Algorytmy uczenia maszynowego i telediagnostyka zaburzeń neurorozwojowych

Przegląd metod, wyników i ograniczeń zastosowań klinicznych, Mateusz Pomianek, wersja 2.0.0 | 2 listopada 2026 r.

Abstrakt. Rozwój psychoinformatyki umożliwił połączenie kwestionariuszy klinicznych, danych pozyskiwanych zdalnie i metod uczenia maszynowego w procesie wykrywania zaburzeń neurorozwojowych. Celem niniejszego opracowania jest krytyczna synteza sposobów pozyskiwania danych, projektowania modeli, interpretowania skuteczności i wdrażania rozwiązań telediagnostycznych dla zespołu nadpobudliwości psychoruchowej z deficytem uwagi (ADHD) oraz zaburzenia ze spektrum autyzmu (ASD). Wyjaśniono, dlaczego pomiar epizodyczny i pomiar cyfrowy odpowiadają na odmienne pytania, jakie ograniczenia wprowadzają samoopis i zjawiska kontekstowe oraz co rzeczywiście oznacza wysoka dokładność klasyfikatora. Szczególną uwagę poświęcono platformie MedicalBIT (n = 1688), badaniu kwestionariuszy ADHD (n = 342) i klasyfikacji ASD u dorosłych z zastosowaniem WEKA i SMOTE. Wyniki zestawiono z wymaganiami prawidłowej walidacji, kalibracji, oceny stronniczości i użyteczności klinicznej. Wniosek podstawowy jest następujący: modele predykcyjne mogą usprawniać przesiew, priorytetyzację skierowań i organizację oceny, ale trafność klasyfikacji wewnątrz zbioru danych nie stanowi dowodu samodzielnej wartości diagnostycznej. Najbezpieczniejsza architektura łączy analizę danych z kontrolą jakości, walidacją zewnętrzną i decyzją odpowiedzialnego specjalisty.

Słowa kluczowe: psychoinformatyka, ADHD, ASD, zaburzenia neurorozwojowe, telediagnostyka, MedicalBIT, uczenie maszynowe, fenotypowanie cyfrowe, walidacja kliniczna, SMOTE, SHAP.

1. Wprowadzenie: od opisu objawów do systemu wspierania decyzji

W rozpoznawaniu zaburzeń neurorozwojowych podstawową trudnością nie jest wyłącznie identyfikacja objawu. Ważniejsze bywa ustalenie jego trwałości, znaczenia funkcjonalnego, nasilenia w różnych środowiskach oraz związku z przebiegiem rozwoju. Jednorazowy wywiad umożliwia zebranie istotnych informacji, lecz przedstawia ograniczony obraz wielowymiarowego procesu. Dziecko może funkcjonować inaczej w domu niż w szkole; ten sam wzorzec aktywności może być interpretowany odmiennie przez rodzica i nauczyciela; zachowanie obserwowane w danym dniu nie zawsze reprezentuje typowy stan osoby badanej.

Cyfryzacja nie usuwa tych trudności, ale zmienia sposób ich badania. Kwestionariusz internetowy skraca drogę od zgłoszenia do pierwszej oceny, a urządzenia mobilne umożliwiają częstszy pomiar niektórych zachowań. Uczenie maszynowe (*machine learning*, ML) pozwala następnie identyfikować zależności między wieloma zmiennymi, które trudno opisać pojedynczą regułą kliniczną. Wspomniane możliwości stanowią istotę psychoinformatyki (*psychoinformatics*), rozumianej jako połączenie metod psychologicznych z informatycznym przetwarzaniem dużych, różnorodnych zbiorów danych [1], [2].

Należy jednak od razu rozróżnić **trzy poziomy zastosowania**: przesiew, wspomaganie oceny diagnostycznej i ustalenie rozpoznania klinicznego. Dodatni wynik przesiewowy oznacza potrzebę dalszej oceny, a nie automatyczne potwierdzenie ADHD lub ASD. Model uczenia maszynowego może przewidywać kategorię używaną w badaniu, lecz ostateczne rozpoznanie wymaga interpretacji klinicznej, różnicowania i oceny funkcjonowania. Zatarcie tych różnic prowadzi do nadmiernie optymistycznej interpretacji wyników algorytmicznych.

W opracowaniu najpierw scharakteryzowano problem pomiaru i źródła danych, następnie przeanalizowano mechanizmy uczenia oraz trzy konkretne rodziny badań, a dopiero później przedstawiono metody oceny, ograniczenia i proponowaną architekturę wdrożeniową. Taki porządek pozwala oddzielić pytanie „co mierzymy?" od pytania „jak model przewiduje?", a oba od pytania „czy model poprawia opiekę?".

2. Granice klasycznego pomiaru psychologicznego

2.1. Samoopis i ocena informatora

Kwestionariusze samoopisowe (*self-report*) oraz oceny rodziców i nauczycieli są ważnymi źródłami informacji, ale nie są neutralnymi rejestratorami zachowania. Osoba odpowiadająca musi sobie przypomnieć zdarzenie, zinterpretować je i przyporządkować do kategorii skali. Na każdym etapie może pojawić się odchylenie pomiarowe. Przy pytaniach o wielotygodniowe zachowania szczególnego znaczenia nabiera **błąd odtwarzania z pamięci** (*recall bias*). W przypadku aktywności cyfrowej trudno wiarygodnie oszacować liczbę sprawdzeń telefonu, czas korzystania z komunikatora lub wielokrotność krótkich sesji.

Drugim mechanizmem jest wpływ bieżącego stanu emocjonalnego i sytuacji społecznej. Zmęczenie, obniżony nastrój, stres lub sytuacja konfliktowa mogą wpływać na ocenę częstości i nasilenia objawów. Osoby badane mogą ponadto świadomie lub nieświadomie dostosowywać wypowiedzi do tego, co uważają za oczekiwane społecznie. Zjawisko to określa się jako **tendencję do odpowiedzi zgodnych z aprobatą społeczną** (*social desirability bias*). Ocena eksperta również ma ograniczenia: jest profesjonalną interpretacją, ale nie staje się z tego powodu pomiarem wolnym od błędu [1], [2].

W psychiatrii rozwojowej rozbieżności między informatorami nie muszą oznaczać, że jeden z nich się myli. Mogą odzwierciedlać rzeczywistą zależność zachowania od kontekstu. Jeżeli objaw występuje w szkole, lecz nie jest zgłaszany w domu, potrzebne jest rozpoznanie przyczyny rozbieżności, a nie mechaniczne uśrednienie ocen. Dlatego przy modelowaniu warto jawnie przechowywać źródło każdej obserwacji oraz mierzyć zgodność pomiędzy informatorami.

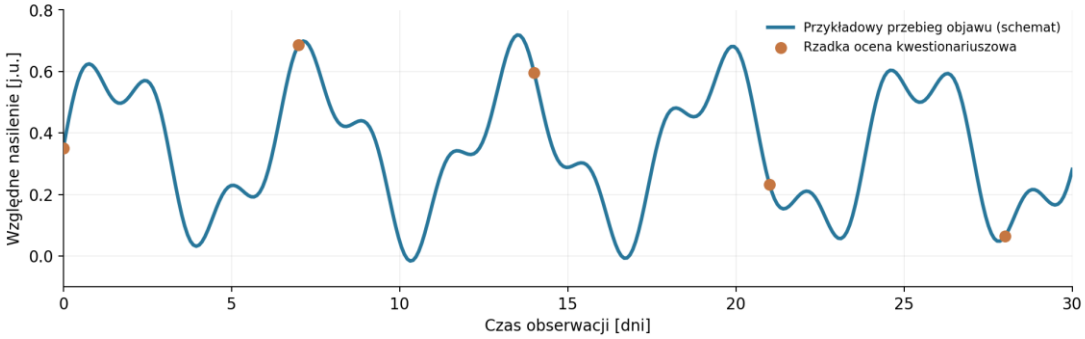
2.2. Problem rozdzielczości czasowej

W tradycyjnych procedurach informacje bywają zbierane w odstępach tygodniowych lub miesięcznych. Taka **niska rozdzielczość czasowa** (*coarse temporal granularity*) utrudnia wykrywanie krótkotrwałych zmian stanu i zależności zachowań od snu, obciążenia szkolnego czy sytuacji rodzinnej. Zwiększanie częstości kwestionariuszy ma jednak koszt: rośnie obciążenie uczestnika, a wielokrotne stosowanie podobnych narzędzi może wywoływać efekty ćwiczenia i pamięci. Dane rejestrowane pasywnie przez urządzenia mogą częściowo wypełnić tę lukę, ale nie zastępują treści psychologicznej wywiadu [1].



Rys. 1. Dwa komplementarne poziomy pomiaru i miejsce klinicznej interpretacji. Schemat autorski.

Rysunek pokazuje, dlaczego dwa zbiory danych mogą mieć różną częstość próbkowania, a mimo to opisywać ten sam proces. Wywiad może lepiej oddawać znaczenie zachowania, natomiast telemetria może lepiej rejestrować jego częstość. Najlepszy projekt badawczy nie uznaje z góry, że więcej obserwacji oznacza większą trafność konstruktu. Dodatkowe pomiary są użyteczne dopiero wtedy, gdy wiadomo, co reprezentują.



Krzywa ilustracyjna, nie przedstawia danych pacjentów ani wyników eksperymentu.

Rys. 2. Ilustracyjny przebieg zmiennego nasilenia objawu i rzadka obserwacja kwestionariuszowa

2.3. Pomiary chwilowe i interpretacja zmian w czasie

Rozwiązaniem pośrednim między rzadkim wywiadem a pełną telemetrią są krótkie oceny wykonywane w naturalnym środowisku życia, określane jako **chwilowa ocena ekologiczna** (*ecological momentary assessment*, EMA). Uczestnik może odpowiadać na kilka pytań w chwili pojawienia się zdarzenia lub o zaplanowanych porach. Takie rozwiązanie zmniejsza zależność od dalekiej retrospekcji, lecz wprowadza inny rodzaj ograniczeń: selektywne nieudzielanie odpowiedzi, reagowanie na pomiar, obciążenie użytkownika i przerwy w obserwacji. Należy porównywać nie tylko liczbę odpowiedzi, lecz także to, w jakich stanach uczestnik w ogóle zdecydował się je udzielić.

Dane podłużne tworzą hierarchiczną strukturę: wiele pomiarów jest zagnieżdżonych w jednej osobie, a osoby są zagnieżdżone w ośrodkach lub środowiskach. Analiza ignorująca tę strukturę może przeceniać informacyjną liczebność próby. Dziesięć tysięcy okien aktywności pochodzących od dziesięciu pacjentów nie stanowi dowodu równoważnego dziesięciu tysiącom niezależnych przypadków. Z tego względu w prognozowaniu należy oddzielać walidację międzyosobniczą od przewidywania przyszłych pomiarów tej samej osoby. Ten drugi problem może być użyteczny w monitorowaniu, ale odpowiada na inne pytanie naukowe niż rozpoznanie u nowego pacjenta.

Dodatkowym problemem jest **autokorelacja** (*autocorrelation*). Zachowania rejestrowane w kolejnych minutach lub godzinach zwykle nie są niezależne. Losowe przypisanie sąsiadujących okien do treningu i testu może prowadzić do zawyżonych wyników nawet bez świadomego kopiowania etykiety. W badaniu prognozy przyszłego nasilenia objawów należy stosować podział chronologiczny i ocenić, jak szybko spada skuteczność predykcji wraz z wydłużaniem horyzontu czasowego. W badaniu różnicującym osoby bardziej właściwy może być podział uczestników, ale w obu przypadkach trzeba jawnie określić jednostkę analizy.

Nie każda zmienna kliniczna powinna być monitorowana ciągle. Informacje o wczesnych etapach rozwoju, historii rodzinnej czy przebiegu ciąży nie zmieniają się co godzinę. Rozdzielenie cech względnie stałych od chwilowych jest więc niezbędne do sensownej architektury danych. Dane epizodyczne, retrospektywne i pasywne mogą być komplementarne, o ile system nie miesza czasu obserwacji, daty wpisania odpowiedzi i okresu, którego ona dotyczy.

2.4. Od zachowania cyfrowego do konstruktu psychologicznego

Sposób korzystania ze smartfona stanowi zapis interakcji z technologią, nie bezpośredni pomiar ADHD, ASD czy depresji. Krótsza aktywność komunikacyjna może wynikać ze spadku motywacji, ale także z urlopu, choroby somatycznej, wymiany telefonu, zmiany harmonogramu pracy albo ograniczenia dostępu do internetu. Analogicznie większa liczba błędów podczas pisania może zależeć od stresu, lecz również od urządzenia, języka, automatycznej korekty lub zmian sprawności motorycznej.

W literaturze opisuje się **fenotypowanie cyfrowe** (*digital phenotyping*), rozumiane tutaj jako ilościowy opis wzorców zachowania na podstawie danych generowanych podczas codziennego korzystania z urządzeń. Zmiennymi mogą być rytm aktywności, czas pierwszej i ostatniej interakcji, liczba sesji, okresy bezczynności, poziom ruchu oraz regularność kontaktów. Powiązanie tych miar z kategorią kliniczną wymaga odrębnej walidacji; korelacja nie świadczy o związku przyczynowym [1], [2].

Dla zaburzeń neurorozwojowych szczególnie istotna jest interpretacja podłużna. Fenotyp zmienia się wraz z wiekiem, środowiskiem i oddziaływaniami terapeutycznymi. Ten sam wskaźnik, np. liczba przełączeń aplikacji, może mieć różne znaczenie u dziecka, nastolatka i dorosłego. W praktyce należy więc weryfikować stabilność predyktorów oraz ich zależność od etapu rozwoju.

3. Źródła danych i proces telediagnostyczny

3.1. Platformy internetowe, kwestionariusze i historia rozwoju

W badaniach nad ADHD i ASD centralne miejsce zajmuje zbieranie danych opisujących historię rozwojową, funkcjonowanie rodzinne oraz aktualne trudności. Źródła mogą obejmować kwestionariusze rodzicielskie, oceny nauczycielskie, wywiad kliniczny, obserwację bezpośrednią, wcześniejszą dokumentację oraz wyniki badania poznawczego. Internetowa platforma MedicalBIT (*Medea Information and Clinical Assessment On-Line*) została opracowana do standaryzowanego pozyskiwania, przechowywania, przetwarzania i eksportowania danych klinicznych w psychopatologii dzieci i młodzieży [3].

Istotnym walorem platformy jest rozdzielenie dwóch etapów. Informacje społeczno-anamnestyczne gromadzi się przed wizytą, natomiast kategorie diagnostyczne zostają wprowadzone przez personel kliniczny po ocenie. Dzięki temu dane wejściowe modelu nie muszą pochodzić bezpośrednio z tego samego pola formularza, które zawiera diagnozę. Nie gwarantuje to jednak braku przecieku informacji: procedury budowy zbioru danych trzeba kontrolować także pod kątem zmiennych pośrednio zależnych od procesu diagnostycznego [3], [4].

W badaniu Grazioli i współautorów z 2024 r. analizowano 1688 dzieci i nastolatków kierowanych z podejrzeniem zaburzeń neurorozwojowych. Wykorzystano dane dotyczące płci, czynników przed- i okołoporodowych, obaw rozwojowych, obciążenia rodzinnego i innych informacji podawanych przez opiekunów. Wyjściowe etykiety ADHD i ASD powiązano z ustaleniami klinicznymi. Badanie miało charakter retrospektywny i jednoośrodkowy, co istotnie ogranicza możliwość prostego przenoszenia wyników na inne populacje [4].

3.2. Dane pasywne i ich odmienne znaczenie

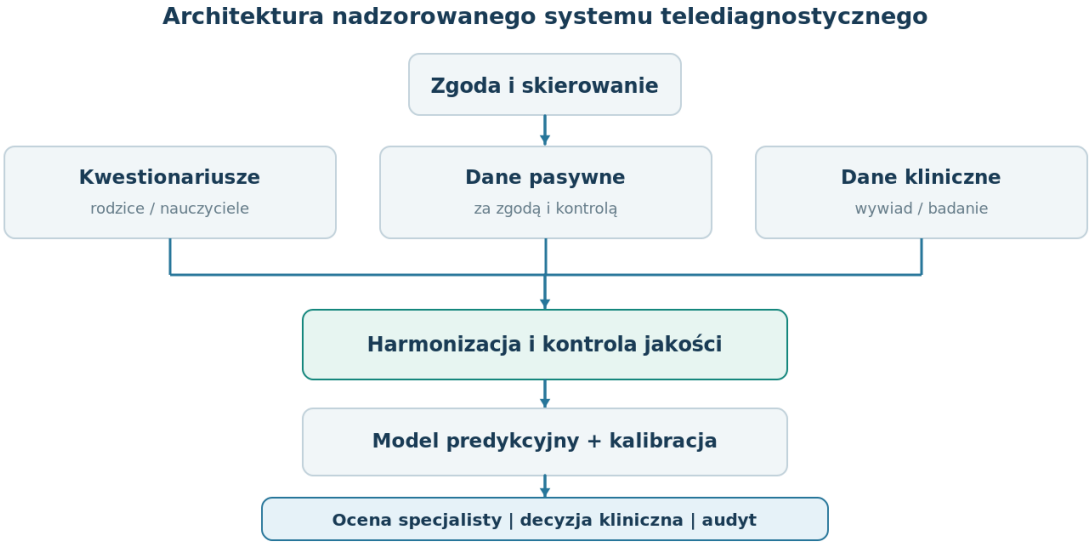
Telemetria urządzeń może uzupełniać kwestionariusze o wskaźniki czasu, aktywności i rytmu zachowania. Przykłady obejmują liczbę krótkich sesji korzystania z telefonu, czas aktywności ekranu, regularność snu wnioskowaną pośrednio z aktywności urządzenia oraz ogólny poziom przemieszczania się. Czujniki ubieralne (*wearable sensors*) wprowadzają kolejne kanały, np. przyspieszenie i aktywność ruchową. Źródła

te mają różne profile błędu: GPS zależy od dostępności sygnału, a pomiar aktywności aplikacji od systemu operacyjnego i sposobu agregacji.

W opracowaniu naukowym należy oddzielać pomiar bezpośredni od cechy pochodnej. Liczba interakcji ze smartfonem to obserwacja techniczna. „Wycofanie społeczne" jest hipotezą interpretacyjną wymagającą odniesienia do kontekstu i innych pomiarów. Rzetelny system powinien umożliwiać prześledzenie, które obserwacje zostały wykorzystane do danej predykcji i jakie transformacje danych zastosowano. Nie jest dopuszczalne przedstawianie wnioskowania psychologicznego jako faktu rejestrowanego przez czujnik.

3.3. Integracja danych z różnych źródeł

Dane heterogeniczne wymagają harmonizacji. Obejmuje ona wspólną identyfikację uczestnika, wersjonowanie kwestionariuszy, obsługę braków danych, kontrolę czasu pomiaru i rozdzielanie informacji dostępnych przed decyzją od informacji powstających po tej decyzji. Braki mogą mieć znaczenie kliniczne: nieukończony kwestionariusz, nieobecność nauczyciela lub brak danych z urządzenia nie powinny być automatycznie interpretowane jako brak objawów.



Rys. 3. Referencyjny przepływ danych i decyzji w nadzorowanej architekturze telediagnostycznej. Schemat autorski.

Zaproponowany schemat ma charakter architektury referencyjnej, a nie opisu gotowego produktu MedicalBIT. Zakłada wyraźną granicę między analizą predykcyjną a decyzją kliniczną. Wynik algorytmu stanowi informację pomocniczą, która powinna być uzupełniona o poziom niepewności, zakres stosowalności modelu i wskazanie danych wejściowych.

4. Jak klasyfikatory uczą się rozpoznawać wzorce

4.1. Modele liniowe i drzewa decyzyjne

Regresja logistyczna (*logistic regression*, LR) modeluje prawdopodobieństwo przynależności do klasy na podstawie liniowej kombinacji cech. Jej zaletą jest przejrzysty punkt odniesienia: współczynniki można analizować w kategoriach ilorazów szans, o ile zmienne są poprawnie zakodowane i uwzględniono niezbędne założenia. LR nie jest jednak „z definicji prostym modelem klinicznym"; interpretacja współczynnika zależy od współzmiennych, doboru próby i sposobu selekcji danych.

Drzewo decyzyjne (*decision tree*, DT) buduje kolejno warunki podziału. W praktyce pozwala prześledzić sekwencję pytań prowadzących do klasyfikacji. Gdy pierwszy węzeł dotyczy wyniku skali ADHD, oznacza to, że w danym zbiorze i sposobie uczenia cecha ta najlepiej ograniczała niejednorodność grup na pierwszym

etapie. Nie oznacza to jednak, że jest najważniejszą przyczyną zaburzenia. Drzewa mogą być niestabilne: niewielka zmiana zbioru danych potrafi zmienić strukturę podziałów.

Las losowy (*random forest*, RF) uśrednia lub agreguje decyzje wielu drzew uczonych na losowo wybieranych próbach i podzbiorach cech. To często zmniejsza wariancję pojedynczego drzewa i umożliwia modelowanie nieliniowych interakcji. Ceną jest trudniejsza interpretacja: predykcja zbioru drzew nie sprowadza się do jednej reguły. W badaniach neurorozwojowych las losowy bywa skuteczny, gdy zależności między zmiennymi anamnestycznymi nie są liniowe [4].

4.2. Klasyfikacja wektorów nośnych i pozostałe metody

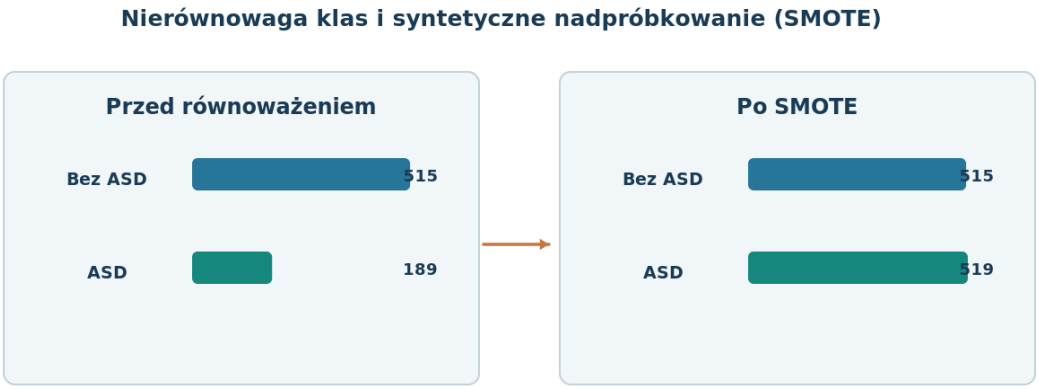
Maszyna wektorów nośnych (*support vector machine*, SVM) uczy granicę oddzielającą klasy w przestrzeni cech, ewentualnie po przekształceniu jądrowym. **Sekwencyjna optymalizacja minimalna** (*sequential minimal optimization*, SMO) nie jest osobną rodziną diagnozowania, lecz algorytmem optymalizacji problemu uczenia SVM. To rozróżnienie jest ważne przy interpretowaniu wyników badań WEKA: klasyfikator oparty na SMO należy porównywać w kontekście implementacji SVM, nie jako abstrakcyjny mechanizm psychometryczny.

Naiwny klasyfikator Bayesa (*naïve Bayes*) wykorzystuje prawdopodobieństwa warunkowe cech i przyjmuje uproszczone założenie ich niezależności przy zadanej klasie. Mimo tego przybliżenia potrafi działać sprawnie na małych i uporządkowanych zbiorach. **IBK** to implementacja metody najbliższych sąsiadów (*k-nearest neighbors*, k-NN) w środowisku WEKA. W tym przypadku etykieta nowej osoby zależy od podobieństwa do wcześniej zaobserwowanych rekordów, dlatego krytyczne znaczenie ma sposób normalizacji, doboru cech i odległości [6].

4.3. Nierównowaga klas, SMOTE i ryzyko przecieku

W zbiorach badań przesiewowych liczebności klas często znacznie się różnią. Jeśli grupa bez rozpoznania jest dominująca, model może osiągać pozornie wysoką dokładność, często przewidując klasę większościową. **Syntetyczne nadpróbkowanie klasy mniejszościowej** (*Synthetic Minority Over-sampling Technique*, SMOTE) tworzy dodatkowe punkty na podstawie obserwacji mniejszościowych, często przez interpolację między ich sąsiadami [8]. Takie punkty są konstrukcjami matematycznymi, a nie nowymi, niezależnymi osobami.

Zasada poprawnej walidacji brzmi: podział na zbiory musi nastąpić **przed** nadpróbkowaniem, imputacją zależną od danych, wyborem cech i dostrajaniem hiperparametrów. W każdej iteracji walidacji krzyżowej SMOTE należy dopasowywać tylko do podzbioru treningowego. W przeciwnym razie próbki syntetyczne mogą przekazywać informację z części testowej do uczącej i zawyżać estymowaną skuteczność. Nie oznacza to, że w każdym opublikowanym badaniu błąd na pewno wystąpił; oznacza potrzebę audytu kolejności operacji [8], [11].



Rys. 4. Zmiana liczebności klas w badaniu ASD dorosłych po zastosowaniu SMOTE

4.4. Interpretowalność, predyktory i przyczynowość

Dla modeli złożonych stosuje się techniki wyjaśniania predykcji, m.in. SHAP (*SHapley Additive exPlanations*) [9]. Pozwalają one przypisać cechom wkład w konkretną predykcję względem przyjętego tła odniesienia. W badaniu MedicalBIT wśród ważnych zmiennych znalazły się płeć, ryzyka przed- i okołoporodowe, obawy rozwojowe i dodatni wywiad rodzinny [4]. Istotność predykcyjna nie jest jednak równoznaczna z przyczynową rolą zmiennej. Przykładowo płeć może częściowo reprezentować różnice w częstości rozpoznawania, sposobie kierowania do diagnostyki lub strukturze próby.

Wnioski z analizy SHAP powinny być uzupełnione o ocenę stabilności w podgrupach, wpływ współzależności cech oraz ryzyko niesprawiedliwego działania modelu. Nie należy usuwać potencjalnych czynników stronniczości wyłącznie po to, by uzyskać pozornie „neutralny" algorytm, bez sprawdzenia konsekwencji dla czułości i swoistości w poszczególnych grupach.

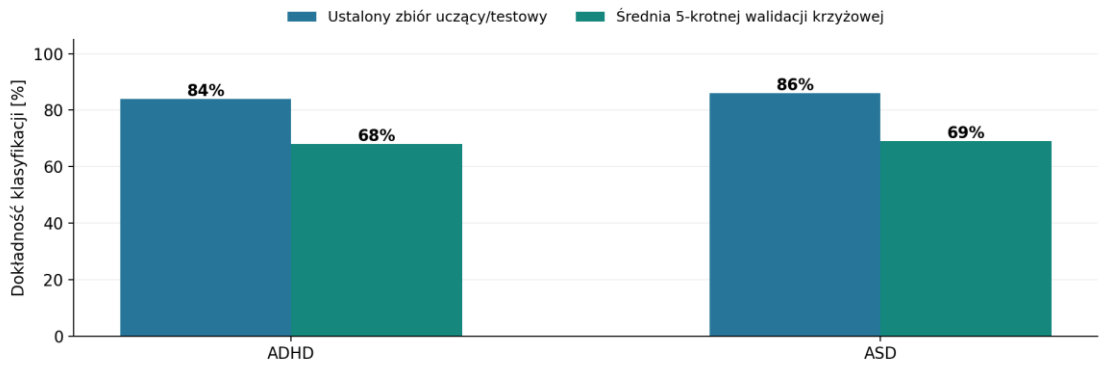
5. Trzy studia przypadków: co rzeczywiście pokazują dane

5.1. MedicalBIT: predykcja na podstawie wywiadu społeczno-anamnestycznego

W retrospektywnym badaniu Grazioli i współautorów analizowano 1688 dzieci i nastolatków skierowanych do diagnostyki zaburzeń neurorozwojowych. Wśród uczestników zidentyfikowano 269 osób z rozpoznaniem ADHD i 270 osób z ASD. Badacze wykorzystali rodzicielskie dane społeczno-anamnestyczne z platformy MedicalBIT i porównali regresję logistyczną, drzewo decyzyjne oraz las losowy [4]. Punkt wyjścia stanowił rzeczywisty kontekst kierowania do specjalistycznego ośrodka, a nie przypadkowa próba populacyjna; to ważne dla interpretacji przyszłego zastosowania modelu.

W ustalonym podziale danych las losowy uzyskał dokładność 84% dla ADHD oraz 86% dla ASD. Dla ADHD czułość wynosiła 93%, a swoistość 75%; dla ASD odpowiednio 95% i 77%. Drzewa decyzyjne osiągnęły dokładności 74% i 79%, natomiast modele regresji logistycznej 61% i 63%. Wyniki pokazują, że w konkretnym zbiorze nieliniowy model zespołowy lepiej uchwycił strukturę informacji niż zastosowane rozwiązania prostsze [4]. Nie wynika z tego jednak ogólna przewaga lasów losowych nad wszystkimi metodami w każdym ośrodku.

Najważniejszy szczegół, łatwy do przeoczenia: autorzy przedstawili także średnie wyniki pięciokrotnej walidacji krzyżowej. Dla lasu losowego wyniosły one 68% (ADHD) oraz 69% (ASD). Odpowiednie średnie czułości wyniosły 73% i 70%, a swoistości 63% i 68%. Wyniki te nie są tym samym oszacowaniem, co 84% i 86% podane dla ustalonego podziału. Rozbieżność pokazuje, jak silnie rezultat może zależeć od schematu oceny i składu prób. Różnica nie jest dowodem celowego zawyżania wyników, ale uniemożliwia przedstawianie jednej liczby jako trwałej „skuteczności" całej metody [4].



Rys. 5. Dokładność lasu losowego w dwóch procedurach oceny MedicalBIT: ustalony podział oraz średnia z pięciu iteracji walidacji krzyżowej. Dane z [4].

Dane dotyczące płci stanowiły istotną zmienną predykcyjną, szczególnie przy ADHD. W analizie ASD ważną rolę przypisano również ryzyku przed- i okołoporodowemu. Nie należy jednak utożsamiać rankingu cech z listą etiologicznych przyczyn ADHD i ASD. Badany zbiór odzwierciedla także praktykę skierowań i diagnostyki w konkretnym ośrodku. Jeżeli model ma być wykorzystany poza tą praktyką, potrzebna jest walidacja na odrębnych danych, z uwzględnieniem grup wieku, płci i innych istotnych charakterystyk [4].

5.2. ADHD: zgodność decyzji modelu z rozpoznaniem specjalisty

W osobnym badaniu dotyczącym telediagnostyki ADHD oceniono 342 dzieci w wieku 3–16 lat skierowanych z podejrzeniem tego zaburzenia. Zebrano zdalne skale wypełniane przez rodziców i nauczycieli oraz wyniki elementów oceny przeprowadzanej stacjonarnie. Autorzy sprawdzili, czy algorytm drzewa decyzyjnego może odtworzyć reguły klasyfikacji stosowane przez klinicystów i jaka jest rola współwystępujących objawów ASD [5].

Model drzewa osiągnął 82% dokładności w analizie decyzji klinicznych, lecz w procedurze walidacji krzyżowej jego dokładność predykcyjna wyniosła 74%. Porównywano go również z lasem losowym i maszyną wektorów nośnych. Najsilniej różnicujące okazały się oceny nasilenia podstawowych objawów ADHD dokonywane przez opiekunów, a objawy ASD mogły zakłócać oszacowanie nasilenia ADHD [5]. Te rezultaty są istotne z dwóch powodów. Po pierwsze, wskazują, że informacja z wywiadu i kwestionariuszy może przyspieszyć wstępną identyfikację ryzyka. Po drugie, ujawniają ograniczenia prostej reguły diagnostycznej w warunkach współwystępowania cech wielu zaburzeń.

Warto oddzielić **zgodność z decyzją kliniczną** od **trafności rozpoznania w sensie niezależnego standardu referencyjnego**. Algorytm uczony na etykietach klinicystów odtwarza wzorzec tych etykiet, a więc może również odziedziczyć ich niejednoznaczność. Wysoka zgodność z decyzjami ekspertów nie dowodzi, że model rozpoznaje odrębny mechanizm biologiczny ADHD. Najsilniej uzasadnionym zastosowaniem jest wsparcie procesu oceny, a nie jego zastąpienie.

5.3. ASD u dorosłych: bardzo wysoka dokładność a ograniczenia danych

W pracy Jaffera i współautorów dotyczącej ASD u dorosłych porównano cztery klasyfikatory: regresję logistyczną, SVM uczony metodą SMO, naiwny Bayes oraz IBK. Zastosowano środowisko WEKA i dane związane z kwestionariuszem przesiewowym ASD dla dorosłych. Wyjściowy zbiór obejmował 515 osób w grupie kontrolnej i 189 przypadków ASD, a następnie zastosowano SMOTE w celu zrównoważenia liczebności klas [6].

Najwyższy wynik osiągnęła metoda SMO: 99,71% dokładności. W opracowaniu raportowano także 100% czułości i 99% swoistości. Pozostałe algorytmy uzyskały 98,74% (naiwny Bayes), 98,16% (regresja

logistyczna) i 97,39% (IBK). W tabeli pomyłek SMO przedstawiono 515 poprawnie sklasyfikowanych osób z grupy kontrolnej oraz 516 z 519 rekordów klasy ASD po równoważeniu. Liczba 519 nie oznacza liczby niezależnych uczestników z ASD w materiale wyjściowym: obejmuje również rekordy wynikające z nadpróbkowania [6].

Ten przykład jest wyjątkowo użyteczny dydaktycznie. Wynik bliski 100% może wyglądać jak rozwiązanie problemu diagnostycznego, lecz o jego znaczeniu rozstrzyga pochodzenie danych, sposób powstania etykiet oraz prawidłowość walidacji. Kwestionariusz przesiewowy, na którego odpowiedziach oparta jest kategoria docelowa, może zawierać cechy bardzo blisko związane z definicją samej etykiety. Dodatkowo nadpróbkowane rekordy nie reprezentują zwiększenia populacji niezależnych pacjentów. Nawet przy poprawnym przeprowadzeniu wewnętrznej walidacji pozostaje pytanie, czy model osiągnąłby zbliżoną trafność na innych osobach, w nowych warunkach klinicznych i dla trudnych przypadków różnicowych.

Uczciwa interpretacja brzmi więc: praca pokazuje, że różne klasyfikatory skutecznie rozdzielały analizowane rekordy w konkretnym zbiorze po przyjętych transformacjach. Nie wykazuje natomiast gotowości modelu do samodzielnego rozpoznawania ASD w populacji ogólnej. To rozróżnienie jest zgodne z późniejszą metaanalizą badań AI w wykrywaniu ASD, w której podkreślono ograniczoną walidację zewnętrzną oraz niepewność uogólnienia [12].

5.4. Syntetyczne podsumowanie porównywanych badań

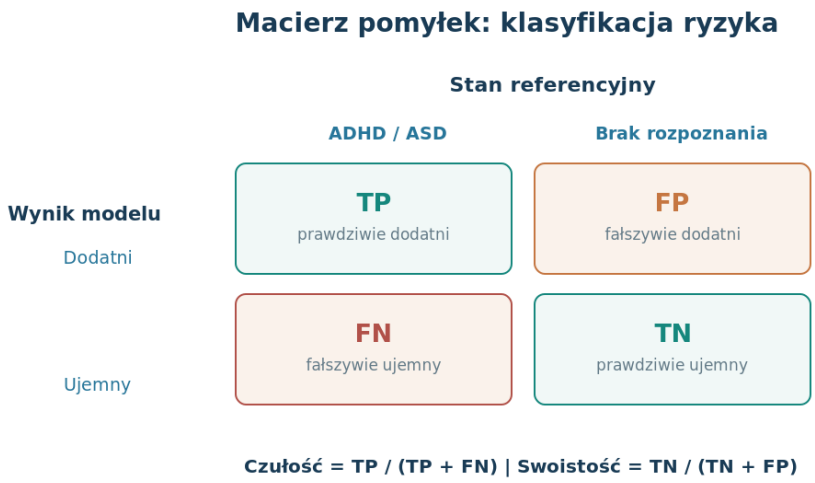
Badanie / populacja	Dane i algorytm	Wynik raportowany	Interpretacja i ograniczenie
MedicalBIT, dzieci i młodzież, n = 1688 [4]	Kwestionariusze społeczno-anamnestyczne; RF	84% ADHD; 86% ASD na ustalonym podziale	Średnio 68% i 69% w 5-krotnej walidacji; badanie jednoośrodkowe
Zdalne kwestionariusze ADHD, n = 342 [5]	Oceny rodziców/nauczycieli; DT	82% zgodności w analizie; 74% w walidacji	Objawy ASD komplikują ocenę nasilenia ADHD
Przesiew ASD u dorosłych, pierwotnie n = 704 [6]	WEKA, SMO/SVM; dane po SMOTE	99,71% na użytym zbiorze	Rekordy syntetyczne i charakter kwestionariusza ograniczają wnioskowanie kliniczne

Bez ujednolicenia definicji zadania, populacji, podziału i punktu odniesienia nie wolno ustawiać tych wartości w jeden ranking „najlepszych algorytmów”. Są to wyniki **różnych eksperymentów**, a nie trzech modeli ocenionych na tych samych osobach.

6. Miary skuteczności: co pokazują liczby, a czego nie pokazują

6.1. Macierz pomyłek

Każda ocena klasyfikatora powinna rozpoczynać się od czterech liczebności: wynik prawdziwie dodatni (TP), prawdziwie ujemny (TN), fałszywie dodatni (FP) i fałszywie ujemny (FN). Dodatni wynik może oznaczać podejrzenie ADHD lub ASD, ale musi być jednoznacznie zdefiniowany. W psychometrii klinicznej istotna jest również jakość standardu referencyjnego, wobec którego przypisano te cztery kategorie.



Rys. 6. Macierz pomyłek i dwa podstawowe wskaźniki wydajności. Schemat autorski.

Dokładność klasyfikacji (*accuracy*, Acc) opisuje odsetek wszystkich poprawnych odpowiedzi modelu:

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

Czułość diagnostyczna (*sensitivity*, *recall*, Se) określa udział prawidłowo wykrytych przypadków pozytywnych:

$$\mathrm{Se} = \frac{TP}{TP + FN}$$

Swoistość diagnostyczna (*specificity*, Sp) określa udział poprawnie odrzuconych przypadków negatywnych:

$$\mathrm{Sp} = \frac{TN}{TN + FP}$$

Należy ją bezwzględnie odróżnić od **precyzji predykcji** (*precision*) lub dodatniej wartości predykcyjnej (*positive predictive value*, PPV):

$$\mathrm{PPV} = \frac{TP}{TP + FP}, \qquad \mathrm{NPV} = \frac{TN}{TN + FN}$$

Swoistość i precision nie są synonimami. Pierwsza dotyczy wszystkich przypadków bez zaburzenia, druga wszystkich dodatnich wskazań modelu. Ich utożsamienie prowadzi do błędnych wniosków szczególnie przy niskim rozpowszechnieniu klasy dodatniej.

6.2. Wpływ częstości występowania zaburzenia na wartość dodatniego wyniku

Wartość predykcyjna testu zależy nie tylko od jego czułości i swoistości, ale również od częstości występowania danego stanu w populacji, oznaczanej przez π. Zależność ta ma postać:

$$\mathrm{PPV} = \frac{\mathrm{Se} \cdot \pi}{\mathrm{Se} \cdot \pi + (1 - \mathrm{Sp})(1 - \pi)}$$

Przy czysto ilustracyjnym założeniu czułości 93%, swoistości 75% i rozpowszechnienia 10% otrzymujemy PPV około 29%. Oznacza to, że w tak zdefiniowanej populacji znaczna część dodatnich sygnałów byłaby fałszywie dodatnia pomimo wysokiej czułości. Nie jest to przeliczenie rzeczywistego wyniku MedicalBIT na częstość ADHD w populacji; to **przykład pokazujący mechanizm zależności od rozpowszechnienia**.

Przenoszenie modelu z poradni specjalistycznej do przesiewu populacyjnego bez rekalkulacji i nowej walidacji mogłoby istotnie zmienić PPV.

6.3. Kalibracja, przedziały ufności i koszty błędów

Zdolność rozdzielania dwóch grup (*discrimination*) nie oznacza poprawnego oszacowania prawdopodobieństwa. Model może dobrze porządkować uczestników od niższego do wyższego ryzyka, a jednocześnie stale zawyżać deklarowane prawdopodobieństwa. Dlatego poza accuracy i AUC warto raportować **kalibrację** (*calibration*), np. wykres obserwowanego odsetka wyników dodatnich względem przewidywanego ryzyka i współczynnik Briera. Podane w publikacji przedziały ufności odnoszą się do konkretnej procedury estymacji, a nie automatycznie do przyszłego wdrożenia [4], [10], [11].

W zastosowaniu przesiewowym pominięcie osoby wymagającej diagnostyki (FN) i niepotrzebne skierowanie do diagnostyki (FP) mają różne koszty. Próg decyzyjny powinien wynikać z celu klinicznego, dostępności świadczeń i akceptowalnego poziomu ryzyka. Nie należy go dobierać wyłącznie tak, aby zmaksymalizować accuracy na posiadanym zbiorze.

7. Trafność, uogólnialność i pułapki projektowania eksperymentu

7.1. Dlaczego wyniki maleją po zmianie schematu walidacji

Nawet gdy model nie zawiera formalnego błędu implementacyjnego, jego skuteczność może zależeć od doboru próby. Podczas uczenia wykorzystuje on zarówno rzeczywiste związki między cechami a etykietą, jak i przypadkowe regularności charakterystyczne dla danego zbioru. Ocena na pojedynczym, korzystnym podziale danych może dawać bardziej optymistyczny wynik niż średnia z powtarzanych podziałów. Różnica między wynikami MedicalBIT w ustalonym podziale i pięciokrotnej walidacji jest dobrym przykładem zależności od procedury oceny [4].

Walidacja krzyżowa (*cross-validation*) ogranicza arbitralność pojedynczego podziału, ale nadal jest oceną **wewnętrzną**. Jeśli wszystkie rekordy pochodzą z jednej placówki i jednego procesu rekrutacji, walidacja krzyżowa nie sprawdza, czy model działa w innym kraju, innej grupie wieku albo w poradni o innej strukturze pacjentów.

7.2. Przeciek informacji i zależności w zbiorze danych

Przeciek danych (*data leakage*) występuje wtedy, gdy algorytm podczas uczenia uzyskuje dostęp do informacji, które w rzeczywistym momencie predykcji nie byłyby dostępne. Może przyjmować postać wykorzystania zmiennej powstałej po diagnozie, normalizacji całego zbioru przed podziałem, nadpróbkowania przed walidacją lub umieszczenia rekordów tej samej osoby zarówno w zbiorze treningowym, jak i testowym. W badaniach longitudinalnych należy dzielić dane według osoby, a często również czasu i ośrodka, zamiast losować pojedyncze okna pomiarowe.

Osobny problem stanowi **powielanie konstruktu etykiety**. Jeżeli rozpoznanie w zbiorze zostało ustalone przede wszystkim na podstawie konkretnego kwestionariusza, a model otrzymuje te same odpowiedzi, wysoka zgodność może w znacznym stopniu oznaczać odtwarzanie reguły skali. Taki wynik może być użyteczny technicznie, lecz nie świadczy o odkryciu nowego wskaźnika biologicznego. W każdym projekcie należy jasno opisać, jak wyznaczono kategorię referencyjną i czy istnieje niezależne potwierdzenie kliniczne [5], [6].

7.3. Walidacja zewnętrzna i użyteczność kliniczna

Walidacja zewnętrzna (*external validation*) polega na ocenie ustalonego modelu w nowych danych, które nie uczestniczyły w dopasowywaniu, selekcji cech ani wyborze progu. Może dotyczyć innej placówki, regionu, okresu lub populacji. Jej celem jest oszacowanie transportowalności modelu. Dopiero następny etap odpowiada na pytanie, czy system rzeczywiście poprawia decyzje, czas oczekiwania, obciążenie specjalistów albo dostępność świadczeń.



Rys. 7. Kolejne poziomy oceny modelu predykcyjnego w telediagnostyce. Schemat autorski.

Rzetelne sprawozdanie powinno zawierać charakterystykę populacji, dostępność cech w momencie decyzji, opis braków, schemat walidacji, metryki, przedziały ufności i analizę wyników w podgrupach. Przydatnym standardem raportowania jest TRIPOD+AI, a do oceny jakości i ryzyka błędu systematycznego służy PROBAST+AI [10], [11]. Te wytyczne nie zastępują badań klinicznych, lecz pomagają odróżnić technicznie poprawnie wykonany eksperyment od wyniku gotowego do przeniesienia do opieki zdrowotnej.

7.4. Co mówi syntetyczna ocena badań nad ASD

Metaanaliza Wanga i współautorów z 2026 r. objęła 26 badań wykrywania ASD z wykorzystaniem AI. Autorzy uzyskali łączną estymację czułości około 0,91 i swoistości około 0,87, ale po ograniczeniu zależności między zbiorami wynik spadał. Zwrócili uwagę na dużą heterogeniczność, dominację badań pediatrycznych, przewagę walidacji wewnętrznej i niewystarczającą liczbę niezależnych ocen zewnętrznych [12]. Ta obserwacja jest zgodna z krytyczną interpretacją wyników badań pojedynczych: wysokie wartości raportowane na danych rozwojowych nie są równoznaczne ze sprawdzoną skutecznością systemu w praktyce.

7.5. Interpretacja podgrup i ryzyko niesprawiedliwych decyzji

Model zbudowany na danych z poradni może odtwarzać nie tylko zależności między objawami a rozpoznaniem, lecz także nierówności w kierowaniu do specjalisty. Jeśli pewna grupa trafia do diagnostyki dopiero przy wyraźnie silniejszych objawach, algorytm może nauczyć się innego wzorca ryzyka niż ten obserwowany w grupie kierowanej wcześniej. Takie efekty nie są prostą konsekwencją „błędnej sztucznej inteligencji". Powstają na przecięciu danych, organizacji opieki i procesu społecznego.

Ocena sprawiedliwości (*fairness*) wymaga analizy działania w podgrupach, lecz same średnie wyniki nie wystarczą. Warto badać osobno czułość, swoistość i kalibrację w grupach wiekowych, płciowych oraz w ośrodkach o odmiennym profilu pacjentów. Nawet identyczna czułość może prowadzić do różnej wartości PPV, jeżeli różni się częstość występowania etykiety. Porównania powinny zawierać przedziały ufności, ponieważ niewielkie podgrupy mogą generować niestabilne estymacje. Zbyt daleko idąca interpretacja

kilkunastu przypadków w podgrupie byłaby metodologicznie równie problematyczna jak całkowite zignorowanie różnic.

W analizie zmiennej płci ważne jest także rozróżnienie pomiędzy predykcją a wyjaśnianiem. Wynik SHAP sugerujący duży wpływ danej zmiennej może wynikać z epidemiologii, przebiegu kierowania do diagnostyki, korelacji z innymi cechami albo struktury próby. Nie jest automatycznie argumentem, aby używać płci jako samodzielnej reguły klinicznej. Poprawna decyzja projektowa wymaga zbadania, czy obecność tej zmiennej poprawia identyfikację rzeczywiście wymagających wsparcia osób oraz czy nie obniża skuteczności w grupach rzadziej rozpoznawanych.

Praktycznie przydatna jest analiza zróżnicowania błędów po zmianie progu decyzyjnego. Jeden globalny próg może zwiększać liczbę pominięć w niektórych grupach; kilka progów może z kolei rodzić inne trudności w uzasadnieniu i komunikowaniu decyzji. Takie dylematy nie mają uniwersalnej odpowiedzi algorytmicznej. Wymagają uprzedniego zdefiniowania celu świadczenia, sposobu nadzoru i kryteriów dopuszczalnego ryzyka. PROBAST+AI przewiduje analizę możliwości zastosowania oraz ryzyka stronniczości na poziomie uczestników, predyktorów, wyniku i analizy [11].

7.6. Utrzymanie modelu w warunkach zmiany praktyki klinicznej

Nawet poprawnie zwalidowany model może z czasem tracić przydatność. Zmienia się praktyka kierowania pacjentów, populacja użytkowników, wersja kwestionariusza, sprzęt pomiarowy i sposób kodowania rozpoznań. Zmiany rozkładu danych określa się ogólnie jako **przesunięcie rozkładu** (*distribution shift*). Występuje też sytuacja bardziej subtelna: identyczny zestaw cech otrzymuje inne znaczenie kliniczne wskutek zmiany praktyk diagnostycznych, co wpływa na zależność między cechami a etykietą.

Nadzór powinien obejmować odsetek brakujących pól, rozkład zmiennych wejściowych, udział osób spoza wcześniej badanych grup oraz miary jakości tam, gdzie dostępny jest niezależny wynik kliniczny. Nie należy jednak automatycznie aktualizować modelu po każdej zmianie średniej wartości cechy. Niektóre zmiany wynikają z sezonowości, różnic w harmonogramie szkoły albo planowych modyfikacji platformy. W systemie z realnymi konsekwencjami klinicznymi nowa wersja powinna być oceniona na danych niezależnych i zatwierdzona zgodnie z przyjętą procedurą zmian.

Szczególnie niebezpieczne jest samowzmacnianie się predykcji. Jeżeli model uzna pewne osoby za grupę niskiego ryzyka i nie trafią one do pełnej oceny, w przyszłym zbiorze danych może zabraknąć informacji o ich rzeczywistych rozpoznaniach. Wówczas system uczy się nie tylko z populacji, ale z własnych wcześniejszych decyzji o tym, komu zapewnić dalszą diagnostykę. Taki mechanizm sprzężenia zwrotnego należy uwzględnić przy planowaniu badań prospektywnych i monitorowaniu brakujących wyników referencyjnych.

8. Telediagnostyka jako usługa: czynniki ludzkie, finansowe i organizacyjne

Telediagnostyka (*teleassessment*, *telediagnostics*) nie sprowadza się do transmisji wideo ani przesłania kwestionariusza przez internet. Jest systemem świadczenia usług, który wymaga odpowiedzialności klinicznej, wsparcia administracyjnego, ochrony danych, jasnego trybu kwalifikacji pacjenta i procedur postępowania w przypadkach niejednoznacznych. Potrzebne są także alternatywne ścieżki dla osób, które nie mogą skorzystać z kanału zdalnego.

W pilotażowym modelu teleneuropsychologii dla dzieci mieszkających w obszarach wiejskich Nowego Meksyku istotnym ograniczeniem były długie odległości do specjalistycznych ośrodków. Badacze opisali bariery harmonogramowania wizyt, komunikacji z rodzinami, refundacji świadczeń oraz dostępności

odpowiednio przygotowanego personelu. Spośród 12 skierowań pięć zakończyło się realizacją konsultacji. Wdrożenie wymagało udziału dwujęzycznego personelu administracyjnego i dostosowania komunikacji do potrzeb rodzin [7].

Oznacza to, że sam wzrost dokładności klasyfikatora nie musi przełożyć się na większą liczbę skutecznie obsłużonych pacjentów. Jeśli system utrudnia dostęp do konsultacji rodzinom o słabej łączności, ograniczonej biegłości cyfrowej lub barierach językowych, może pogłębić nierówności. Dobre wdrożenie powinno obejmować wskaźniki dostępności, kompletności dokumentacji, czasu od skierowania do wizyty, rezygnacji z procesu oraz oceny zrozumiałości komunikacji.

Kwestie ekonomiczne wymagają rozróżnienia kosztu przetwarzania danych od kosztu całego procesu. Zdalny formularz może być tani, ale weryfikacja niespójnych odpowiedzi, kontrole bezpieczeństwa, naprawa awarii, opieka nad pacjentem i przygotowanie personelu generują nakłady. Analiza kosztowo-skuteczności powinna uwzględniać także liczbę niepotrzebnych skierowań i koszt pominięcia osoby wymagającej pomocy.

9. Proponowana architektura badań i wdrożenia

Najbardziej racjonalny system wspomagający ocenę ADHD i ASD powinien oddzielać zadania administracyjne, pozyskiwanie danych, obliczanie predykcji oraz odpowiedzialną decyzję kliniczną. Platforma zbierania danych powinna zapisywać czas, wersję formularza i źródło obserwacji. Warstwa analityczna odpowiada za kontrolę kompletności, generowanie cech i oszacowanie ryzyka. Warstwa decyzyjna prezentuje wynik wraz z niepewnością i informacją o zakresie, w którym model był walidowany. Upoważniony specjalista podejmuje decyzję dotyczącą dalszego postępowania.

W analizie podłużnej szczególnie ważne jest modelowanie zmienności wewnątrzosobniczej (*within-person variability*) oddzielnie od różnic między osobami. Przykładowo nagła zmiana rytmu aktywności w odniesieniu do własnej historii użytkownika może być bardziej informacyjna niż porównanie jego wartości z populacyjną średnią. Jednak tylko kontrolowane badanie może rozstrzygnąć, czy konkretny sygnał ma znaczenie diagnostyczne, prognostyczne lub terapeutyczne. Sama dostępność tysięcy punktów pomiarowych nie daje podstaw do takich wniosków.

Projekt powinien uwzględniać minimalizację danych: pozyskiwać tylko informacje niezbędne do jasno zdefiniowanego celu, odróżniać identyfikatory techniczne od danych klinicznych oraz dokumentować uprawnienia dostępu. Rejestrowanie informacji o komunikacji, lokalizacji czy aktywności urządzenia niesie szczególne ryzyko prywatności. Dopuszczalność prawna i etyczna zbierania tych danych nie może być zastąpiona samą techniczną możliwością ich pobrania.

System potrzebuje ponadto **mechanizmu odmowy predykcji** (*abstention*): gdy dane są niepełne, pochodzą z nieobsługiwanej populacji lub niepewność jest zbyt wysoka, poprawnym wynikiem może być skierowanie do oceny eksperckiej bez wygenerowania liczbowej prognozy. Należy gromadzić rejestr zmian modelu, wersję danych wejściowych, parametry kalibracji i wyniki audytu. Każda nowa wersja powinna być oceniona na niezmienionym zbiorze regresyjnym i na niezależnej próbie, zanim zastąpi wersję dotychczasową [10], [11].

9.1. Cztery przykłady interpretacji wyniku w praktyce

Przypadek pierwszy: wysokie ryzyko z niepełną historią rozwoju. Rodzic wypełnia formularz internetowy, a model zgłasza wysokie prawdopodobieństwo ADHD. Z powodu brakujących danych okołoporodowych i niewysłanej oceny nauczyciela wynik nie powinien być traktowany jako pełna diagnoza.

Użyteczny system wskazuje brakujące źródła informacji i proponuje priorytet weryfikacji, nie zaś samodzielne przypisanie rozpoznania.

Przypadek drugi: sprzeczne obserwacje domowe i szkolne. Odpowiedzi rodzica wskazują nasilone trudności, lecz nauczyciel nie obserwuje analogicznego profilu. Uśrednienie dwóch ocen może zamaskować różnicę środowiskową. Właściwy proces zachowuje oba kanały i pozwala klinicyście ustalić, czy rozbieżność wynika z wymagań sytuacyjnych, historii dziecka, różnej interpretacji zachowania czy innych czynników.

Przypadek trzeci: pozornie trafny sygnał cyfrowy. Uczestnik w danym tygodniu rzadziej korzysta z telefonu, a model wiąże zmianę ze spadkiem aktywności. Wywiad ujawnia jednak pobyt bez dostępu do urządzenia. Dodatkowa informacja nie dyskwalifikuje automatycznie telemetrii; pokazuje, że bez odpowiedniego kontekstu jedna i ta sama zmiana może prowadzić do całkowicie odmiennych interpretacji psychologicznych.

Przypadek czwarty: nowe środowisko wdrożeniowe. Model zbudowany w specjalistycznej poradni jest stosowany w szerokim przesiewie szkolnym. Nawet jeśli jego czułość i swoistość pozostałyby zbliżone, mniejsza częstość występowania rozpoznania mogłaby znacznie obniżyć dodatnią wartość predykcyjną. W efekcie wzrosłaby proporcja fałszywych alarmów. Stąd konieczność niezależnej walidacji, kalibracji i dostosowania sposobu komunikowania wyników.

Powyższe scenariusze są przykładami dydaktycznymi, a nie relacjami z konkretnych przypadków klinicznych. Ich zadaniem jest rozdzielenie oceny statystycznej, jakości wejścia i znaczenia dla ścieżki pacjenta. Uczenie maszynowe jest najbardziej użyteczne wtedy, gdy poprawia **kolejny właściwy krok postępowania**, zamiast generować izolowaną etykietę.

9.2. Bezpieczeństwo danych i ograniczenia wykorzystania telemetrii

Dane dotyczące zdrowia psychicznego, rozwoju oraz zachowań cyfrowych są szczególnie wrażliwe. Nawet pozornie anonimowe sekwencje aktywności mogą zawierać charakterystyczne wzorce czasowe i lokalizacyjne. Ochrona prywatności nie powinna więc sprowadzać się do zamiany imienia na losowy identyfikator. Potrzebne są ograniczenie zakresu pobieranych danych, ścisłe uprawnienia dostępu, możliwość audytu, kontrola okresu przechowywania i jasno określone procedury wykorzystania danych do wtórnego treningu.

W przypadku dzieci dodatkowe znaczenie ma rozróżnienie informacji przekazanych przez rodziców, nauczycieli i samego uczestnika. Każdy kanał ma inne podstawy pozyskania, inny zakres wiedzy i inne ryzyko ujawnienia informacji. System powinien umieć utrzymać rozdzielność tych źródeł oraz dokumentować zmiany zgód i uprawnień. W procesie projektowania warto pytać nie tylko, czy da się zgromadzić daną cechę, ale także czy jej wartość dla celu klinicznego jest wystarczająca, by uzasadnić pozyskanie.

Nie istnieje uniwersalna reguła, zgodnie z którą większy model AI lub większy zakres gromadzonych informacji prowadzi do lepszego rozpoznania. W wielu zastosowaniach prostszy, dobrze skalibrowany model oparty na ograniczonym zestawie sprawdzonych danych może być łatwiejszy do utrzymania, walidowania i nadzorowania. Optymalizacja powinna obejmować cały proces świadczenia, nie tylko maksymalną wartość wybranej miary klasyfikacyjnej.

10. Projekt badania porównawczego: od hipotezy do protokołu

Przygotowanie współczesnego badania psychoinformatycznego warto rozpocząć od zdefiniowania punktu końcowego. Możliwymi celami są: wczesne wykrycie potrzeby konsultacji, przewidywanie przyszłego wyniku oceny, identyfikacja narastania trudności w trakcie obserwacji albo klasyfikacja już rozpoznanych przypadków. Nie należy mieszać tych zadań, ponieważ wymagają odmiennych etykiet, horyzontu czasowego i sposobu oceny.

Jako pragmatyczne porównanie proponuje się cztery warianty, liczone na identycznej populacji: (A) sam kwestionariusz i reguła kliniczna, (B) kwestionariusz z modelem liniowym, (C) kwestionariusz z lasem losowym, (D) uzupełnienie wariantu C o jasno określone, dostępne za zgodą dane podłużne. W takim układzie można ustalić, czy złożoność modelu i dodatkowa telemetria przynoszą rzeczywistą korzyść, zamiast tylko zwiększać liczbę parametrów.

Wstępne przetwarzanie należy włączyć do każdego folda walidacji. Uczestnicy, a nie pojedyncze wiersze z ich pomiarami, tworzą podstawową jednostkę podziału. Wszystkie warianty powinny korzystać z tych samych definicji etykiet, okresów obserwacji i procedur doboru progu. Oceniać należy zarówno odsetek błędów, jak i kalibrację oraz potencjalne różnice między podgrupami.

Obszar protokołu	Minimalny wymóg	Dlaczego jest ważny
Cel i punkt końcowy	Jedna, jawna definicja wyniku klinicznego	Eliminuje mieszanie przesiewu z rozpoznaniem
Jednostka podziału	Osoba, a przy potrzebie także ośrodek i czas	Ogranicza przeciek danych
Przetwarzanie	Imputacja, selekcja, SMOTE tylko wewnątrz treningu	Zapobiega zawyżeniu wyniku
Miary	Czułość, swoistość, PPV/NPV, kalibracja i przedziały ufności	Oddaje konsekwencje błędów
Analiza podgrup	Wiek, płeć, źródło skierowania, warunki techniczne	Ujawnia niejednorodność i stronniczość
Walidacja zewnętrzna	Osobna próba z innej placówki lub okresu	Sprawdza transportowalność
Ocena wdrożeniowa	Czas do konsultacji, niepotrzebne skierowania, satysfakcja	Sprawdza rzeczywistą wartość kliniczną

Protokół powinien także przewidywać badanie negatywnych wyników: grup nieobsługiwanych, braków danych, zmiany praktyk klinicznych, awarii czujników i odchyleń od założonego rozkładu. Każdy udokumentowany problem produkcyjny powinien zostać przekształcony w reprodukowalny test regresyjny. Zanim model zostanie wdrożony, należy przeprowadzić audyt metodologiczny zgodny z odpowiednimi elementami TRIPOD+AI i PROBAST+AI [10], [11].

11. Dyskusja i wnioski

Analizowane badania dostarczają dobrych argumentów za cyfrową organizacją zbierania danych i nadzorowanym wykorzystaniem modeli predykcyjnych. MedicalBIT pokazuje użyteczność kwestionariuszy internetowych do tworzenia uporządkowanego procesu przesiewowego i budowy baz danych. Badanie 342 dzieci z podejrzeniem ADHD wskazuje, że algorytm może modelować część zależności wykorzystywanych przez klinicystów i równocześnie ujawniać wpływ objawów ASD na oszacowanie ADHD. Eksperyment WEKA na danych ASD dorosłych obrazuje, jak wysoka może być dokładność rozdzielania klas w zamkniętym problemie obliczeniowym [4]–[6].

Nie jest jednak uzasadnione traktowanie tych wyników jako trzech niezależnych potwierdzeń, że automatyczna diagnoza ADHD i ASD została rozwiązana. Poszczególne prace wykorzystują inne populacje, zadania, transformacje danych i schematy walidacji. Szczególnie wymowne są różnice między wynikiem na ustalonym podziale a walidacją krzyżową w MedicalBIT oraz zależność wyniku SMO od charakterystyki kwestionariuszowego zbioru danych i nadpróbkowania klas.

Również koncepcja fenotypowania cyfrowego wymaga ostrożności. Gęsta rejestracja aktywności może ujawnić zmienność w czasie, ale sygnał urządzenia nie jest sam w sobie wskaźnikiem diagnostycznym. Aby

uniknąć błędów interpretacyjnych, trzeba oddzielić dane obserwowalne, konstrukty psychologiczne i kliniczne decyzje. Zysk z analizy cyfrowej zależy zatem od prawidłowego projektu pomiaru równie mocno jak od wyboru algorytmu.

Ostateczna rekomendacja ma charakter wielowarstwowy: standaryzować pozyskiwanie danych, zachować informację o ich źródle, projektować walidację z uwzględnieniem uczestnika i ośrodka, kontrolować stronniczość i kalibrację, a wynik modelu włączać do procesu odpowiedzialnej oceny specjalistycznej. **Narzędzie może wspierać rozpoznawanie ryzyka i organizację świadczeń, lecz skuteczność diagnostyczną trzeba potwierdzić poza środowiskiem, w którym model powstał.**

Bibliografia

[1] A. Markowetz, K. Błaszkiewicz, C. Montag, C. Switala and T. E. Schlaepfer, "Psycho-Informatics: Big Data shaping modern psychometrics," *Medical Hypotheses*, vol. 82, no. 4, pp. 405–411, 2014, doi: 10.1016/j.mehy.2013.11.030.

[2] C. Montag, É. Duke and A. Markowetz, "Toward Psychoinformatics: Computer Science Meets Psychology," *Computational and Mathematical Methods in Medicine*, vol. 2016, art. 2983685, 2016, doi: 10.1155/2016/2983685.

[3] P. Colombo, S. Busti Ceccarelli, S. Pacchiarini, S. Cribellati and M. Molteni, "A Web Platform for Standardized Data Acquisition, Processing, and Export in the Child Psychopathology Clinical Routine (MedicalBIT): Design and Implementation Study," *JMIR Formative Research*, vol. 6, no. 7, art. e36757, 2022, doi: 10.2196/36757.

[4] S. Grazioli *et al*., "Use of Machine Learning Models to Differentiate Neurodevelopment Conditions Through Digitally Collected Data: Cross-Sectional Questionnaire Study," *JMIR Formative Research*, vol. 8, art. e54577, 2024, doi: 10.2196/54577.

[5] S. Grazioli *et al*., "Exploring telediagnostic procedures in child neuropsychiatry: addressing ADHD diagnosis and autism symptoms through supervised machine learning," *European Child & Adolescent Psychiatry*, vol. 33, pp. 139–149, 2024, doi: 10.1007/s00787-023-02145-4.

[6] S. Jaffer, I. Abdulazez, N. Al-Qazzaz and T. Yousif, "Data Mining for Autism Spectrum Disorder detection among Adults," *Al-Nahrain Journal for Engineering Sciences*, vol. 25, no. 4, pp. 142–151, 2022, doi: 10.29194/NJES.25040142.

[7] A. R. Sherwood and B. MacDonald, "A Teleneuropsychology Consultation Service Model for Children with Neurodevelopmental and Acquired Disorders Residing in Rural State Regions," *Archives of Clinical Neuropsychology*, vol. 35, no. 8, pp. 1196–1203, 2020, doi: 10.1093/arclin/acaa099.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[9] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.

[10] G. S. Collins *et al*., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, art. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[11] K. G. M. Moons *et al*., "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, art. e082505, 2025, doi: 10.1136/bmj-2024-082505.

[12] Q. Wang, R. Luo and Q. Wen, "Artificial intelligence for detecting autism spectrum disorder: a systematic review and meta-analysis," *Frontiers in Public Health*, vol. 14, art. 1923851, 2026, doi: 10.3389/fpubh.2026.1923851.

Aneks A. Słownik pojęć i rozróżnień

Pojęcie	Znaczenie w opracowaniu
Psychoinformatyka (*psychoinformatics*)	Interdyscyplinarne badanie i modelowanie zjawisk psychologicznych z zastosowaniem metod informatycznych
Telediagnostyka (*telediagnostics*)	Zdalne pozyskiwanie informacji i realizacja wybranych elementów procesu oceny klinicznej
Fenotypowanie cyfrowe (*digital phenotyping*)	Ilościowy opis wzorców zachowania na podstawie danych pochodzących z technologii cyfrowych
Trafność predykcyjna (*predictive validity*)	Zdolność predyktora do przewidywania właściwie zdefiniowanego punktu końcowego
Uogólnialność (*generalizability*)	Zdolność utrzymania właściwości predykcyjnych poza zbiorem, na którym model oceniano
Kalibracja (*calibration*)	Zgodność przewidywanego ryzyka z częstością zdarzeń obserwowaną w danych
Przeciek danych (*data leakage*)	Przenikanie do procesu uczenia informacji niedostępnej w rzeczywistym momencie predykcji
Czynniki zakłócające (*confounding factors*)	Zmienne, które mogą zniekształcać zależność między badanym predyktorem a wynikiem
Swoistość (*specificity*)	Odsetek poprawnych rozpoznań klasy negatywnej wśród wszystkich prawdziwie negatywnych przypadków
Precyzja (*precision*, PPV)	Odsetek prawdziwie dodatnich wyników wśród wszystkich wskazań dodatnich

Aneks B. Minimalna karta modelu klinicznego

Wersjonowana karta modelu jest warunkiem kontroli zmian i reprodukowalności. Poniższy zestaw pól stanowi minimum praktyczne, a nie pełny zamiennik TRIPOD+AI i PROBAST+AI [10], [11].

Specyfikacja	Ocena	Nadzór
Cel i populacja	Rozdzielenie treningu/testu	Znane scenariusze niepowodzenia
Kryteria wyłączenia	Walidacja wewnętrzna	Uprawnienia i rejestry dostępu
Definicja etykiety	Walidacja zewnętrzna	Audyt i wersjonowanie danych
Wymagane cechy	Czułość, swoistość, PPV, NPV	Tryb zgłaszania regresji
Przetwarzanie danych	Kalibracja i przedziały ufności	Procedura wyłączenia modelu
Parametry modelu	Wyniki w podgrupach	Odpowiedzialność za decyzję

Zasada interpretacji karty modelu

Karta modelu nie jest jedynie dokumentacją techniczną dla programisty. Stanowi informację o tym, czy określony wynik można bezpiecznie wykorzystać dla konkretnej osoby. W rubryce „cel i populacja” należy zapisać, czy system wspiera przesiew dzieci już skierowanych do poradni, czy ma służyć ogólnemu badaniu populacyjnemu. Podobieństwo tych dwóch zastosowań jest pozorne: ich profile ryzyka, częstość rozpoznań i koszty błędów mogą być różne.

W części dotyczącej walidacji trzeba zapisać oddzielnie wynik z pojedynczego podziału, wynik walidacji wewnętrznej oraz ewentualne rezultaty z niezależnej próby. Jeżeli publikacja podaje wyłącznie wewnętrzne estymacje, karta powinna zawierać jednoznaczną informację „brak walidacji zewnętrznej”, a nie domniemanie, że wynik będzie powtarzalny w nowej placówce. Podobnie przy brakujących danych o kalibracji należy zaznaczyć, że przewidywane prawdopodobieństwa nie zostały ocenione pod kątem zgodności z częstością zdarzeń.

Wreszcie system powinien podawać nie tylko wynik, ale także sytuacje, w których wyniku nie należy interpretować: brak wymaganych formularzy, wiek spoza zakresu badań, istotne zmiany warunków pomiaru albo niewystarczające podobieństwo do danych treningowych. Takie ograniczenia nie świadczą o słabości projektu. Stanowią konieczny element jego rzetelności klinicznej.