich analizy i porównania. Celem tego algorytmu jest także weryfikacja czy wszystkie modele ze zbioru Rashomon w podobny sposób modelują dane. Takie podejście zapewnia większą wiarygodność do analizy danych poprzez szerszą eksplorację wielu modeli, zamiast polegania wyłącznie na jednym modelu.

Proponuję algorytm `Rashomon_DETECT` w celu automatycznego znalezienia podzbioru najbardziej zróżnicowanych modeli na podstawie dobrze znanych w dziedzinie XAI profili. Ponadto przedstawiam nową miarę służącą do porównania profili - Partial Disparity Index (PDI). Na diagramie 4.1 przedstawiam obrazowo proponowany proces, który opisuję w dalszej części tej pracy.

W kontekście medycyny, wiele badań wykorzystujących uczenie maszynowe nadal priorytetowo traktuje jeden model optymalizujący wybraną miarę jakości [34]. Chociaż Alex John London [78] twierdzi, że rozbieżności między modelami mogą być tolerowane, jeśli zostaną solidnie sprawdzone klinicznie, to należy wskazać, że w wielu opublikowanych badaniach dotyczących modeli predykcyjnych zastosowanych do danych medycznych występuje duże ryzyko obciążenia (ang. *bias*) [135]. Dlatego interpretacja



Rysunek 4.1: Proces porównania modeli na podstawie ich zachowań za pomocą algorytmu `Rashomon_DETECT`. Pole 'uwaga' oznacza potencjalnie niestabilne modele, które potrzebują dalszej ewaluacji i porównania z wiedzą domenową. Pole 'potwierdzone' oznacza modele z podobnymi zależnościami. Metodę tę można wykorzystać do znalezienia wiarygodnego modelu dla badanego zjawiska.

modelu jest niezbędna [138], a metody mające na celu eksplorację i wyjaśnienie modelu powinny być stosowane w procesie modelowania. Cynthia Rudin [100] wskazała, że analiza zbioru Rashomona jest ważna w kontekście zastosowań metod uczenia maszynowego. Określiła, że umiejętność znalezienia zbiorów Rashomona i wyświetlania ich użytkownikom rozwiązuje prawdopodobnie najtrudniejszy otwarty problem wyjaśnialnego uczenia maszynowego – interakcję międzyludzką. Rozwiązanie tego problemu może mieć ogromne znaczenie czy modele uczenia maszynowego będą mogły być stosowane przy podejmowaniu decyzji o dużej stawce. Zaproponowany przeze mnie algorytm `Rashomon_DETECT` jest krokiem w kierunku uproszczenia analizy modeli ze zbioru Rashomona.

4.2. Proces budowy modelu

Standardowy proces modelowania uczenia maszynowego używany był przy modelowaniu problemów medycznych w rozdziałach 3.1 i 3.3. Taki proces skutkuje wybraniem jednego modelu, który najlepiej spełnia wybrane kryterium jakościowe. Niestety proces ten, choć powszechnie stosowany przy analizach zbiorów medycznych, prowadzi do rezygnacji z głębszej analizy modeli nieco gorszych od optymalnego. W analizach

z rozdziałów 3.1, 3.2 pokazałam, że warto jest porównać więcej modeli przy pomocy metod XAI, gdyż nie zawsze model optymalizujący funkcję straty przedstawia najlepiej zależności w danych. W przypadku złożonych zależności pojawiających się w modelowanych zjawiskach, wnioskowanie oparte tylko na jednym analizowanym modelu może prowadzić do błędnych decyzji lub niepełnych wniosków. Ten problem jest szczególnie istotny, gdy mamy do czynienia ze zbiorem modeli, które są prawie tej samej jakości, ale znacznie różnią się sposobem, w jaki opisują dane. Zbiór ten może być bardzo liczny i wymagać bardziej kompleksowej analizy. To skłania do zaproponowania nowego sposobu postępowania przy wyborze modelu predykcyjnego. Proponuję poszerzenie procedury o uwzględnienie i analizowanie modeli ze zbioru Rashomon zamiast tylko jednego modelu. W rezultacie tak przeprowadzonej analizy, dostarczony model może być jednym z modeli ze zbioru Rashomon, niekoniecznie tym, który optymalizuje funkcję straty. Widzę jednak również możliwość dostarczenia np. zestawu najbardziej różnych modeli, które dają inne spojrzenia na badane zjawisko. Proponuję poszerzenie procesu wyboru modelu, który pozwala na uzyskanie szerszego spojrzenia na badany problem i wyjaśnienie zależności w danych. W tym celu definiuję nową miarę służącą do porównania modeli uczenia maszynowego oraz proponuję nowy proces wyboru modelu. Zbiór modeli Rashomon może być bardzo liczny, co utrudnia analizę porównawczą wszystkich modeli. Dlatego proponuję algorytm znajdujący najbardziej różne modele z tego zbioru, które warto przeanalizować i porównać.

Takie podejście poprawia proces wyboru modelu, szczególnie w kontekście zastosowań medycznych i podejmowania decyzji w opiece zdrowotnej. Rozszerzenie procesu poprzez uwzględnienie zestawu dobrych modeli sprawdzając i porównując k najbardziej odrębnych modeli z wygenerowanego zbioru, może doprowadzić do bardziej szczegółowych i wiarygodnych wniosków.

Gdy najbardziej zróżnicowane modele nie różnią się znacząco w wyjaśnieniach danych, wynik ten może zwiększyć zaufanie do modeli. Jeżeli z kolei widoczne są oczywiste rozbieżności, wybrany zostanie model, który nie musi być modelem optymalizującym funkcję straty. W zależności od scenariusza klinicznego i zaobserwowanych cech mogą pojawić się alternatywne kryteria wyboru. Mogą one obejmować modele, które są bardziej zgodne z wiedzą dziedzinową, wyróżniają się w leczeniu nietypowych pacjentów (wartości odstające) lub wykazują zwiększoną stabilność i solidność. Ponadto dostrzegam potencjał dostarczenia zbioru modeli, z których każdy prezentuje

odmienne perspektywy na analizowane zjawisko. Taki zestaw modeli może pomóc klinicystom w wyborze modelu najodpowiedniejszego dla danego pacjenta, dopasowując podejście medyczne do indywidualnych potrzeb.

4.3. Definicje

W celu przedstawienia proponowanego algorytmu przedstawiam następującą notację. Niech $\mathbf{M} = [\mathbf{X}\ \mathbf{Y}]$ będzie macierzą składającą się z n obserwacji $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ i p niezależnych zmiennych ($\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}] \in \mathcal{X}$), oraz zmiennej zależnej w postaci wektora: $\mathbf{Y} \in \mathcal{Y}$.

Przestrzeń modeli predykcyjnych jest postaci: $\mathcal{F} = \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$, gdzie f to funkcje mierzalne. Niech przestrzeń $\mathcal{L} : (\mathcal{F} \times \mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ zawiera funkcje straty użyte do ewaluacji jakości modeli predykcyjnych.

[Model referencyjny] Model referencyjny $f_{ref} \in \mathcal{F}$ jest jednym z modeli wybranych z przestrzeni modeli predykcyjnych, jest on modelem odniesienia w stosunku do innych modeli. Model ten spełnia wybrane kryterium, np. minimalizuje funkcje straty $\mathcal{L}$ po wszystkich modelach f z rodziny $\mathcal{F}$,

$$f_{ref} = \text{argmin}_{f \in \mathcal{F}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim(\mathcal{X},\mathcal{Y})} \mathcal{L}(f, \mathbf{X}, \mathbf{Y}), \tag{4.1}$$

lub dowolny, inny referencyjny model, na przykład taki który aktualnie jest używany do danego problemu w praktyce. Wtedy zbiór modeli Rashomon można zdefiniować następująco:

[Zbiór Rashomon] Dla funkcji straty $\mathcal{L}$ i modelu referencyjnego $f_{ref} \in \mathcal{F}$, niech $\epsilon > 0$, zbiór Rashomon można zdefiniować jako:

$$R_{\mathcal{F}}(\mathcal{L}, \epsilon, f_{ref}) = \{f \in \mathcal{F} \mid \mathbb{E}\mathcal{L}(f, \mathbf{X}, \mathbf{Y}) \leq \mathbb{E}\mathcal{L}(f_{ref}, \mathbf{X}, \mathbf{Y}) + \epsilon\} \subset \mathcal{F}, \tag{4.2}$$

Funkcja straty $\mathcal{L}$ może być dowolną funkcją, która wskazuje jakość predykcyjną modelu, np. błąd średniokwadratowy (RMSE), pole pod krzywą ROC (AUC), czy dokładność (ACC). W przeprowadzonych eksperymentach, zbiór najlepszych modeli jest zdefiniowany na podstawie miary AUC, a modele porównywane są przy pomocy *profili*.

[Profil] Dla danego modelu $f \in \mathcal{F}$ i dla zmiennej $j \in \{1, 2, \ldots, p\}$, *profilem* nazywamy funkcję $g_f^j : \mathcal{D}_j \to \mathbb{R}$ która opisuje relację pomiędzy wartościami $j-tej$ zmiennej a predykcjami modelu f.

Jest wiele metod konstruowania takich profili. Najczęściej używanymi profilami są PDP [43], ALE (ang. *accumulated local effects*) [1], czy profile SHAP [81].

Dla profilu PDP użytego w dalszej części tej pracy, funkcja $g_f^j(z)$ jest zdefiniowana jako

$$g_{f,PDP}^j(z) = \mathbb{E}_{\mathcal{X}^{(-j)}}[f(X^{(1)}, \ldots, X^{(j-1)}, z, X^{(j+1)}, \ldots, X^{(p)})]$$

Jest to wartość oczekiwana predykcji modelu dla ustalonej wartości zmiennej j w punkcie z, przy rozkładzie łącznym $\mathcal{X}^{(-j)}$, a estymatorem jest:

$$\widehat{g_{f,PDP}^j}(z) = \frac{1}{n} \sum_{i=1}^{n} f(x_i^{(1)}, \ldots, x_i^{(j-1)}, z, x_i^{(j+1)}, \ldots, x_i^{(p)}).$$

4.3.1. Porównywanie różnic między profilami

Badając różnice między modelami na podstawie profili należy wybrać miarę, która określi ilościowo różnicę między profilami. Istnieje wiele miar odległości [39], które wskazują jak bardzo dwie krzywe różnią się. W przypadku profili wprowadzę nową miarę, dostosowaną do uchwycenia istotnych różnic dla profili modeli predykcyjnych.

W analizie funkcjonalnej [39] jedną z naturalnych możliwości jest metryka L2, czyli odległość euklidesowa pomiędzy funkcjami. Mierzy ona odległość pomiędzy profilami

$g^j_{f_1}, g^j_{f_2} : \mathcal{D}_j \to \mathbb{R}$ dla modeli f_1, f_2 i zmiennej objaśniającej $X^{(j)}$ w następujący sposób:

$$d_{L^2}(g^j_{f_1}, g^j_{f_2}) = \sqrt{\int_{\mathcal{D}_j} \left(g^j_{f_1}(z) - g^j_{f_2}(z)\right)^2 \mathrm{d}z}. \tag{4.3}$$

Jednak w wielu przypadkach podobieństwo profili zależy raczej od ich kształtów niż od wartości predykcyjnych. Aby uchwycić tę różnicę kształtu, można zastosować miarę uwzględniającą monotoniczność. Do tego celu służy miara L^2 między pochodnymi pierwszego rzędu:

$$d_{L^2,der}(g^j_{f_1}, g^j_{f_2}) = \sqrt{\int_{\mathcal{D}_j} \left(\frac{\partial}{\partial X^{(j)}} g^j_{f_1}(z) - \frac{\partial}{\partial X^{(j)}} g^j_{f_2}(z)\right)^2 \mathrm{d}z}. \tag{4.4}$$

W dziedzinie modeli liniowych często stosowanych w kontekście medycznym pojawia się scenariusz, w którym profile dla tej samej zmiennej mogą wykazywać niewielkie różnice ze względu na różne siły regularyzacji, często zachowując niezmienną zależność monotoniczności. Proponuję jednak własne podejście oparte na doświadczeniu, że podobne profile to te o podobnych kształtach i niekoniecznie podobnych wartościach predykcyjnych. Na przykład na Rysunku 3.3 uznaję za różne profile dla zmiennej smkyears dla modelu BACH i PLCO lub BACH i LCRAT, podczas gdy dla tej zmiennej profile dla modeli LCRAT i PLCO nie różnią się znacząco. Innym przykładem mogą być modele liniowe z regularyzacją, w których profile dla tej samej zmiennej mogą wyglądać różnie w zależności od siły regularyzacji, ale najczęściej zachowują tę samą zależność monotoniczności (znak współczynnika odpowiadającego zmiennej nie zmienia się). Postrzegając przesunięcie w górę lub w dół jako nieistotną zmianę w profilach, opieram nową metodę na porównaniu tego, jak często profil ma różne nachylenia. Interesujące jest sprawdzenie, czy jeden profil rośnie, a drugi maleje dla danej wartości zmiennej. Taka sytuacja oznacza, że profile wskazują różne zależności między zmienną a predykcją w rozpatrywanych modelach. Miarę tą definiuję jako wskaźnik rozbieżności profilu (ang. *profile disparity index, PDI*), a jej formuła jest następująca:

$$PDI(g^j_{f_1}, g^j_{f_2}) = \frac{1}{\sup \mathcal{D}_j - \inf \mathcal{D}_j} \int_{\mathcal{D}_j} \mathbb{1}\left[\operatorname{sgn}\left(\frac{\partial}{\partial X^{(j)}} g^j_{f_1}(z)\right) \neq \operatorname{sgn}\left(\frac{\partial}{\partial X^{(j)}} g^j_{f_2}(z)\right)\right] \mathrm{d}z.$$

Może być także estymowana w oparciu o profile empiryczne $\widehat{g^j_{f_1}}, \widehat{g^j_{f_2}}$ jako

$$\widehat{PDI}(\widehat{g^j_{f_1}}, \widehat{g^j_{f_2}}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[\text{sgn}\left(\text{der}(\widehat{g^j_{f_1}})[i]\right) \neq \text{sgn}\left(\text{der}(\widehat{g^j_{f_2}})[i]\right)\right], \tag{4.5}$$

gdzie $\text{der}(\widehat{g^j_{f_k}})$ jest wektorem pochodnych wyznaczonych w m kolejnych punktach profilu dla k^{tego} modelu i $\text{der}(\widehat{g^j_{f_k}})[i]$ jest i^{tym} elementem tego wektora. Należy zauważyć, że wartości profilu wyznaczane są w praktyce dla wybranych wartości zmiennej objaśniającej, np. dla argumentów z siatki lub wszystkich wartości występujących w próbie uczącej. Natomiast pochodne można wyznaczyć metodą Uogólnionej Ortogonalnej Pochodnej Lokalnej (GOLD) [26].

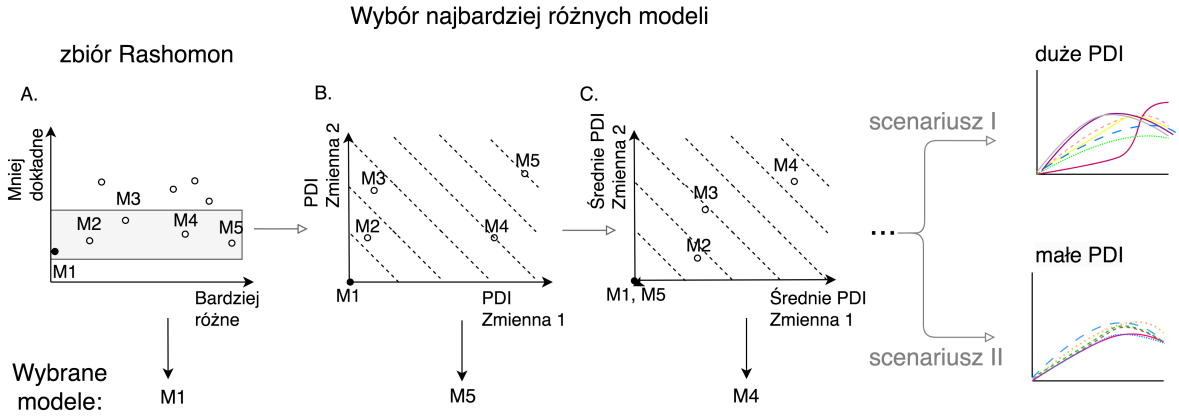
Interpretacja miary PDI jest bardzo intuicyjna. Jest to procent rozbieżności pomiędzy dwoma profilami i przyjmuje wartości z przedziału $[0, 1]$. Wartość 0 oznacza, że dwie krzywe są takie same, natomiast 1 oznacza, że dwie krzywe różnią się radykalnie.

W scenariuszach obejmujących zmienne kategoryczne zastosowanie instrumentów pochodnych nie jest wykonalne. Aby zaradzić temu ograniczeniu, w takich przypadkach algorytm stosuje odległości wektorowe jako alternatywną metodę pomiaru różnic. Takie podejście pomaga wypełnić lukę w zaproponowanej metodologii i zapewnia kompleksową analizę różnych typów zmiennych.

4.3.2. Porównanie modeli ze zbioru Rashomon

Identyfikacja modeli z ekstremalnymi zachowaniami ze zbioru Rashomon może mieć kluczowe znaczenie i jest ważnym krokiem w kierunku zbadania grupy modeli, zamiast analizowania tylko jednego. Znalezienie k najbardziej różnych modeli jest podobne do NP-trudnego problemu różnorodności MAXSUM [47]. Najprostszym sposobem byłoby opisać wszystkie możliwe podzbiory k-elementowe. W takim przypadku złożoność jest wykładnicza $O(2^M)$, gdzie M reprezentuje liczbę wszystkich modeli. W związku z tym takie rozwiązanie jest w zasadzie niemożliwe dla scenariuszy ze świata rzeczywistego z dużą liczbą liczba modeli. Dlatego proponuję rozwiązanie heurystyczne, algorytm `Rashomon_DETECT`, który rozwiązuje ten problem. Jego celem jest wybranie podzbioru najbardziej różnych ze zbioru utworzonych modeli $\hat{\mathcal{F}}$. Działanie algorytmu opiera się na obliczeniach miary różności dla par modeli w zachłanny sposób. Miarą w algorytmie

może być jedna z przedstawionych miar w sekcji 4.3.1. Schemat działania algorytmu zaprezentowany jest na Rysunku 4.2.



Rysunek 4.2: Graficzny schemat działania algorytmu `Rashomon_DETECT`. A.) Każda kropka reprezentuje pojedynczy model, a jasnoszary prostokąt zawiera modele ze zbioru Rashomona wyznaczone w stosunku do modelu M1, który wykazuje najniższą funkcję straty. Oś X przedstawia średnią miarę PDI pomiędzy M1 a kolejnymi modelami, oś Y przedstawia funkcję straty. B.) Panel ilustruje poszukiwanie modelu najbardziej różniącego się od modelu referencyjnego z poprzedniego kroku. Chociaż koncepcja została pokazana dla dwóch zmiennych, może być rozszerzona na p zmiennych. Kropki reprezentują wartości PDI pomiędzy M1 a kolejnymi modelami, a linie wskazują modele w jednakowej odległości od modelu referencyjnego pod względem PDI. C.) Panel ilustruje poszukiwanie modelu najbardziej różniącego się od punktu odniesienia będącego średnią PDI dla modeli referencyjnych z wcześniejszych kroków (w tym przypadku M1 i M5). Podejście to można dalej przenieść na kolejne etapy, w zależności od wielkości k. Scenariusz I wskazuje potencjalnie zróżnicowane modele, które wymagają dodatkowej walidacji w oparciu o wiedzę dziedzinową. Scenariusz II wskazuje modele o podobnych zachowaniach. Źródło: publikacja autora [70].

Algorytm rozpoczyna się od dodania do wyników zestawu modeli, pierwszego modelu, którym jest model referencyjny. Zwykle model ten minimalizuje zadaną funkcję straty. W kolejnych krokach dodawany jest do zestawu model, który najbardziej się różni (wg. wybranej miary, np. PDI) od ostatnio dodanych modeli. Miara liczona jest dla pojedynczej zmiennej, zatem wartość dla konkretnego modelu to średnia liczona po wszystkich zmiennych.

Require: $\hat{\mathcal{F}}$ – rodzina wyestymowanych modeli, $\mathcal{L}$ – funkcja straty, $\epsilon > 0$, $k \geq 2$ – liczba modeli, które chcemy znaleźć, g – metoda estymacji modeli, d – miara licząca różnice między profilami

Ensure: $\mathcal{R}$ s.t. $|\mathcal{R}| = k$ – zbiór najbardziej odelgłych modeli ze zbioru Rashomon

1: $f_{ref} \leftarrow \text{argmin}_{f \in \mathcal{F}} \mathcal{L}(f)$ ▷ zacznij od modelu referencyjnego lub wyznacz go
2: $R \leftarrow \{f \in \hat{\mathcal{F}} \mid \mathcal{L}(f) \leq \mathcal{L}(f_{ref}) + \epsilon\}$ ▷ wyznacz modele które będą rozważane
3: **if** $k \geq |R|$ **then** ▷ za mało modeli dla podanego k
4: $\quad \mathcal{R} \leftarrow R$
5: **else**
6: $\quad \mathcal{R} \leftarrow \{f_{ref}\}$
7: $\quad R \leftarrow R \setminus \{f_{ref}\}$
8: $\quad$ **while** $|\mathcal{R}| \leq k$ **do**
9: $\quad\quad f^* \leftarrow \text{argmax}_{f \in R} \frac{1}{|\mathcal{R}| \cdot p} \sum_{i=1}^{|\mathcal{R}|} \sum_{j=1}^{p} d(\widehat{g^j_{f_{\mathcal{R}_i}}}, \widehat{g^j_f})$ ▷ wybierz model z największą średnią miarą rozbieżności profili do modeli w $\mathcal{R}$
10: $\quad\quad \mathcal{R} \leftarrow \mathcal{R} \cup \{f^*\}$
11: $\quad\quad R \leftarrow R \setminus \{f^*\}$
12: $\quad$ **end while**
13: **end if**
14: **return** $\mathcal{R}$

Algorytm działa na zbiorze utworzonych modeli $\hat{\mathcal{F}}$ i opiera się na obliczeniu średnich różnic profili dla par modeli. Te wartości różnic są obliczane dla każdej zmiennej, a następnie uśredniane po wszystkich zmiennych. W przypadkach, gdy modele posiadają różne zestawy zmiennych, wszelkie brakujące profile są uzupełniane przy użyciu stałej funkcji $\mathbf{0}$.

Początkowe kroki algorytmu (linie 1-2) polegają na identyfikacji zbioru modeli Rashomona R. Następnie sprawdza się, czy zadanie nie jest trywialne, a konkretnie czy oczekiwana liczba wybranych modeli k jest większa lub równa krotności zbioru Rashomona (wiersz 3). Jeżeli ten warunek jest spełniony, do zmiennej wynikowej zostaje przypisany cały zbiór R (linia 4). Jeżeli jednak warunek nie jest spełniony, algorytm przechodzi do kolejnej, kluczowej fazy.

W tej części algorytmu proces rozpoczyna się od dodania jako pierwszego wyni-

kowego modelu – modelu referencyjnego (linia 6). Zazwyczaj model referencyjny odpowiada modelowi minimalizującemu zadaną funkcję straty. W kolejnych krokach algorytm identyfikuje model najbardziej różniący się od już wybranych, wykorzystując wybraną miarę (wiersz 9). Kryterium wyboru to średnia wartość różnicy dla modeli już znajdujących się w zestawie. Na koniec zbiór wynikowy zostaje uzupełniony o zidentyfikowany model (wiersz 10), a następnie model ten zostaje usunięty ze zbioru rozpatrywanych modeli (wiersz 11). To podejście iteracyjne jest kontynuowane aż do wybrania żądanej liczby modeli (wiersz 8).

Oprócz wyznaczonego zbioru najróżniejszych modeli, wynikiem algorytmu są obliczone profile i miary różnic między nimi. Wyniki te mogą zostać wykorzystane do dalszej analizy modeli, np. uzyskane wartości mogą pomóc w klastrowaniu modeli w odpowiednie zbiory modeli.

`Rashomon_DETECT` wykazuje złożoność $\mathcal{O}(gM^2kp)$, gdzie g oznacza rozmiar siatki wykorzystywanej do estymacji profilu, p to liczba zmiennych używanych przez modele, a k oznacza liczbę modeli do znalezienia (także liczbę iteracji). Jednakże w przypadku dużego zbioru modeli warto zastosować uproszczoną formę algorytmu, która działa zachłannie, włączając kolejne modele jedynie na podstawie porównania z ostatnio dodanym. Algorytm taki działa w złożoności $\mathcal{O}(gMkp)$. Jednak takie uproszczenie może prowadzić do gorszych wyników, potencjalnie wybierając modele skupione w klastrach, które różnią się tylko między sobą.

4.4. Eksperymenty

W tej części opisuję procedury przeprowadzone w celu oceny skuteczności proponowanego rozwiązania. Eksperymenty opisane w sekcji 4.4.1 służą porównaniu miar rozbieżności profili w różnych syntetycznie wygenerowanych scenariuszach. W kolejnym eksperymencie, opisanym w sekcji 4.4.2 wykorzystałam rzeczywisty zestaw danych HLH do wykonania algorytmu `Rashomon_DETECT` i przeprowadzenia dogłębnej analizy uzyskanych wyników, pokazując pełny przypadek użycia proponowanego podejścia. W sekcji 4.4.3, przedstawiony jest eksperyment porównujący wyniki algorytmu `Rashomon_DETECT` na zbiorach medycznych i stanowi ocenę skuteczności algorytmu.

4.4.1. Porównanie miar odległości na syntetycznych zbiorach danych

Przedstawiona symulacja wykorzystuje osiem scenariuszy obejmujących pary profili o różnych relacjach. W tym eksperymencie celem jest zademonstrowanie zachowania omawianych w rozdziale 4.3.1 miar. Wyniki w tej sekcji są niezależne od zastosowanego algorytmu i odnoszą się jedynie do kształtu profili. Warto zauważyć, że podczas gdy wartości PDI są ograniczone do przedziału [0, 1], pozostałe miary mogą przyjmować dowolne wartości z nieograniczonego przedziału $[0, \infty)$, w zależności od wartości rozważanych profili. Dlatego nie można przeprowadzić bezpośredniego porównania wartości miar w różnych scenariuszach. Zamiast tego analizowane są wartości w różnych scenariuszach i ich względne relacje (rankingi).

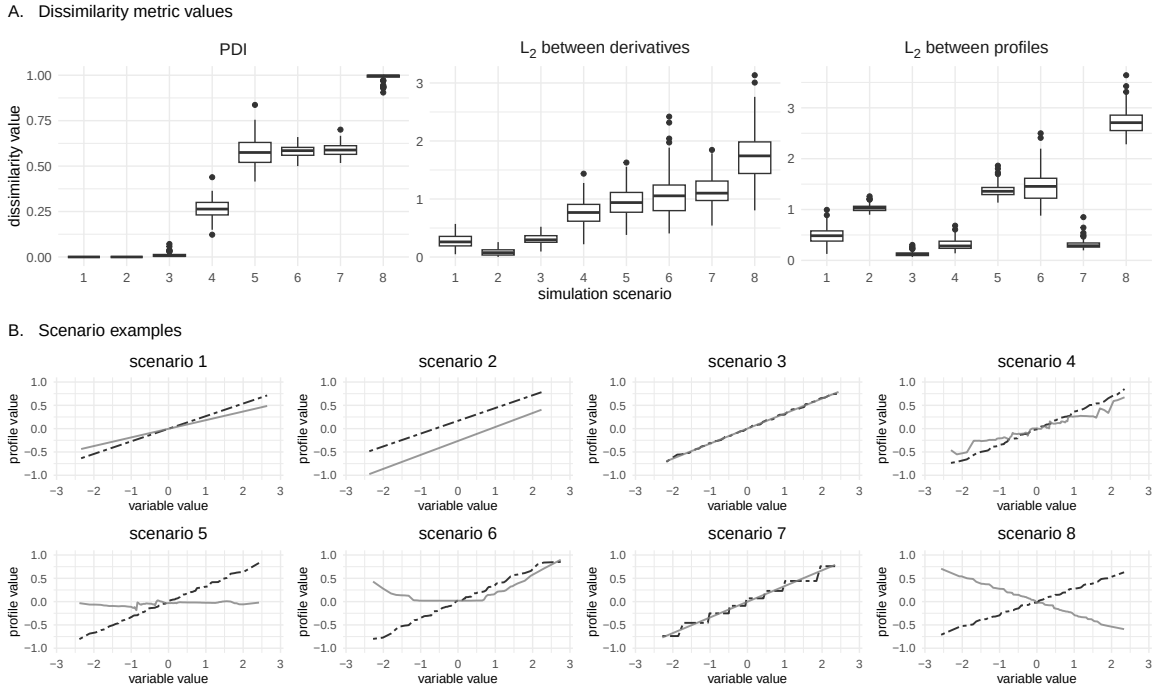
W każdym scenariuszu wygenerowanych zostało 100 par profili, które różnią się ze względu na szum losowy. W procesie generowania wykorzystywane są dostosowane do tego celu modele uczenia maszynowego, zapewniające wiarygodność powstałych profili uwzględniających oscylacje i szum.

Wyniki eksperymentu przedstawiono na rysunku 4.3. Panel A przedstawia rozkład wartości metryk pomiędzy scenariuszami, natomiast panel B przedstawia przykłady analizowanych scenariuszy. Miara PDI skutecznie rozróżnia scenariusze, w których profile wykazują podobne zależności (np. scenariusze 1-4) i te, w których widoczne są istotne różnice (np. scenariusze 5-6 i 8). Pozostałe miary wykazują mniej wyraźne różnice pomiędzy tymi dwiema grupami. Warto zauważyć, że mediana wartości każdej miary osiąga szczyt w scenariuszu 8, gdzie pokazane są odwrotne zależności profili. W przypadku PDI większość wartości osiąga maksimum 1, z niewielkimi odchyleniami wynikającymi z szumu występującego w profilach.

W scenariuszu 7 profil schodkowy porównywany jest z profilem liniowym. Chociaż profil nieliniowy jest podobny do scenariusza 3 i wykazuje podobne tendencje, charakteryzuje się większymi oscylacjami wokół linii. Prowadzi to do wyższych wartości metryki PDI i metryki opartej na drugiej pochodnej w porównaniu z innymi scenariuszami. Wynika to głównie ze względu na przedziały stałych wartości dających pochodne 0. Natomiast metryka euklidesowa postrzega tę sytuację inaczej, dając stosunkowo niskie wartości. Porównywalną, ale mniej wyraźną zależność obserwuje się w scenariuszu 4. W konsekwencji miara euklidesowa może być bardziej odpowiednia dla profili o dużych

oscylacjach (np. wysoki szum, funkcje schodkowe wynikające z pojedynczych, płytkich drzew).

Eksperyment podkreśla brak jednej najlepszej miary służącej do oceny różnic w profilach. Zamiast tego należy dobrać odpowiednią miarę do analizowanego problemu. Chociaż miara PDI sprawdza się w większości scenariuszy, należy pamiętać, że żadna miara nie gwarantuje optymalnej wydajności w każdym kontekście, szczególnie w skomplikowanych scenariuszach podejmowania decyzji medycznych.



Rysunek 4.3: A) Rozkłady wartości metryk dla analizowanych scenariuszy. Wyższe wartości odzwierciedlają większą różnicę profili w oparciu o odpowiednią miarę. Należy zauważyć, że osie Y różnią się ze względu na odrębne zakresy wartości teoretycznych każdej miary. B) Przykładowe pary profili wygenerowane dla każdego badanego scenariusza. Źródło: publikacja autora [70].

4.4.2. Studium przypadku - hemophagocytic lymphohistiocytosis

W tej sekcji przedstawię działanie algorytmu `Rashomon_DETECT` na rzeczywistym zbiorze danych medycznych. Zbiór składa się z 19 zmiennych charakteryzujących 101 dorosłych pacjentów z zespołem hemofagocytarnym (HLH, hemophagocytic lymphohistiocytosi). HLH to rzadki zespół skrajnego stanu zapalnego, który nieleczony prawie zawsze kończy się śmiercią. Nawet przy optymalnym leczeniu śmiertelność jest wysoka, ale jeśli początkowy stan zapalny zostanie wyciszony, a następnie kontrolowany, z czasem rokowanie jest znacznie lepsze. Zatem sześciomiesięczny czas przeżycia jest dobrym predyktorem sukcesu leczenia. HLH może być wyzwalany przez różne czynniki (nowotwory, infekcje, choroby autoimmunologiczne). Ponadto stopień wpływu na różne narządy i parametry jest różny u poszczególnych pacjentów. Ta różnorodność sprawia, że bardzo trudno jest przewidzieć wynik leczenia. Głównym celem zadania prognostycznego jest przewidzenie, czy pacjent przeżyje sześć miesięcy po postawieniu diagnozy.

Przeprowadziłam porównanie algorytmów uczenia maszynowego, które przewidują, czy pacjenci przeżyją 6 miesięcy. Do porównania włączyłam następujące metody: maszyny wektorów podpierających (ang. *support vector machines, SVM*), drzewo decyzyjne, model wzmocnienia gradientowego oraz lasu losowego. Podzieliłam zestaw danych na zestawy testowe i treningowe aby ocenić wydajność predykcji modeli za pomocą miary AUC. Najlepszą wydajność ma jeden z modeli lasu losowego (AUC równy 0,81 na zbiorze danych testowych) i jest to model referencyjny. Jednak w badaniu uwzględniłam również inne modele, które należą do zbioru Rashomon: pięć modeli lasu losowego i cztery modele wzmocnienia gradientowego. W dalszej części tego rozdziału używam nazw modeli `rf` dla lasów losowych i `gbm` dla wzmocnienia gradientowego z odpowiednią liczbą sugerującą, który pod względem wielkości AUC jest dany model. Pozostałe modele (drzew decyzyjnych i maszyn wektorów podpierających) znalazły się poza zbiorem Rashomon.

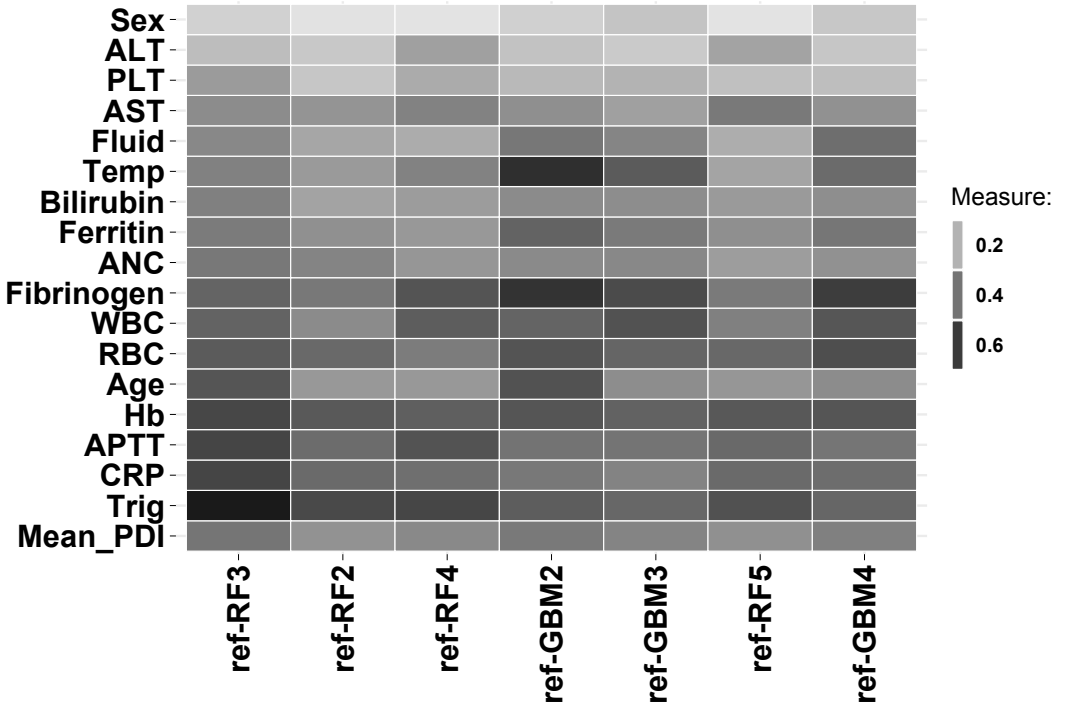
Mając zbiór modeli Rashomon, zdecydowałam się wybrać podzbiór składający się z 3, najbardziej zróżnicowanych modeli. Wybór odpowiedniej wielkości szukanego podzbioru zależy od wielkości zestawu Rashomon. W analizowanym przypadku zestaw Rashomon składa się tylko z 9 modeli. W tym celu przeprowadziłam algorytm

Rashomon_DETECT i wykorzystałam miarę PDI.

Model rf1 jest referencyjnym modelem z najwyższym AUC. W drugim kroku dla par: referencyjnego modelu i każdego modelu ze zbioru Rashomon obliczyłam miary PDI i ich średnią wartość (uśredniając po wszystkich zmiennych). Modelem najbardziej różnym okazał się gbm1 W kolejnym kroku algorytmu najbardziej różny model został zidentyfikowany jako rf3. Rysunek 4.4 przedstawia średnie wartości PDI w ostatnim kroku algorytmu Rashomon_DETECT. Wartość ref oznacza wartość średniego PDI dla wybranych modeli we wcześniejszych krokach (w tym przypadku rf1 i gbm1).

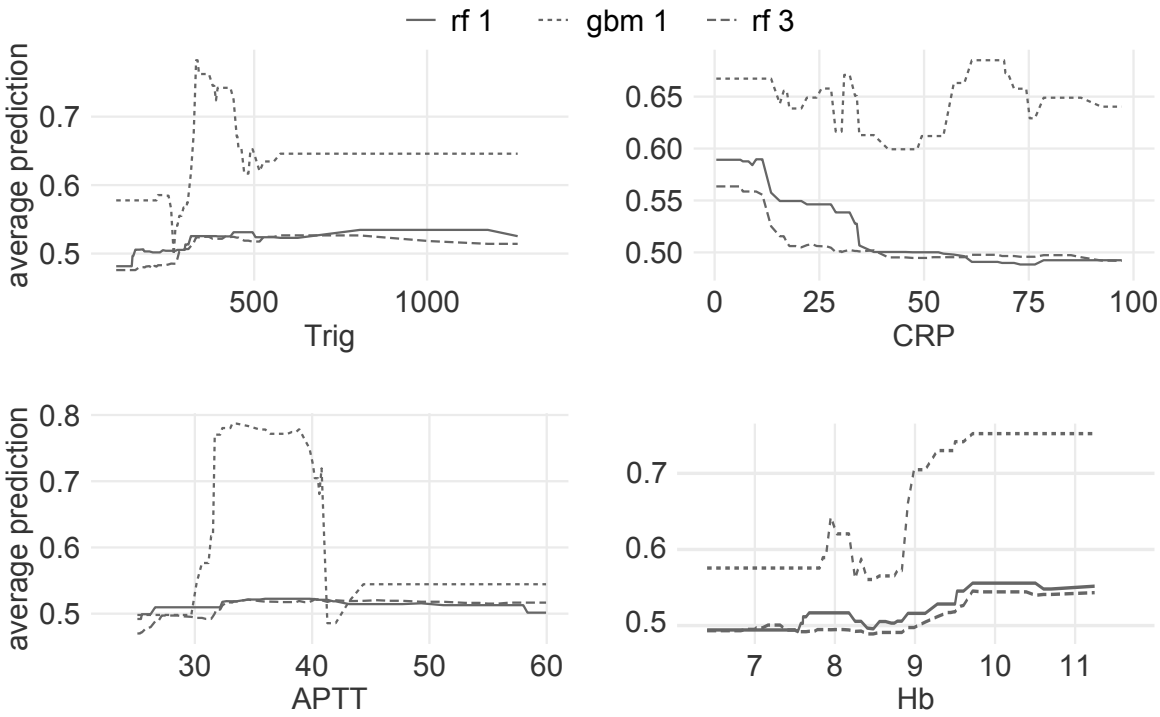
Na podstawie Rysunku 4.4 najwyższe wartości PDI (przekraczające 50%) obserwuje się dla zmiennych: Trig, CRP i APTT. Te zmienne zostały zatem zaprezentowane na Rysunku 4.5, który przedstawia wykresy cząstkowej zależności PDP. Wykresy te potwierdzają różnice pomiędzy profilami, które zostały zidentyfikowane za pomocą miary PDI. Zaprezentowane przykłady pokazują, że przewidywania dotyczące średniego przeżycia pacjenta przy pomocy modelu gbm1 wykazuje wiele oscylacji, podczas gdy modele rf1 i rf3 są znacznie bardziej stabilne. Wahania te są sprzeczne z medyczną wiedzą na temat zmiennych Trig i CRP, podczas gdy pozostałe modele wskazują wiarygodne tendencje dla tych zmiennych. Białko CRP o wyższych wartościach wiąże się z większym stanem zapalnym w organizmie i potencjalnie koinfekcją bakteryjną. Jego niskie wartości są prawidłowo kojarzone z lepszym rokowaniem. Niewielki wzrost przeżywalności przy rosnącym stężeniu trójglicerydów wydaje się nieco przeciwstawne intuicji. Można to błędnie zinterpretować jako wadę modeli rf1 i rf2. Natomiast jest to związane z konstrukcją baza danych - pacjenci byli diagnozowani na podstawie kryteriów HLH- kryteria z 2004 r., a jednym z nich jest hipertriglicerydemia i/lub hipofibrynogemia. Zatem u niektórych pacjentów występuje wysoki poziom trójglicerydów spełniający kryterium HLH zamiast niskiego poziomu fibrynogenu. Jednak niski poziom fibrynogenu stwarza większe zagrożenie związane z koagulopatią i ryzykiem krwawienia. Zaletą oscylacji modelu gbm1 jest z kolei spadek w przewidywaniu przeżycia wraz z wydłużeniem czasu krwawienia APTT przez około 40 sekund. W przypadku Hb różnice w modelach wynikają głównie z innych wartości predykcji. Natomiast wszystkie trzy modele pokazują, zgodnie z oczekiwaniami, że wyższe wartości zmiennej wiążą się z wyższymi wartościami predykcji.

To zastosowanie algorytmu Rashomon_DETECT do danych medycznych pokazuje, że kilka prawie tak samo dobrych modeli pod względem jakości, może opisywać naturę



Rysunek 4.4: Porównanie średniej miary PDI dla par `ref` (wartość referencyjna średniego PDI dla wybranych we wcześniejszych krokach modeli) i pozostałych modeli, które nie zostały jeszcze wybrane. Wykres przedstawia wartości w ostatnim, trzecim kroku algorytmu. Im ciemniejsze pole, tym wyższa miara, co oznacza większe różnice między modelami. Następujące skróty odnoszą się do nazw zmiennych: `ALT` (aminotransferaza alaninowa), `ANC` (bezwzględna liczba neutrofili), `APTT` (czas częściowej trombiny aktywowanej), `AST` (aminotransferaza asparaginianowa) , `Age` (wiek), `Bilirubin` (bilirubina), `CRP` (białko C-reaktywne), `Ferritin` (ferrytyna), `Fibrinogen` (fibronektyna), `Fluid` (płyny), `Hb` (hemoglobina), `PLT` (liczba płytek krwi), `RBC` (liczba czerwonych krwinek), `Trig` (trójglicerydy). Źródło: publikacja autora [70].

danych na różne sposoby. Procedura ta pozwala na bezpieczniejsze wnioskowanie o problemach medycznych, ponieważ bierze pod uwagę kilka modeli z najbardziej różnymi spojrzeniami na zjawisko. Pozwala ona badaczowi lub praktykowi wybrać jeden najlepiej zachowujący się model zgodnie z wiedzą dziedzinową lub użyć kilku modeli, które mają różne, ale nadal dobre zachowania.



Rysunek 4.5: Wykresy cząstkowej zależności PDP dla zmiennych z największymi wartościami miary PDI. Źródło: publikacja autora [70].

4.4.3. Porównanie wyników `Rashomon_DETECT` na rzeczywistych zbiorach danych

Celem tego badania porównawczego jest ocena skuteczności algorytmu `Rashomon_DETECT` zastosowanego w modelach uczenia maszynowego obliczone na podstawie zbiorów danych medycznych. Rozważane zbiory danych obejmują różne problemy medyczne, pochodzą z badania przeprowadzonego na platformie SeFNet. Zostały wybrane następujące zbiory danych: PIMA, COVID, ILPD i Heart. Zbiór danych PIMA zawiera informacje o 768 przypadkach pacjentów opisanych przez 9 zmiennych. Celem eksperymentu jest przewidzenie, czy pacjent ma cukrzycę. Zbiór danych COVID zawiera informacje o 603 przypadkach opisanych 19 zmiennymi. Celem modelowania w tym przypadku jest predykcja czy pacjent, który został przyjęty do szpitala cierpi na Covid-19. Zbiór danych ILPD [95] zawiera informacje o 583 przypadkach pacjentów opisanych 11 zmiennymi. Celem modelowania jest wykrycie pacjentów z chorobami

Measure	COVID	PIMA	ILPD	Heart
PDI	0.42	0.37	0.42	0.19
	0.39	0.35	0.32	0.16
	0.31	0.13	0.16	0.16
L^2	0.08	0.1	0.7	0.08
	0.06	0.09	0.57	0.06
	0.03	0.04	0.07	0.06
L^2 między	0.13	0.09	0.06	0.08
pochodnymi	0.1	0.07	0.05	0.06
	0.07	0.04	0.02	0.06

Tabela 4.1: Średnie wartości miar odległości dla wszystkich par modeli z podzbioru optymalnych modeli. Żródło: publikacja autora [70].

wątroby. Zbiór danych Heart zawiera informacje o 1190 pacjentach opisanych przez 12 zmiennych. Celem eksperymentu jest przewidzenie choroby wieńcowej. Szczegółowe informacje na temat zbiorów danych można znaleźć w [134].

Dla każdego zbioru danych przeprowadzany jest podobny proces poszukiwania optymalnych modeli. Przeprowadzone podejście obejmuje następujące etapy: 5-krotna walidacja krzyżowa podczas szukania hiperparametrów i budowania zbiorów Rashomona na podstawie średniego wyniku z tej walidacji krzyżowej. Etap szukania optymalnych hiperparametrów obejmuje eksplorację siatki parametrów, koncentrując się na dwóch klasach modeli uczenia maszynowego: lasu losowego i wzmocnienia gradientowego. Zbiory Rashomona obejmują wszystkie modele, których średnie pole pod krzywą (AUC) z 5-krotnej walidacji krzyżowej nie jest mniejsze niż zadany epsilon.

Celem tego badania jest określenie skuteczności algorytmu `Rashomon_DETECT` w identyfikacji najbardziej zróżnicowanych profili w ramach rozpatrywanych zestawów modeli. Algorytm `Rashomon_DETECT` został wykonany trzykrotnie na każdym zestawie danych, za każdym razem z inną miarą. Tabela 4.4.3 przedstawia wyniki wielkości miar uzyskane podczas przeprowadzenia tego eksperymentu. Wyniki pokazują, że zbiór Heart charakteryzuje się niższymi wartościami PDI w porównaniu do innych zbiorów danych, co zostało potwierdzone w dalszej części tej pracy (panel B na Rysunku 4.7). Zbiór danych ILPD charakteryzuje największe wartości miary L^2 pomiędzy profilami, podczas gdy zbiór danych dotyczących COVID osiągnął najwyższe wartości L^2 między pochodnymi pierwszego rzędu.

Skuteczność algorytmu mierzona jest poprzez identyfikację rzeczywistych rozbieżności w profilach, przedstawiając na wykresach indywidualne profile dla wszystkich modeli i podkreślając kolorem modele, w których algorytm wskazał istotne różnice. Wykresy 4.6 i 4.7 przedstawiają cząstkowe zależności PDP dla każdego modelu. Potwierdzają dokładność wyników uzyskanych przy pomocy algorytmu `Rashomon_DETECT`.

Przedstawione w ten sposób wyniki potwierdzają, że algorytm w przeważającej mierze identyfikuje modele o najbardziej różnych profilach. Na każdym rysunku najbardziej odstające krzywe to te wykryte przez algorytm. Największe różnice powstają pomiędzy różnymi typami modeli. W przypadku zbioru danych PIMA wykres potwierdza, że algorytm pomyślnie zidentyfikował odpowiednie modele, tj. te które najbardziej się różnią. Najbardziej zróżnicowany od modelu zielonego to niebieski, który reprezentuje model wzmocnienia gradientowego. Zielone i czerwone krzywe z kolei odpowiadają dwóm różnym lasom losowym. Modele lasu losowego wykazują również różnice w przypadku pojedynczych zmiennych (np. insulina, czy BMI). Algorytm zidentyfikował najbardziej różne modele dla zbiorów ILPD i COVID. W obu tych przypadkach, krzywa niebieska różni się najbardziej w porównaniu z modelem referencyjnym, reprezentowanym przez zieloną krzywą. Inne modele zachowują się bardzo podobnie. Wykres pokazany dla zbioru danych Heart także potwierdza skuteczność algorytmu, gdyż modele czerwony i zielony różnią się najmocniej.

Profile przedstawione na rysunkach 4.6 i 4.7 ilustrują wynik algorytmu przy użyciu najbardziej pasującej miary dla każdego zestawu danych, zgodnie z wynikami eksperymentu z sekcji 4.4.1.

W przypadku zbioru danych Heart, w którym większość profili wykazuje oscylacje zgodne ze scenariuszem 4 i scenariuszem 7, przyjęto miarę euklidesową. Profile wyprowadzone ze zbioru danych PIMA w dużej mierze przypominają scenariusz 6, prowadzący do zastosowania miary PDI. Profile powiązane ze zbiorami danych ILPD i COVID wykazują podobieństwa do scenariuszy 6 i 4, co także prowadzi do zastosowania miary PDI.

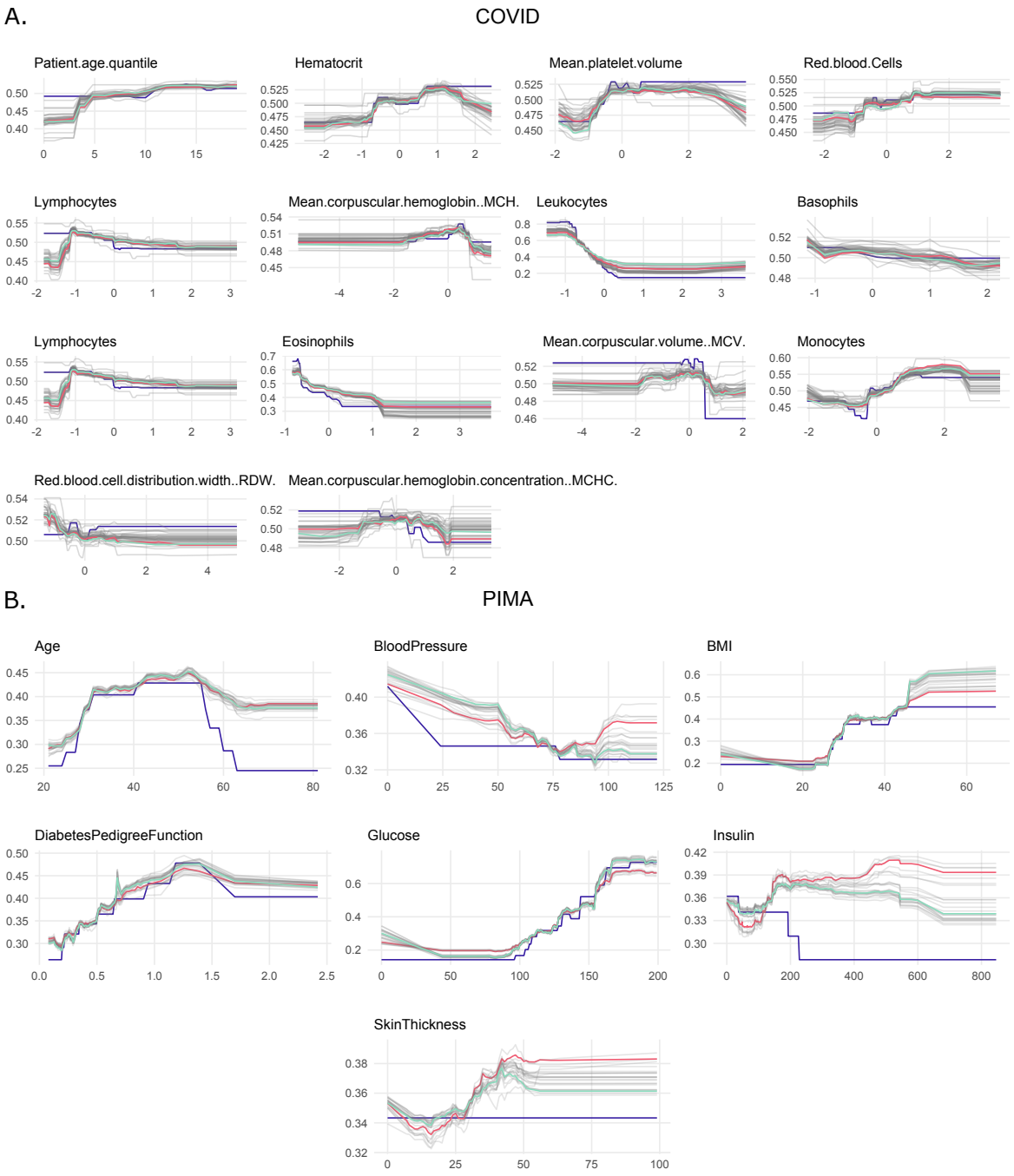
Podsumowując, badanie to podkreśla znaczenie uwzględnienia wyjaśnień w ramach zbiorów Rashomona. Stosując algorytm `Rashomon_DETECT` możemy znaleźć najbardziej zróżnicowane modele, wzbogacając tym samym nasze rozumienie złożoności modelowania predykcyjnego. Podejście to stanowi krok w kierunku zwiększenia niezawodności i możliwości interpretacji modeli uczenia maszynowego w różnych dziedzinach. Na

przykład, modele zidentyfikowane na potrzeby zbioru danych PIMA wykazują znaczne rozbieżności w podejściu do wieku i insuliny.

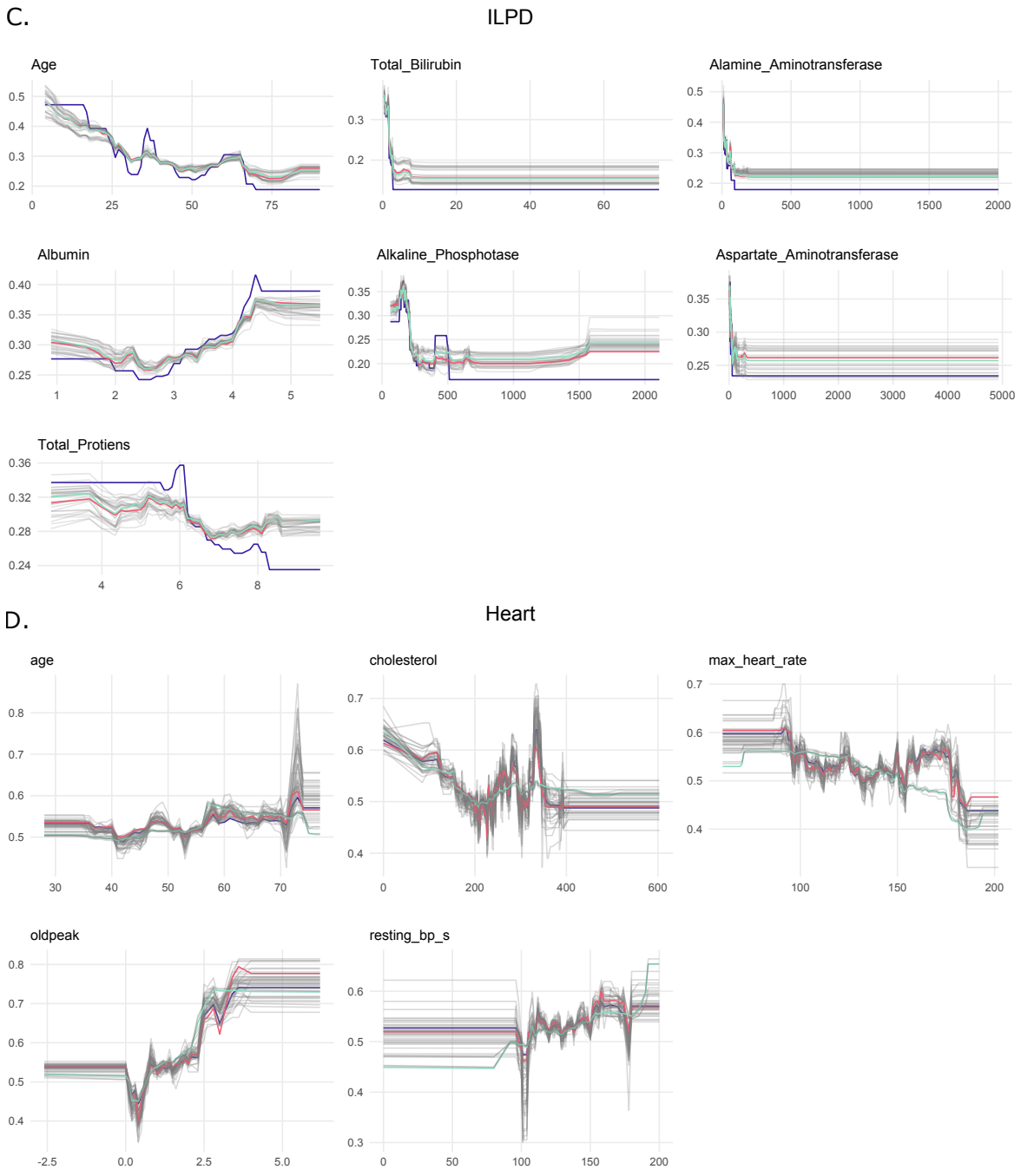
4.5. Podsumowanie

W tym badaniu przedstawiam nowy proces eksploracji modeli ze zbioru Rashomon. Takie podejście można zastosować w przypadku dowolnego problemu wykorzystując uczenie maszynowe, aby uzyskać szerszą perspektywę wyjaśniania danych. Wybór modeli uczenia maszynowego ze zbioru Rashomon i ich analiza może dać bardziej wiarygodne wyniki niż analiza tylko jednego modelu. Zwłaszcza, gdy istnieje wiele niemal optymalnych modeli, należy porównać ich zachowania. Zamiast porównywania tylko jakości modeli proponuję porównanie ich profili. Jeśli interpretacje tych modeli są różne, wyniki i ostateczna decyzja powinny zostać wzmocnione, na przykład wiedzą dziedzinową lub innymi metodami XAI. Co więcej, takie szersze spojrzenie daje większe zaufanie do ostatecznego modelu.

Ponadto pokazuję możliwość agregacji cząstkowej zależności PDP poprzez obliczenie miary PDI, co ułatwia interpretację wyników. Otrzymane rezultaty pokazują, że istniejące metody XAI mogą zostać rozszerzone o nowy algorytm szukania optymalnego modelu i miarę służącą do porównania wyjaśnień modeli. Hipoteza 4 z sekcji 1.3 została zatem potwierdzona.



Rysunek 4.6: Wykresy cząstkowej zależności PDP dla ciągłych zmiennych przedstawione na zbiorach danych: A. COVID, B. PIMA. Każda krzywa przedstawia profil dla pojedynczego modelu ze zbioru Rashomon. Trzy najbardziej różne modele wykryte dla każdego zbioru danych są zaznaczone kolorem czerwonym, zielonym i niebieskim. Źródło: publikacja autora [70].



Rysunek 4.7: Wykresy cząstkowej zależności PDP dla ciągłych zmiennych przedstawione na zbiorach danych: C. ILPD, D. Heart. Każda krzywa przedstawia profil dla pojedynczego modelu ze zbioru Rashomon. Trzy najbardziej różne modele wykryte dla każdego zbioru danych są zaznaczone kolorem czerwonym, zielonym i niebieskim. Źródło: publikacja autora [70].

Rozdział 5

Podsumowanie

W niniejszej rozprawie zweryfikowałam empirycznie wszystkie założone hipotezy badawcze. Pokazałam, że dzięki metodom wyjaśnialnej sztucznej inteligencji można zwiększyć jakość wnioskowania na podstawie modelowania. W szczególności mtody XAI pomogły w zwiększeniu dokładności modelowania predykcyjnego, zwalidowaniu istniejących modeli, zrozumieniu i zbudowaniu zaufania do modelowania. Ponadto przedstawiłam nową metodą XAI, która wspomaga wnioskowanie na podstawie modeli uczenia maszynowego.

W moich badaniach przedstawiłam metody wyjaśnialnego uczenia maszynowego zastosowane do rzeczywistych problemów medycznych. Wyniki zostały opublikowane w czasopismach naukowych. Zaproponowane modele przedstawione w pracy oraz te nad którymi pracowałam, a nie stanowią integralnej części tej pracy, stanowiły punkt wyjścia do rozważań i rozwijania nowej metody badawczej. Ponadto większość z tych modeli stosowana jest w dużych ośrodkach badawczych. Na podstawie moich badań, zaobserwowałam, że opisując badane zjawisko, często możemy znaleźć wiele równie dobrych modeli predykcyjnych. Wykorzystanie modeli w medycynie jest obarczone dużym ryzykiem. Skłoniło mnie to do pracy nad modyfikacją procesu wyboru modelu predykcyjnego, rozszerzając proces o analizę całego zbioru niemal optymalnych modeli, zbioru modeli Rashomon. Proponuję porównanie modeli o podobnej jakości na podstawie jednej z metod wyjaśnialnego uczenia maszynowego, metody wyliczania profili, przedstawiających zależności między zmiennymi a predykcjami modelu. W wyniku tej pracy powstał algorytm służący do szukania podzbioru najbardziej różnych modeli

spośród równie skutecznych modeli predykcyjnych, ułatwiając ich porównanie i weryfikację. Taka modyfikacja procesu budowy modelu zwiększa wiarygodność modelowania oraz zaufanie do wyników modelowania. Artykuł, w którym opisuję to badanie jest zamieszczony na arXiv i w trakcie recenzji do czasopisma IEEE Journal of Biomedical and Health Informatics.

Moje badania mają charakter interdyscyplinarny, wnosząc wkład zarówno w dziedzinę informatyki, ze szczególnym uwzględnieniem uczenia maszynowego, jak i w rozwój medycyny. W kontekście informatyki, badania te przyczyniły się do opracowania nowego algorytmu, który zwiększa jakość wnioskowania na podstawie modeli predykcyjnych. Dzięki temu możliwe jest budowanie bardziej zaawansowanych i transparentnych systemów sztucznej inteligencji. W obszarze medycyny, moje badania mają na celu poprawę procesów diagnostycznych i prognostycznych, zwłaszcza w kontekście chorób takich jak rak płuca czy sepsa. Wykorzystanie zaawansowanych modeli uczenia maszynowego do analizy danych medycznych umożliwia lepsze zrozumienie czynników wpływających na wyniki leczenia, co może prowadzić do bardziej spersonalizowanych i skutecznych strategii terapeutycznych.

W moich dalszych badaniach będę skupiać się na rozwoju uczenia maszynowego i jego zastosowaniach w medycynie. Chciałabym rozwinąć zaproponowany algorytm `Rashomon_DETECT` i dalej weryfikować czy podejście analizy wielu niemal optymalnych modeli może być stosowane w medycynie.

Bibliografia

[1] Dan Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B*, 82(4):1059–1086, 2020.

[2] Peter B Bach, Michael W Kattan, Mark D Thornquist, Mark G Kris, Ramsey C Tate, Matt J Barnett, Lillian J Hsieh, and Colin B Begg. Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute*, 2003.

[3] Hubert Baniecki, Dariusz Parzych, and Przemyslaw Biecek. The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery*, pages 1–37, 2023.

[4] Andrew L Beam and Isaac S Kohane. Translating artificial intelligence into clinical care. *JAMA*, 316(22):2368–2369, December 2016.

[5] Brett K Beaulieu-Jones, William Yuan, Gabriel A Brat, Andrew L Beam, Griffin Weber, Marshall Ruffin, and Isaac S Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit. Med.*, 4(1):62, March 2021.

[6] Lorenzo Belenguer. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI Ethics*, 2(4):771–787, February 2022.

[7] Tomasz Berus, Anna Markiewicz, Katarzyna Kobylinska, Przemyslaw Biecek, Jolanta Orlowska-Heitzman, Bozena Romanowska-Dixon, and Piotr Donizy. Down-regulation of polo-like kinase-1 (plk-1) expression is associated with poor clinical outcome in uveal melanoma patients. *Folia Histochemica et Cytobiologica*, 58(2):

108–116, 2020. ISSN 1897-5631. doi: {}. URL https://journals.viamedica.pl/folia_histochemica_cytobiologica/article/view/FHC.a2020.0017.

[8] Przemyslaw Biecek. DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, 19(84):1–5, 2018.

[9] Przemyslaw Biecek. Model development process, 2019.

[10] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. ISBN 9780367135591.

[11] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016.

[12] Leo Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[13] Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3): 199–215, 2001.

[14] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[15] Jacob Calvert, Thomas Desautels, Uli Chettipally, Christopher Barton, Jana Hoffman, Melissa Jay, Qingqing Mao, Hamid Mohamadlou, and Ritankar Das. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg (Lond)*, 8:50–55, May 2016.

[16] Jacob S Calvert, Daniel A Price, Uli K Chettipally, Christopher W Barton, Mitchell D Feldman, Jana L Hoffman, Melissa Jay, and Ritankar Das. A computational approach to early sepsis detection. *Comput Biol Med*, 74:69–73, May 2016.

[17] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

[18] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, M. Sturm, and Noémie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[19] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. *The CRISP-DM 1.0 Step-by-step data mining guide*. 1999.

[20] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

[21] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, Wei Xie, Gail L Rosen, Benjamin J Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M Cofer, Christopher A Lavender, Srinivas C Turaga, Amr M Alexandari, Zhiyong Lu, David J Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K Wiley, Marwin H S Segler, Simina M Boca, S Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S Greene. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface*, 15(141):20170387, April 2018.

[22] Council of European Union. Council regulation (EU) no 2021/0106(cod), 2024. http://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0012.02/DOC_1&format=PDF.

[23] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020.

[24] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *J. Big Data*, 6(1), December 2019.

[25] Harry J. de Koning, Carlijn M. van der Aalst, Pim A. de Jong, Ernst T. Scholten, Kristiaan Nackaerts, Marjolein A. Heuvelmans, Jan-Willem J. Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, Susan van 't Westeinde, Mathias Prokop, Willem P. Mali, Firdaus A.A. Mohamed Hoesein, Peter M.A. van Ooijen, Joachim G.J.V. Aerts, Michael A. den Bakker, Erik Thunnissen, Johny Verschakelen, Rozemarijn Vliegenthart, Joan E. Walter, Kevin ten Haaf, Harry J.M. Groen, and Matthijs Oudkerk. Reduced lung-cancer mortality with volume ct screening in a randomized trial. *New England Journal of Medicine*, 382(6):503–513, 2020. doi: 10.1056/NEJMoa1911793. URL `https://doi.org/10.1056/NEJMoa1911793`. PMID: 31995683.

[26] Pascal R. Deboeck. Estimating Dynamical Systems: Derivative Estimation Hints From Sir Ronald A. Fisher. *Multivariate Behavioral Research*, 45:725–745, 2010.

[27] Ben Dickson. Inside darpa's effort to create explainable artificial intelligence, 2019. URL `https://bdtechtalks.com/2019/01/10/darpa-xai-explainable-artificial-intelligence/`.

[28] Jiayun Dong and Cynthia Rudin. Exploring the Cloud of Variable Importance for the Set of All Good Models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

[29] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[30] Aberle DR, Adams AM, Berg CD, Black WC, Fagerstrom RM Clapp JD, Gareen IF, Gatsonis C, Marcus PM, and Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[31] Joanna Drozd-Sokolowska, Jan Maciej Zaucha, Przemyslaw Biecek, Agnieszka Giza, Katarzyna Kobylinska, Monika Joks, Tomasz Wrobel, Beata Kumiega, Wanda Knopinska-Posluszny, Wojciech Spychalowicz, Joanna Romejko-Jarosinska, Joanna Fischer, Wieslaw Wiktor-Jedrzejczak, Monika Dlugosz-Danecka, Sebastian Giebel, and Wojciech Jurczak. Type 2 diabetes mellitus

compromises the survival of diffuse large b-cell lymphoma patients treated with (R)-CHOP - the PLRG report. *Sci. Rep.*, 10(1):3517, February 2020.

[32] Joanna Drozd-Sokołowska, Krzysztof Mądry, Joanna Barankiewicz, Katarzyna Kobylińska, Przemysław Biecek, Jagoda Rytel, Ewa Karakulska-Prystupiuk, Kamila Skwierawska, Aleksander Salomon-Perzyński, Tomasz Stokłosa, and Grzegorz Władysław Basak. SARS-CoV-2 infection in patients treated with azacitidine and venetoclax for acute leukemia: A report of a case series treated in a single institution. *Chemotherapy*, 68(1):16–22, 2023.

[33] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning, 2019.

[34] Randall J. Ellis, Ryan M. Sander, and Alfonso Limon. Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine*, 6:100068, 2022. ISSN 2666-5212.

[35] EU Expert Group. Ethics guidelines for trustworthy ai. Online, 2019. URL `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`.

[36] European Commission. On artificial intelligence - a european approach to excellence and trust. Online, 2020. URL `https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en`.

[37] General Data Protection Regulation European Union. Regulation (eu) 2016/679 of the european parliament and of the council, 2016. URL `"https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2838-1-1"`.

[38] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, Christa Schorr, Steven Simpson, W Joost Wiersinga, Fayez Alshamsi, Derek C Angus, Yaseen Arabi, Luciano Azevedo, Richard Beale, Gregory Beilman, Emilie Belley-Cote, Lisa Burry, Maurizio Cecconi, John

Centofanti, Angel Coz Yataco, Jan De Waele, R Phillip Dellinger, Kent Doi, Bin Du, Elisa Estenssoro, Ricard Ferrer, Charles Gomersall, Carol Hodgson, Morten Hylander Møller, Theodore Iwashyna, Shevin Jacob, Ruth Kleinpell, Michael Klompas, Younsuck Koh, Anand Kumar, Arthur Kwizera, Suzana Lobo, Henry Masur, Steven McGloughlin, Sangeeta Mehta, Yatin Mehta, Mervyn Mer, Mark Nunnally, Simon Oczkowski, Tiffany Osborn, Elizabeth Papathanassoglou, Anders Perner, Michael Puskarich, Jason Roberts, William Schweickert, Maureen Seckel, Jonathan Sevransky, Charles L Sprung, Tobias Welte, Janice Zimmerman, and Mitchell Levy. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Med*, 47(11):1181–1247, October 2021.

[39] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis*. Springer New York, 2006.

[40] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. Model class reliance: Variable importance measures for any machine learning model class, from the 'rashomon' perspective. *Journal of Computational and Graphical Statistics*, 2018. URL `http://arxiv.org/abs/1801.01489`.

[41] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, 2019.

[42] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[43] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29:1189–1232, 2000.

[44] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[45] María Vega García and José L. Aznarte. Shapley additive explanations for no2 forecasting. *Ecological Informatics*, 2020.

[46] Piotr Gawrysiak, Muraszkiewicz Mieczysław, Nowak Robert, and Politechnika Warszawska Oficyna Wydawnicza. *Odpowiedzialna Sztuczna Inteligencja*. 2024.

[47] Jay B. Ghosh. Computational aspects of the maximum diversity problem. *Oper. Res. Lett.*, 19(4):175–181, 1996. doi: 10.1016/0167-6377(96)00025-9.

[48] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. 2018.

[49] OS Glotzer, T Fabian, A Chandra, and CT Bakhos. Non-small cell lung cancer therapy: safety and efficacy in the elderly. *Drug Healthc Patient Saf*, 22:113–121, 2013.

[50] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, 2014.

[51] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095. URL `https://doi.org/10.1080/10618600.2014.907095`.

[52] Arieh Gomolin, Elena Netchiporouk, Robert Gniadecki, and Ivan V Litvinov. Artificial intelligence applications in dermatology: Where do we stand? *Front. Med. (Lausanne)*, 7:100, March 2020.

[53] Alicja Gosiewska and Przemyslaw Biecek. Do not trust additive explanations, 2019.

[54] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. 51(5), 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL `https://doi.org/10.1145/3236009`.

[55] Masaki Hamamoto and Masashi Egi. Model-agnostic Ensemble-based Explanation Correction Leveraging Rashomon Effect. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–08, 2021.

[56] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *Metabolism*, 69:S36 – S40, 2017. ISSN 0026-0495. doi: https://doi.org/10.1016/j.metabol.2017.01.011. URL `http://www.sciencedirect.com/science/article/pii/S002604951730015X`. Insights Into the Future of Medicine: Technologies, Concepts, and Integration.

[57] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: Data mining, inference, and prediction, second edition. *Springer Series in Statistics*, 2009.

[58] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923, 2017. URL `http://arxiv.org/abs/1712.09923`.

[59] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 9(4):e1312, July 2019.

[60] Hsiang Hsu and Flavio Calmon. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. In *Advances in Neural Information Processing Systems*, volume 35, pages 28988–29000. Curran Associates, Inc., 2022.

[61] Stephanie Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 2020.

[62] Christoph Jansen, Thomas Penzel, Stephan Hodel, Stefanie Breuer, Martin Spott, and Dagmar Krefting. Network physiology in insomnia patients: Assessment of relevant changes in network topology with interpretable machine learning models. *Chaos*, 29(12):123129, December 2019.

[63] Bernhard O. Josephus, Ardianto H. Nawir, Evelyn Wijaya, Jurike V. Moniaga, and Margaretha Ohyver. Predict mortality in patients infected with covid-19 virus based on observed characteristics of the patient using logistic regression. *Procedia Computer Science*, 179:871–877, 2021. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.01.076. URL `https://www.sciencedirect.com/science/article/pii/S187705092100106X`. 5th International Conference on Computer Science and Computational Intelligence 2020.

[64] Hormuzd Katki, Stephanie Kovalchik, Christine Berg Li Cheung, and Anil Chaturvedi. Development and validation of risk models to select ever-smokers for ct lung cancer screening. *JAMA*, 2016.

[65] Hormuzd Katki, Stephanie Kovalchik, Lucia Petito, Li Cheung, Eric Jacobs, Ahmedin Jemal, Christine Berg, and Anil Chaturvedi. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. *Annals of Internal Medicine*, 2018.

[66] Nicholas Kissel and Lucas Mentch. Forward Stability and Model Path Selection. *arXiv preprint arXiv:2103.03462v1*, 2021.

[67] W A Knaus, E A Draper, D P Wagner, and J E Zimmerman. APACHE II: a severity of disease classification system. *Crit Care Med*, 13(10):818–829, October 1985.

[68] Katarzyna Kobylińska, Tomasz Mikołajczyk, Mariusz Adamek, Tadeusz Orłowski, and Przemysław Biecek. Explainable machine learning for modeling of early postoperative mortality in lung cancer. In Mar Marcos, Jose M. Juarez, Richard Lenz, Grzegorz J. Nalepa, Slawomir Nowaczyk, Mor Peleg, Jerzy Stefanowski, and Gregor Stiglic, editors, *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*, pages 161–174, Cham, 2019. Springer International Publishing. ISBN 978-3-030-37446-4.

[69] Katarzyna Kobylińska, Tadeusz Orłowski, Mariusz Adamek, and Przemysław Biecek. Explainable machine learning for lung cancer screening models. *Applied Sciences*, 12(4), 2022. ISSN 2076-3417. doi: 10.3390/app12041926. URL `https://www.mdpi.com/2076-3417/12/4/1926`.

[70] Katarzyna Kobylińska, Mateusz Krzyziński, Rafał Machowicz, Mariusz Adamek, and Przemysław Biecek. Exploration of the rashomon set assists trustworthy explanations for medical data, 2023.

[71] Simon Meyer Lauritsen, Mads Ruben Burgdorff Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *ArXiv*, abs/1912.01266, 2019.

[72] Junghwan Lee, Cong Liu, Junyoung Kim, Zhehuan Chen, Yingcheng Sun, James R. Rogers, Wendy K. Chung, and Chunhua Weng. Deep learning for rare disease: A scoping review. *Journal of Biomedical Informatics*, 135:104227, 2022. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2022.104227. URL `https://www.sciencedirect.com/science/article/pii/S1532046422002325`.

[73] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, 139:107539, 2023. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2022.107539. URL `https://www.sciencedirect.com/science/article/pii/S0747563222003594`.

[74] Anna Lemańska-Perek, Dorota Krzyżanowska-Gołąb, Tomasz Skalec, and Barbara Adamik. Plasma and cellular forms of fibronectin as prognostic markers in sepsis. *Mediators Inflamm*, 2020:8364247, August 2020.

[75] Anna Lemańska-Perek, Dorota Krzyżanowska-Gołąb, Katarzyna Kobylińska, Przemysław Biecek, Tomasz Skalec, Maciej Tyszko, Waldemar Gozdzik, and Barbara Adamik. Explainable artificial intelligence helps in understanding the effect of fibronectin on survival of sepsis. *Cells*, 11(15), 2022. ISSN 2073-4409. doi: 10.3390/cells11152433. URL `https://www.mdpi.com/2073-4409/11/15/2433`.

[76] Dan Liu, Longxiang Su, Gencheng Han, Peng Yan, and Lixin Xie. Prognostic value of procalcitonin in adult patients with sepsis: A systematic review and Meta-Analysis. *PLoS One*, 10(6):e0129450, June 2015.

[77] Alex John London. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent. Rep.*, 49(1):15–21, January 2019.

[78] Alex John London. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1):15–21, 2019. doi: https://doi.org/10.1002/hast.973.

[79] Peter ED Love, Weili Fang, Jane Matthews, Stuart Porter, Hanbin Luo, and Lieyun Ding. Explainable artificial intelligence: Precepts, methods, and opportunities for research in construction, 2023.

[80] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774. Curran Associates, Inc., 2017.

[81] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[82] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*, 2(10):749–760, October 2018.

[83] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2018.

[84] Marco Matarese, Francesco Rea, and Alessandra Sciutti. How much informative is your xai? a decision-making assessment task to objectively measure the goodness of explanations, 2023.

[85] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022.

[86] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018.

[87] C. Molnar. Interpretable machine learning. a guide for making black box models explainable. 2018. URL https://christophm.github.io/interpretable-ml-book/.

[88] Mieczysław Muraszkiewicz, Robert Nowak (informatyka), and Politechnika Warszawska Oficyna Wydawnicza. *Sztuczna inteligencja dla inżynierów: metody ogólne*. 2022.

[89] Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles, 2024.

[90] Daniel Nevo and Ya'acov Ritov. Identifying a Minimal Class of Models for High–dimensional Data. *Journal of Machine Learning Research*, 18(24):1–29, 2017.

[91] Favour Olaoye, Kaledio Potter, and Lucas Doris. Explainable ai: Interpreting and understanding machine learning models. *Journal of Machine Learning Research*, 03 2024.

[92] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016. ISBN 0553418815.

[93] G Sarah Power and David A Harrison. Why try to predict ICU outcomes? *Curr Opin Crit Care*, 20(5):544–549, October 2014.

[94] Vineet K Raghu, Wei Zhao, Jiantao Pu, Joseph K Leader, Renwei Wang, James Herman, Jian-Min Yuan, Panayiotis V Benos, and David O Wilson. Feasibility of lung cancer prediction from low-dose ct scan and smoking factors using causal models. *Thorax*, 74(7):643–649, 2019. ISSN 0040-6376. doi: 10.1136/thoraxjnl-2018-212638. URL https://thorax.bmj.com/content/74/7/643.

[95] Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu, and Prof. N. B. Venkateswarlu. A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, 3:101–114, 2011.

[96] M Reichsoellner, R B Raggam, J Wagner, R Krause, and M Hoenigl. Clinical evaluation of multiple inflammation biomarkers for diagnosis and prognosis for patients with systemic inflammatory response syndrome. *J Clin Microbiol*, 52 (11):4063–4066, September 2014.

[97] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[98] Greg Ridgeway. Generalized boosted models: A guide to the gbm package. 2006. URL `https://api.semanticscholar.org/CorpusID:12781809`.

[99] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistics Surveys*, 16:1 – 85, 2022.

[100] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Amazing things come from having many good models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[101] Guadalupe Ruiz Martín, José Prieto Prieto, Jorge Veiga de Cabo, Luisa Gomez Lus, José Barberán, Jose M González Landa, and Cristina Fernández. Plasma fibronectin as a marker of sepsis. *Int J Infect Dis*, 8(4):236–243, July 2004.

[102] Sergio Sanchez-Martinez, Oscar Camara, Gemma Piella, Maja Cikes, Miguel Ángel González-Ballester, Marius Miron, Alfredo Vellido, Emilia Gómez, Alan G Fraser, and Bart Bijnens. Machine learning for clinical decision-making: Challenges and opportunities in cardiovascular imaging. *Front. Cardiovasc. Med.*, 8: 765693, 2021.

[103] Ramesh Lal Sapra, Siddharth Mehrotra, Shailendra Lalwani, Vivek Mangla, Naimish Mehta, and Samiran Nundy. Predicting mortality following gastrointestinal surgery. *Current Medicine Research and Practice*, 8(1):8–12, 2018. ISSN 2352-0817. doi: https://doi.org/10.1016/j.cmrp.2017.12.001. URL `https://www.sciencedirect.com/science/article/pii/S2352081717301976`.

[104] Christian Scheeder, Florian Heigwer, and Michael Boutros. Machine learning and image-based profiling in drug discovery. *Current Opinion in Systems Biology*, 10:43–52, 2018. ISSN 2452-3100. doi: https://doi.org/10.1016/j.coisb.2018.05.004. URL `http://www.sciencedirect.com/science/article/pii/S2452310018300027`. Pharmacology and drug discovery.

[105] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the Existence of Simpler Machine Learning Models. In *Conference on Fairness, Accountability, and Transparency*. ACM, 2022.

[106] Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

[107] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. A survey on medical explainable ai (xai): Recent progress, explainability approach, human interaction and scoring system. *Sensors*, 22(20), 2022. ISSN 1424-8220. doi: 10.3390/s22208068. URL `https://www.mdpi.com/1424-8220/22/20/8068`.

[108] David W Shimabukuro, Christopher W Barton, Mitchell D Feldman, Samson J Mataraso, and Ritankar Das. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res*, 4(1):e000234, November 2017.

[109] Galit Shmueli. To Explain or to Predict? *Statistical Science*, 25(3):289 – 310, 2010.

[110] Manu Siddhartha, Paramita Maity, and Rajendra Nath. Explanatory artificial intelligence (xai) in the prediction of post-operative life expectancy in lung cancer patients. *International Journal of Scientific Research*, 2019.

[111] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, Richard S Hotchkiss, Mitchell M Levy, John C Marshall, Greg S Martin, Steven M Opal, Gordon D Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C Angus. The

third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, February 2016.

[112] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis, 2020.

[113] Tomasz Skalec, Barbara Adamik, Katarzyna Kobylinska, and Waldemar Gozdzik. Soluble urokinase-type plasminogen activator receptor levels as a predictor of kidney replacement therapy in septic patients with acute kidney injury: An observational study. *Journal of Clinical Medicine*, 11(6), 2022. ISSN 2077-0383. doi: 10.3390/jcm11061717. URL `https://www.mdpi.com/2077-0383/11/6/1717`.

[114] Gavin Smith, Roberto Mansilla, and James Goulding. Model Class Reliance for Random Forests. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

[115] Mateusz Staniak and Przemysław Biecek. Explanations of model predictions with live and breakdown packages. *The R Journal*, 10(2):395–409, 2018. doi: 10.32614/RJ-2018-072. URL `https://doi.org/10.32614/RJ-2018-072`.

[116] Martin C. Tammemägi. Selecting lung cancer screenees using risk prediction models—where do we go from here. *Translational Lung Cancer Research*, 7(3), 2018. ISSN 2226-4477. URL `http://tlcr.amegroups.com/article/view/21997`.

[117] Martin C. Tammemägi, Hormuzd A Katki, William G Hocking, Timothy R Church, Neil Caporaso, Paul A Kvale, Anil K Chaturvedi, Gerard A Silvestri, Tom L Riley, John Commins, and Christine D Berg. Selection criteria for lung-cancer screening. *The New England Journal of Medicine*, 2013.

[118] Ziqi Tang, Kangway V Chuang, Charles DeCarli, Lee-Way Jin, Laurel Beckett, Michael J Keiser, and Brittany N Dugger. Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.*, 10(1):2173, May 2019.

[119] Abdalrahman Tawhid, Tanya Teotia, and Haytham Elmiligi. Chapter 13 - machine learning for optimizing healthcare resources. In Pardeep Kumar, Yugal Kumar, and Mohamed A. Tawhid, editors, *Machine Learning, Big Data, and IoT for*

Medical Informatics, Intelligent Data-Centric Systems, pages 215–239. Academic Press, 2021. ISBN 978-0-12-821777-1. doi: https://doi.org/10.1016/B978-0-12-821777-1.00020-3. URL https://www.sciencedirect.com/science/article/pii/B9780128217771000203.

[120] Armin W. Thomas, Hauke R. Heekeren, Klaus-Robert Müller, and Wojciech Samek. Analyzing neuroimaging data through recurrent deep learning models, 2019.

[121] Hans-Christian Thorsen-Meyer, Annelaura B Nielsen, Anna P Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Kirstine Belling, Søren Brunak, and Anders Perner. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2020.

[122] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, nov 2021. doi: 10.1109/tnnls.2020.3027314. URL https://doi.org/10.1109%2Ftnnls.2020.3027314.

[123] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 359–380. PMLR, 09–10 Aug 2019. URL https://proceedings.mlr.press/v106/tonekaboni19a.html.

[124] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56, 2019. URL https://api.semanticscholar.org/CorpusID:57574615.

[125] Erik trumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18, 2010.

[126] Theja Tulabandhula and Cynthia Rudin. Robust Optimization using Machine Learning for Uncertainty Sets. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2014.

[127] Lior Turgeman, Jerrold H. May, and Roberta Sciulli. Insights from a machine learning model for predicting the hospital length of stay (los) at the time of admission. *Expert Systems with Applications*, 78:376–385, 2017. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.02.023. URL https://www.sciencedirect.com/science/article/pii/S0957417417301057.

[128] Heidi Vainio-Pekka, Mamia Ori otse Agbese, Marianna Jantunen, Ville Vakkuri, Tommi Mikkonen, Rebekah A. Rousi, and Pekka Abrahamsson. The role of explainable ai in the research field of ai ethics. *ACM Transactions on Interactive Intelligent Systems*, 13:1 – 39, 2023. URL https://api.semanticscholar.org/CorpusID:258990627.

[129] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102470. URL https://www.sciencedirect.com/science/article/pii/S1361841522001177.

[130] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. 41(3), 2014. ISSN 0219-1377. doi: 10.1007/s10115-013-0679-x. URL https://doi.org/10.1007/s10115-013-0679-x.

[131] Anna Waszczuk-Gajda, Jolanta Małyszko, David H Vesole, Magdalena Feliksbrot-Bratosiewicz, Kamila Skwierawska, Katarzyna Krzanowska, Katarzyna Kobylińska, Przemysław Biecek, Emilian Snarski, Anna Rodziewicz-Lurzyńska, Paweł Kozłowski, Agnieszka Stefaniak, Joanna Drozd-Sokołowska, Mateusz Ziarkiewicz, Pyush Vyas, Piotr Boguradzki, Krzysztof Mądry, Jarosław Biliński, Agnieszka Tomaszewska, Martyna Maciejewska, Elżbieta Urbanowska, Beata Blajer, Małgorzata Król, Maria Król, Hanna Zborowska, Artur Jurczyszyn, Jadwiga Dwilewicz-Trojaczek, Wieslaw W Jedrzejczak, and Grzegorz W Basak. Negative impact of borderline creatinine concentration and glomerular filtration

rate at baseline on the outcome of patients with multiple myeloma treated with autologous stem cell transplant. *Transplant. Proc.*, 52(7):2186–2192, September 2020.

[132] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson, editors. *The What-If Tool: Interactive Probing of Machine Learning Models*, 2019. URL `https://arxiv.org/pdf/1907.04135`.

[133] Rüdiger Wirth and Jochen Hipp. CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, volume 1, pages 29–40, 2000.

[134] Katarzyna Woźnica, Piotr Wilczyński, and Przemysław Biecek. SeFNet: Bridging Tabular Datasets with Semantic Feature Nets, 2023.

[135] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Elena Albu, Arshi, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, 369, 2020. doi: 10.1136/bmj.m1328.

[136] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang "Anthony" Chen. Chexplain. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr 2020. doi: 10.1145/3313831.3376807. URL `http://dx.doi.org/10.1145/3313831.3376807`.

[137] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the Whole Rashomon Set of Sparse Decision Trees. In *Advances in Neural Information Processing Systems*, pages 14071–14084. Curran Associates, Inc., 2022.

[138] Chang Ho Yoon, Robert Torrance, and Naomi Scheinerman. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, 48(9):581–585, 2022. ISSN 0306-6800. doi: 10.1136/medethics-2020-107102.

[139] Yu Zhang, Sadia Khalid, and Li Jiang. Diagnostic and predictive performance of biomarkers in patients with sepsis in an intensive care unit. *J Int Med Res*, 47 (1):44–58, November 2018.

[140] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang, and Min Xu. Respond-cam: Analyzing deep models for 3d imaging data by visualizations, 2018.

[141] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.

[142] Kun Zhu, Hongyuan Lin, Xichun Yang, Jiamiao Gong, Kang An, Zhe Zheng, and Jianfeng Hou. An in-hospital mortality risk model for elderly patients undergoing cardiac valvular surgery based on lasso-logistic regression and machine learning. *Journal of Cardiovascular Development and Disease*, 10(2), 2023. ISSN 2308-3425. doi: 10.3390/jcdd10020087. URL `https://www.mdpi.com/2308-3425/10/2/87`.

[143] Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1740–1747, 2018. doi: 10.1109/BIBM.2018.8621359.

[144] A. Zuin, G. Marulli, C. Breda, Schiavon M. Rebusso A. Bulf, R., F. Di Chiara, and F. Rea. Pneumonectomy for lung cancer over the age of 75 years: is it worthwhile? *Interact Cardiovasc Thorac Surg. 2010;10:931–935*, 2010.

[145] İbrahim Kök, Feyza Yıldırım Okay, Özgecan Muyanlı, and Suat Özdemir. Explainable artificial intelligence (xai) for internet of things: A survey. *IEEE Internet of Things Journal*, 10:14764–14779, 2022. URL `https://api.semanticscholar.org/CorpusID:249605318`.