AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI I ELEKTRONIKI
KATEDRA AUTOMATYKI

ROZPRAWA DOKTORSKA

ZASTOSOWANIE NARZĘDZI STATYSTYCZNYCH I MATEMATYCZNYCH METOD SZTUCZNEJ INTELIGENCJI DO PREDYKCJI WYSTĄPIENIA DYSPLAZJI OSKRZELOWO-PŁUCNEJ U NOWORODKÓW

MGR INŻ. PAWEŁ STOCH

Promotor:
Prof. dr hab. inż. Wiesław Wajs

Kraków, 2007

Streszczenie

W pracy poruszony został temat predykcji dysplazji oskrzelowo-płucnej - przewlekłego powikłania wcześniactwa u noworodków. Badaną grupę stanowiły dzieci urodzone przedwcześnie o bardzo małej urodzeniowej masie ciała (poniżej 1500 g). Podobnie jak w innych pracach dotyczących tego tematu użyte zostały podstawowe dane typu statycznego takie jak m.in. urodzeniowa masa ciała, wiek płodowy, oraz dane typu dynamicznego - zmienne w czasie hospitalizacji (np. zastosowanie surfaktantu, wentylacja mechaniczna). Dane te pozyskane zostały ze szpitalnej bazy danych Oddziału Intensywnej Terapii Noworodka Polsko-Amerykańskiego Instytutu Pediatrii. W celu poprawy trafności predykcji dodatkowo wykorzystano informacje zebrane przy pomocy specjalnie na potrzeby pracy skonstruowanego systemu rejestracji danych transmitowanych przez urządzenia medyczne.

W większości dotychczasowych prac dotyczących predykcji dysplazji oskrzelowo-płucnej stosowano głównie narzędzia statystyczne. W tej pracy oprócz regresji logistycznej wykorzystano również sztuczne sieci neuronowe. W celu znalezienia optymalnego zbioru zmiennych predykcyjnych zastosowane zostały proste metody selekcji postępującej i eliminacji wstecznej oraz algorytm genetyczny. Przy użyciu wyżej wymienionych metod trafność prognozy dysplazji oskrzelowo-płucnej dla badanej populacji wynosi ok. 85-90%. Dzięki wykorzystaniu powyższych narzędzi możliwe jest skonstruowanie systemu wspomagania decyzji pomagającego lekarzowi w określeniu stopnia ryzyka rozwoju dysplazji oskrzelowo-płucnej u noworodków.

Spis treści

Spis rysunków

Spis tabel

Spis skrótów i oznaczeń

NIS - Neonatal Information System
OITN - Oddział Intensywnej Terapii Noworodka
SpO_2 - oxygen saturation of arterial hemoglobin - wysycenie hemoglobiny tlenem

BPD - Bronchopulmonary Dysplasia - dysplazja oskrzelowo-płucna
PDA - Patent Ductus Arteriosis - przetrwały przewód tętniczy
AA - Alveolar-Arterial Ratio - wskaźnik włośniczkowo-pęcherzykowy
MASAUR - urodzeniowa masa ciała
WIEKPL - wiek płodowy
RESPIMV - zastosowanie respiratora w pierwszym tygodniu życia
SURFACT - podanie surfaktantu w pierwszym tygodniu życia
SPO2MEAN - średnia wartość wysycenia hemoglobiny tlenem
SPO2DEV - odchylenie standardowe wysycenia hemoglobiny tlenem
LOW85 - procent czasu, dla którego wysycenie hemoglobiny tlenem jest mniejsze od 85%
HIGH94 - procent czasu, dla którego wysycenie hemoglobiny tlenem jest większe od 94%
BPMMEAN - średnia ilość uderzeń serca na minutę

1. Wstęp

Opracowanie modelu zjawisk biologicznych, szczególnie w obszarze nauk medycznych, jest zadaniem niezwykle złożonym. Badane zjawiska są z reguły uwarunkowane wieloczynnikowo. Obserwowane zależności w większości przypadków mają charakter nieliniowy. Dynamiczny rozwój narzędzi teleinformatycznych umożliwia obecnie przesyłanie i gromadzenie dużej liczby danych, jak również konstruowanie coraz bardziej skomplikowanych modeli matematycznych opisujących procesy badane w medycynie. W wielu szpitalach instalowane są komputerowe bazy danych, w których gromadzi się dane istotne w diagnostyce i leczeniu. Jedną z najważniejszych zalet takiego sposobu przechowywania informacji jest łatwy i szybki dostęp do nich. Dynamiczny wzrost mocy obliczeniowej dzisiejszych komputerów umożliwia wykorzystanie tych danych, przy użyciu nowoczesnych narzędzi matematycznych do rozwiązania wielu złożonych problemów. Przykładem takiego zastosowania może być wykorzystanie narzędzi sztucznej inteligencji do modelowania zjawisk biologicznych.

Podstawowymi problemami pojawiającymi się przy wykorzystywaniu zgromadzonych danych medycznych jest ich kompletność i wiarygodność. Dane gromadzone w szpitalnych bazach danych zazwyczaj pochodzą z różnych źródeł. Jako przykład przytoczyć można tutaj funkcjonowanie bazy danych na Oddziale Intensywnej Terapii Noworodka Polsko-Amerykańskiego Instytutu Pediatrii Collegium Medicum Uniwersytetu Jagiellońskiego. Znaczna część informacji gromadzonej w tej bazie wprowadzana jest ręcznie przez personel szpitala. Oczywiście w pewnych przypadkach ma to uzasadnienie, gdyż nie istnieje inna metoda wprowadzenia tych danych do elektronicznej bazy. Przykładem mogą być wielkości określane w momencie urodzenia noworodka, takie jak urodzeniowa masa ciała, wiek płodowy, płeć, czy punktacja w skali Apgar. Jednakże, dodatkowo przechowywane są informacje dotyczące przebiegu hospitalizacji - zarówno wyniki badań, jak i dane rejestrowane przez aparatury diagnostyczno - monitorujące. W takim przypadku ręczne wprowadzanie odczytanych z urządzeń medycznych wielkości jest zbyt czasochłonne. Ponadto zebrane dane, odwzorowując stan zdrowia pacjenta z pewnej chwili niekoniecznie reprezentatywnej dla ogólnego stanu zdrowia pacjenta, obarczone są dużym błędem dyskretyzacji.

Rozwiązaniem tego problemu jest automatyczny zapis danych uzyskiwanych z urządzeń medycznych. Znaczna część używanych obecnie urządzeń medycznych umożliwia transmisję

mierzonych wielkości w sposób elektroniczny. Problemem jednak jest duża liczba stosowanych protokołów transmisji, brak jednolitego standardu formatu transmitowanych danych dla urządzeń różnych producentów i brak ogólnie dostępnego oprogramowania umożliwiającego dowolną manipulację otrzymanymi danymi. Podejmowane były wprawdzie różne inicjatywy w tym zakresie, jednak daleko jeszcze do pełnej standaryzacji w tej dziedzinie [52],[39].

W niniejszej pracy zaproponowana została metoda konstrukcji systemu gromadzenia transmitowanych przez urządzenia medyczne parametrów określających stan zdrowia noworodka. Podłączenie urządzeń do takiego systemu umożliwia prezentację tych danych w czasie rzeczywistym na zdalnym komputerze z wykorzystaniem sieci komputerowej. Wyświetlenie raportów, przebiegów, trendów monitorowanych wartości, jak również alarmów wysyłanych przez urządzenia medyczne znacznie ułatwia ocenę stanu zdrowia noworodka. Co więcej, zgromadzenie informacji transmitowanej przez pulsoksymetry, respiratory, czy urządzenia służące do oznaczania gazometrii krwi skutkuje uzyskaniem zupełnie nowego strumienia danych, zawierającego informacje uzupełniające w stosunku do tych zgromadzonych wcześniej w bazie danych Oddziału Intensywnej Terapii Noworodka. Otrzymane w ten sposób dane pozwalają poprawić wyniki modelowania wielu zjawisk biologicznych.

Na przykładzie problemu predykcji dysplazji oskrzelowo-płucnej u noworodków można wykazać, że dane uzyskane w wyniku ciągłego zapisu i gromadzenia informacji transmitowanych przez urządzenia medyczne, w znaczący sposób mogą poprawić jakość uzyskiwanych rezultatów.

Dysplazja oskrzelowo-płucna (ang. *BronchoPulmonary Dysplasia* - BPD) jest przewlekłym powikłaniem wcześniactwa związanym z uszkodzeniem dróg oddechowych i pęcherzyków płucnych w wyniku działania tlenu, dodatniego ciśnienia w drogach oddechowych i czynników zapalnych [18]. Dla potrzeb badań naukowych i porównania wyników leczenia dysplazję oskrzelowo-płucną definiuje się najczęściej jako tlenozależność po 28 dniu życia. Krytyczną rolę w pierwszych dniach życia przedwcześnie urodzonego dziecka odgrywają płuca. Nasilenie zaburzeń oddychania, występujące w pierwszych dniach życia, decyduje o przeżyciu i dalszych losach dziecka. W powstaniu dysplazji decydującą rolę odgrywa samo wcześniactwo, a także zespół zaburzeń oddychania (ang. *Respiratory Distress Syndrome* - RDS) i wentylacja mechaniczna. Istnieje jednak wiele czynników współdziałających, które mogą nasilać zaburzenia oddychania i sprzyjać wystąpieniu BPD.

Należy do nich zakażenie wewnątrzmaciczne, które samo w sobie może być przyczyną przedwczesnego porodu i prowadzić do wrodzonego zapalenia płuc, jak również przetrwały przewód tętniczy (ang. *Patent Ductus Arteriosus* – PDA). Wystąpienie dysplazji oskrzelowo-płucnej zależy więc od wielu czynników. Wczesne (w ciągu kilku pierwszych dni życia) prognozowanie wystąpienia BPD daje możliwość podjęcia działań mających na celu ograniczenie ryzyka. Przykładami takich działań są leczenie surfaktantem i szybka rezygnacja z wentylacji mechanicznej. Kolejnym możliwym do podjęcia krokiem zapobiegającym wystąpieniu BPD jest wczesne podwiązanie przewodu tętniczego. Występowanie BPD można także ograniczyć przez zwalczanie infekcji. Działania te, w zależności od czasu ich wprowadzenia i zakresu, mogą przynieść mniejsze lub większe ograniczenie ryzyka wystąpienia dysplazji oskrzelowo-płucnej.

Przeszukując jedną z najbardziej kompleksowych i uznanych w środowisku medycznym baz bibliografii medycznej MEDLINE (Medical Literature Analysis and Retrieval System Online) [35] można znaleźć przykłady wielu prac poruszających problematykę oceny ryzyka występowania dysplazji oskrzelowo-płucnej u noworodków ([6],[7],[12],[41],[45],[46],[54]). Autorzy tych prac przebadali od kilkudziesięciu do nawet kilku tysięcy noworodków o bardzo niskiej urodzeniowej masie ciała (ang. *very low birth weight infants* - VLBWI), badając przy użyciu metod statystycznych wpływ różnych wielkości na ryzyko rozwoju dysplazji oskrzelowo-płucnej w badanej grupie dzieci. Rozpatrywanymi parametrami były zarówno parametry typu „statycznego” (jak urodzeniowa masa ciała, wiek płodowy, płeć, itp.), jak i parametry „dynamiczne” reprezentujące pewne zdarzenia w czasie hospitalizacji, np. podanie surfaktantu, zmiany sposobu wentylacji mechanicznej, itp. Przykładowo J. Tapia i in. [54] analizując dane 1825 noworodków o bardzo niskiej urodzeniowej masie ciała doszli do wniosku, że większa wartość urodzeniowej masy ciała noworodka i wieku płodowego oraz żeńska płeć noworodka wiążą się z mniejszym ryzykiem rozwoju dysplazji oskrzelowo-płucnej. Natomiast czynnikami zwiększającymi to ryzyko są m.in. mechaniczna wentylacja i PDA. G. Cunha i in. [6] wzięli pod uwagę grupę 153 noworodków o masie poniżej 1500 g. Na podstawie badań tych dzieci, które dożyły 28 dnia, doszli do podobnych wniosków : niewielka urodzeniowa masa ciała i wiek płodowy były skorelowane z podwyższonym ryzykiem dysplazji oskrzelowo-płucnej. Czynniki takie jak płeć, punktacja Apgar, wiek matki i inne spośród wziętych pod uwagę nie miały wpływu na ryzyko rozwinięcia dysplazji. Z kolei R. Somech i in. [45] stwierdzili różnicę w wartości punktacji Apgar dla pierwszej minuty pomiędzy grupami dzieci z dysplazją i bez dysplazji. Autorzy w wyżej wymienionych pracach nie budują modelu predykcyjnego, który pozwalałby określić ryzyko rozwoju

dysplazji oskrzelowo-płucnej u konkretnego dziecka. Koncentrują się raczej na określaniu statystycznej istotności poszczególnych rozważanych czynników na ryzyko rozwoju BPD w badanej populacji noworodków.

Przeszukując bazę MEDLINE nie udało się znaleźć przykładów wykorzystania danych transmitowanych w sposób ciągły przez urządzenia medyczne do predykcji dysplazji oskrzelowo-płucnej. Zastosowanie tych wielkości jest metodą umożliwiającą polepszenie zdolności predykcyjnej rozpatrywanych modeli. W mojej pracy, podobnie jak w wyżej wymienionych publikacjach, do predykcji BPD zastosowane zostały metody statystyczne (w tym przypadku regresja logistyczna), jak również sztuczne sieci neuronowe (ang. *artificial neural networks* – ANN). Narzędzia statystyczne do niedawna były najpopularniejszymi narzędziami służącymi do analizy danych medycznych. Sztuczne sieci neuronowe są w ostatnich latach coraz powszechniej wykorzystywanym w nauce narzędziem analitycznym. Ich działanie, w dużym przybliżeniu, zbliżone jest do koncepcji działania mózgu. Ogromną zaletą sieci neuronowych jest fakt, że pozwalają one na poszukiwanie modeli dla słabo poznanych zjawisk i procesów, ponieważ sieć w procesie uczenia sama określa zależności pomiędzy parametrami, tworząc model wyłącznie na podstawie dostępnych zbiorów danych. Jak wykazano w literaturze sieci neuronowe posiadają zdolności do aproksymacji dowolnych zależności nieliniowych. Takie modelowanie zjawisk w oparciu o obserwacje empiryczne jest szczególnie przydatne w naukach biologicznych, w których często mamy do czynienia z bardzo złożonymi zależnościami.

Oprócz opisanych powyżej narzędzi, w mojej pracy wykorzystane zostały również metody poszukiwania optymalnego zbioru wielkości predykcyjnych. Metody te polegają na przeszukaniu wszystkich możliwych kombinacji dla niewielkich ilości danych lub użyciu algorytmów heurystycznych w przypadku większej ich ilości. Zastosowanie ich umożliwi w przyszłości konstrukcję systemu informatycznego wykorzystującego dane kolejnych hospitalizowanych pacjentów do budowy coraz bardziej skomplikowanych narzędzi predykcyjnych. Narzędzia te korzystając z coraz większej zgromadzonej bazy wiedzy w rezultacie umożliwią wspomaganie lekarza w podejmowaniu decyzji w kierunku właściwego leczenia.

2. Cel i tezy pracy

Celem pracy jest wykorzystanie narzędzi statystycznych i matematycznych metod sztucznej inteligencji do predykcji wystąpienia odległego powikłania dla wcześniactwa tj. dysplazji oskrzelowo-płucnej, na podstawie analizy zarówno statycznych jak i dynamicznych parametrów stanu zdrowia dzieci (takich jak m.in. urodzeniowa masa ciała, wiek płodowy, przetrwały przewód tętniczy, wskaźnik włośniczkowo-pęcherzykowy, wysycenie hemoglobiny tlenem).

Dane wykorzystane w pracy pobrane zostały ze szpitalnej bazy danych NIS (Neonatal Information System) zainstalowanej na Oddziale Intensywnej Terapii Noworodka Szpitala Uniwersyteckiego w Krakowie - Prokocimiu. Ponadto dla celów niniejszej pracy skonstruowano system rejestracji danych medycznych, którego zadaniem jest gromadzenie danych transmitowanych w sposób ciągły przez urządzenia medyczne. Uzyskane w ten sposób dane zastosowano w niniejszej pracy w celu wykazania poprawności następujących tez :

- **Analiza przebiegu parametrów statycznych i dynamicznych stanu zdrowia noworodków w pierwszym tygodniu życia pozwala na predykcję wystąpienia dysplazji oskrzelowo-płucnej.**

- **Analiza parametrów uzyskanych w wyniku ciągłego monitorowania noworodka umożliwia poprawę jakości predykcji dysplazji oskrzelowo-płucnej.**

3. Struktura pracy

W pracy wyróżniono cztery części.

Część pierwsza przedstawia opis metod pozyskiwania danych medycznych z bazy danych funkcjonującej w szpitalu oraz ze skonstruowanego i zainstalowanego na OITN systemu rejestracji danych transmitowanych przez urządzenia medyczne. Ponadto, w części tej przedstawiona została analiza badanej grupy pacjentów wraz z opisem wykorzystywanych w pracy parametrów stanu zdrowia noworodków.

Część druga prezentuje od strony teoretycznej metody badawcze wykorzystane do predykcji dysplazji oskrzelowo-płucnej u noworodków, tzn. model regresji logistycznej oraz sztuczne sieci neuronowe.

W części trzeciej opisano wyniki przeprowadzonych badań, porównanie rezultatów osiągniętych w wyniku zastosowania modeli statystycznych i sieci neuronowych.

Ostatnia, czwarta część stanowi podsumowanie, wnioski i końcowe uwagi.

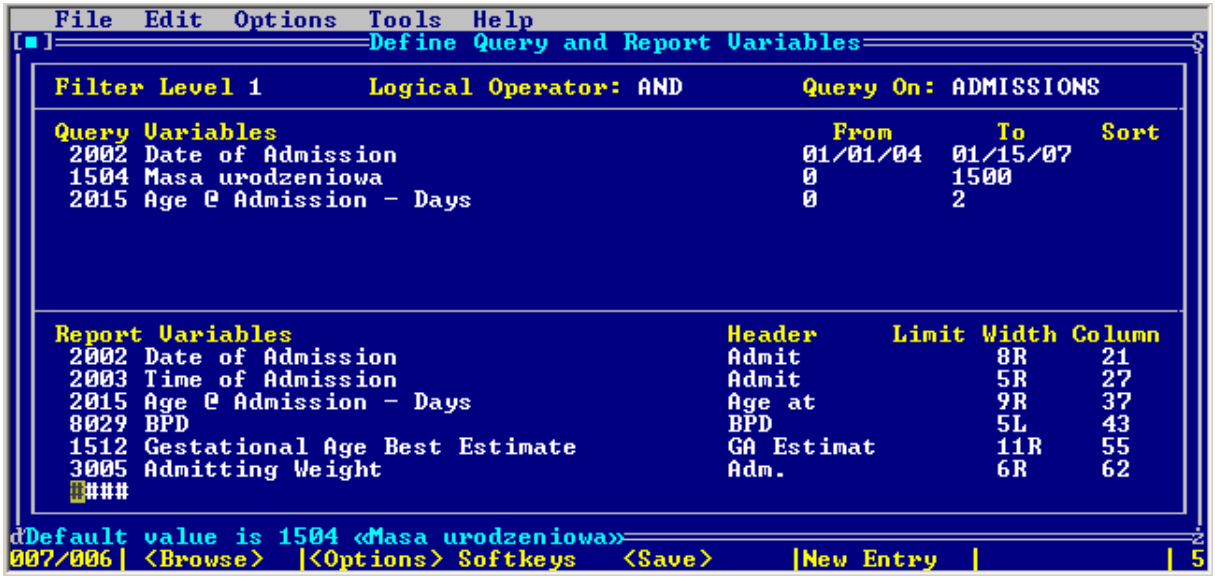
4. Materiał

4.1. Źródła danych

Materiał badawczy wykorzystany w prezentowanej pracy został zgromadzony w trakcie prawie trzyletniej obserwacji pacjentów na Oddziale Intensywnej Terapii Noworodka Polsko-Amerykańskiego Instytutu Pediatrii Collegium Medicum Uniwersytetu Jagiellońskiego. Na oddziale tym od wielu już lat wykorzystywana jest komputerowa baza danych NIS (Neonatal Information System), służąca do przechowywania informacji o wszystkich hospitalizowanych dzieciach. Drugim źródłem danych wykorzystanych w pracy jest system rejestracji danych medycznych monitorowanych w sposób ciągły.

4.1.1. Baza NIS

Baza danych NIS była wykorzystywana jako źródło danych medycznych w wielu pracach [26],[27],[28],[48],[59]. Funkcjonująca już od wielu lat i gromadząca dane dotyczące historii leczenia wszystkich hospitalizowanych dzieci baza oparta jest o system AREV (Advanced Revelation) działający pod kontrolą systemu operacyjnego DOS. Do wprowadzania i pozyskiwania danych służy interfejs tekstowy. Jego wygląd przedstawiony jest na rys.1.



Rys. 1. Wygląd interfejsu tekstowego bazy NIS

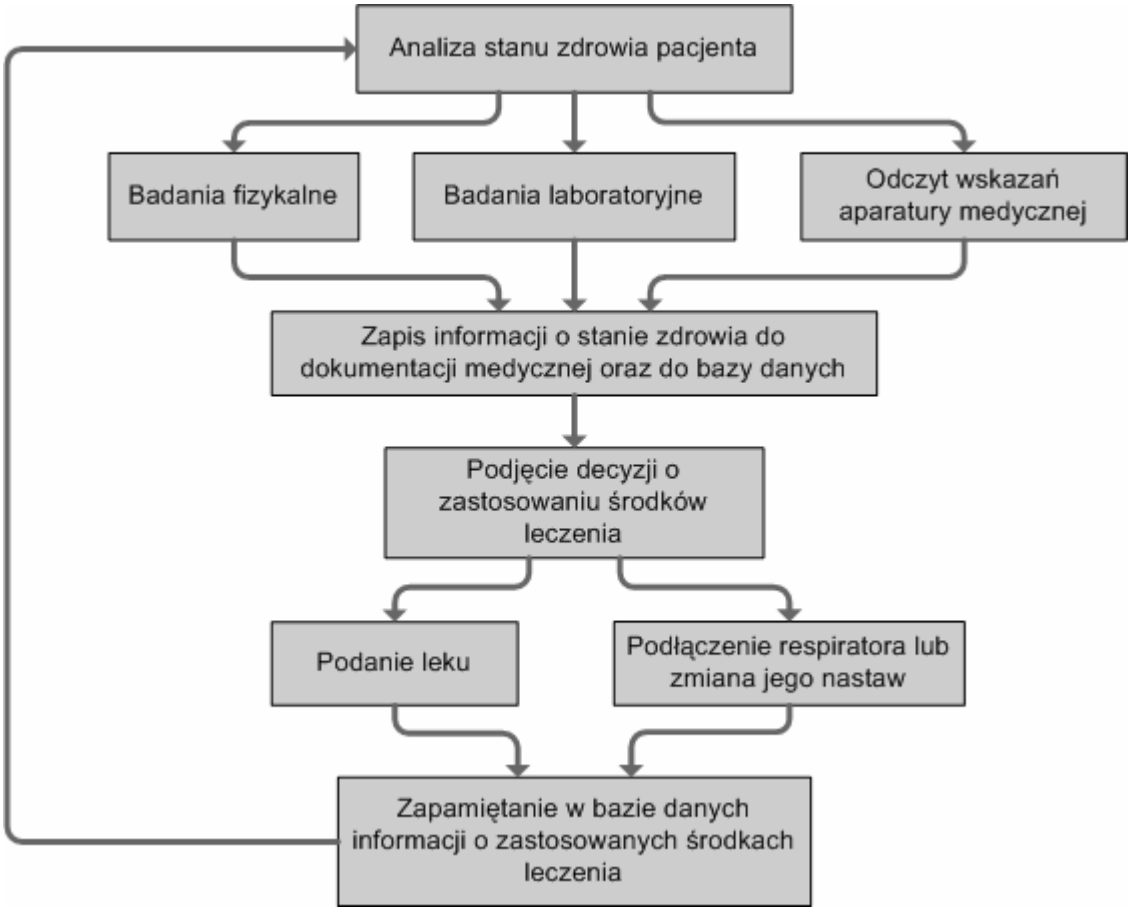
Ze względu na komercyjny charakter tego narzędzia nie jest możliwe uzyskanie dokładniejszych danych o charakterze struktury bazy danych, jak i o rodzaju metod używanych w celu pobrania z niej żądanej przez użytkownika informacji. Można stwierdzić,

że mechanizmy konstruowania zapytań są podobne do składni języka SQL. Rezultatem działania zapytań są pliki tekstowe, z których można uzyskać niezbędne dane.

Proces gromadzenia informacji opisującej stan zdrowia pacjenta, graficznie przedstawiony na rys.2, można opisać w następujących punktach [48] :

1. Przeprowadzane badania fizykalne, laboratoryjne i rejestracja wskazań aparatury medycznej dostarczają informacji o parametrach opisujących aktualny stan zdrowia pacjenta. Dokonywane one są kilka razy dziennie z jednoczesnym zapisem informacji w dokumentacji historii choroby pacjenta i rejestracją danych w szpitalnej bazie danych.
2. W oparciu o uzyskaną informację podejmowana jest decyzja o stosowanych środkach leczenia: dokonywane są ustawienia respiratorów, a także podejmowana jest decyzja o podaniu leku pacjentowi. Dane o tych faktach rejestrowane są w bazie danych. Działania medyczne z punktów 1 i 2 następują kolejno po sobie, dostarczając do bazy danych informacji o historii określonego przypadku chorobowego.
3. Z informacji zawartych w systemie NIS, dla potrzeb niniejszej pracy pobierane są dane pacjentów o charakterze statycznym (takie jak np. urodzeniowa masa ciała, wiek płodowy, itp.), oraz o charakterze skokowym (np. podanie leków typu surfaktant, zmiana nastaw respiratorów, itp.)

Podstawową wadą metody wykorzystywanej w procesie gromadzenia informacji w bazie danych NIS jest fakt ręcznego wprowadzania tych danych. O ile w przypadku danych o charakterze stałym (np. wpisywanych w momencie przyjęcia na oddział) ma to znaczenie niewielkie, o tyle w przypadku danych ciągłych (odczyty wskazań aparatury medycznej) częstotliwość odczytów jest sprawą kluczową. Dyskretyzacja danych o charakterze ciągłym ma negatywny wpływ na dokładność prognozy parametrów medycznych. Ponadto duże znaczenie ma kwestia łatwego dostępu do informacji zawartych w bazie danych. Problemy pozyskiwania danych dotyczą głównie starszych systemów bazodanowych (np. takich jak NIS). Nowsze systemy oparte najczęściej na relacyjnych bazach danych typu Oracle, mysql, PostgreSQL, itp. umożliwiają łatwy dostęp do zgromadzonej informacji. Dalszy rozwój narzędzi wspomagających proces podejmowania decyzji przez lekarza uzależniony jest od łatwości integracji systemów przechowujących dane medyczne.



Rys. 2. Sposób gromadzenia informacji o stanie zdrowia pacjenta i mechanizm podejmowania decyzji o sposobie leczenia

4.1.2. System rejestracji danych medycznych monitorowanych w sposób ciągły

Drugim źródłem danych medycznych wykorzystywanych w pracy jest system rejestracji parametrów medycznych monitorowanych w sposób ciągły przez urządzenia medyczne. System ten został stworzony w celu zwiększenia dokładności i częstotliwości odczytu parametrów, które co prawda są przechowywane w bazie NIS, ale są wprowadzane zbyt rzadko lub z niewystarczającą dokładnością oraz w celu gromadzenia danych, które nie są rejestrowane w żaden inny sposób. W systemie tym zbierane są dane z urządzeń medycznych podłączonych na Oddziale Intensywnej Terapii Noworodka Polsko-Amerykańskiego Instytutu Pediatrii Collegium Medicum Uniwersytetu Jagiellońskiego, jednak jego modułowa architektura umożliwia zainstalowanie podobnego systemu na dowolnym innym oddziale intensywnej terapii wyposażonym w odpowiednie urządzenia medyczne.

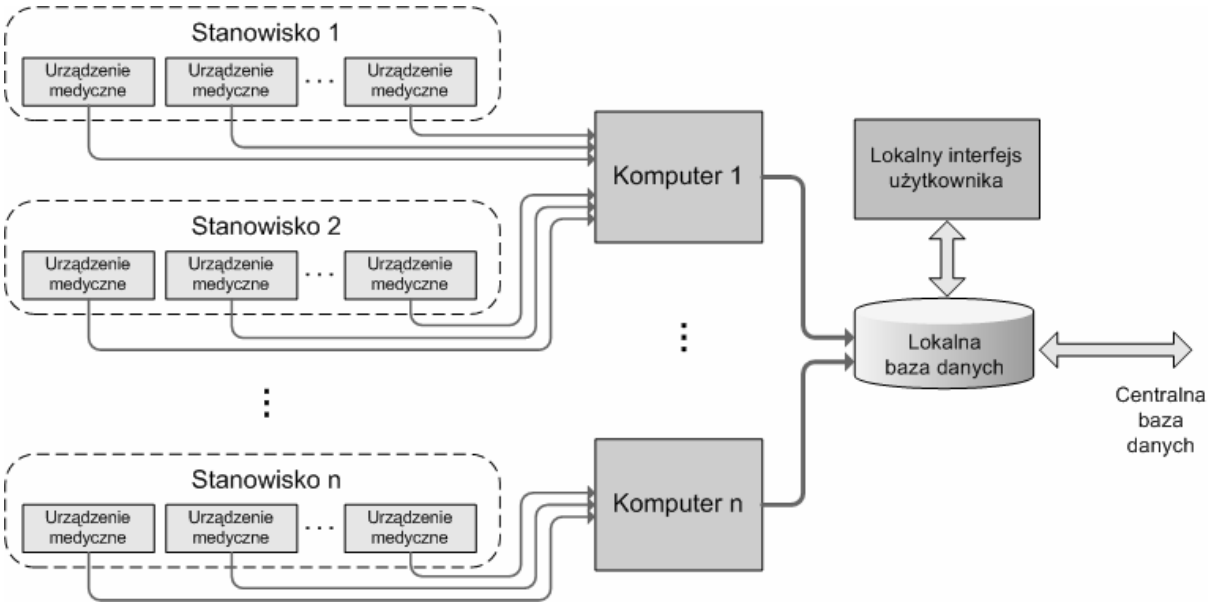
Podstawowym problemem pojawiającym się przy próbie konstrukcji takiego systemu jest różnorodność sposobów przesyłania danych cyfrowych przez urządzenia medyczne [39],[52]. Brak standaryzacji w tej dziedzinie prowadzi do sytuacji, w której każdy producent

implementuje swój własny protokół transmisji danych, niestety rzadko kiedy dostępny w postaci jawnej. Do odczytania i interpretacji tych danych konieczny jest zakup innego dedykowanego do tego celu urządzenia tego samego producenta.

Urządzenia medyczne wykorzystują zasadniczo dwa rodzaje interfejsów do transmisji danych:

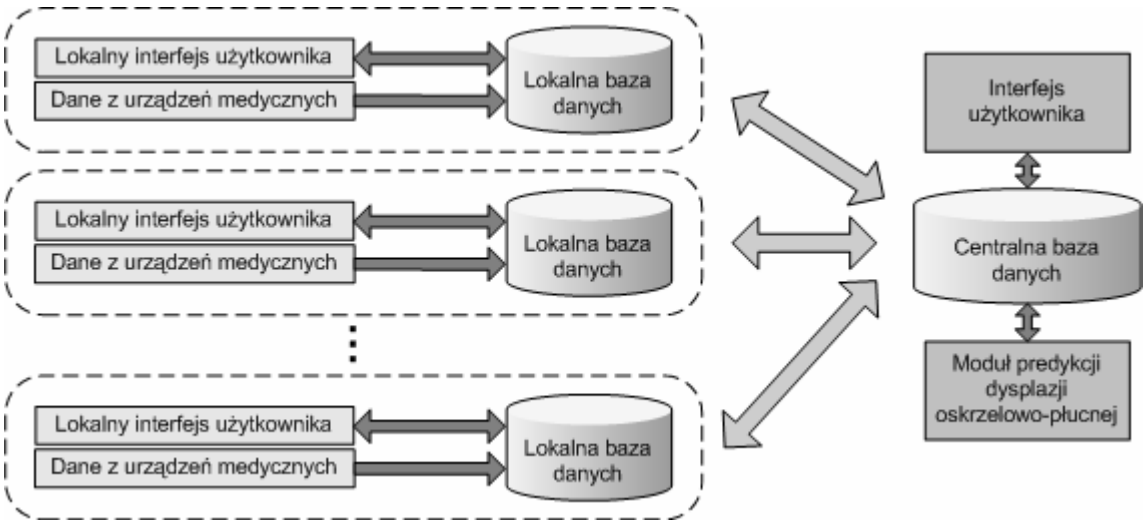
- interfejs szeregowy (RS-232),
- interfejs sieciowy LAN (Ethernet).

System zaprojektowano w sposób modułowy. Istniejący moduł, którego schemat pokazano na rys. 3 zainstalowany został na Oddziale Intensywnej Terapii Noworodka Szpitala Dziecięcego w Krakowie - Prokocimiu. Dane zebrane z tego modułu gromadzone są w jednym miejscu - centralnej bazie danych i wykorzystywane do dalszej analizy (rys. 4). Do transmisji danych pomiędzy lokalnymi modułami a centralną bazą danych wykorzystano sieć Internet.



Rys. 3. Schemat modułu rejestracji danych medycznych

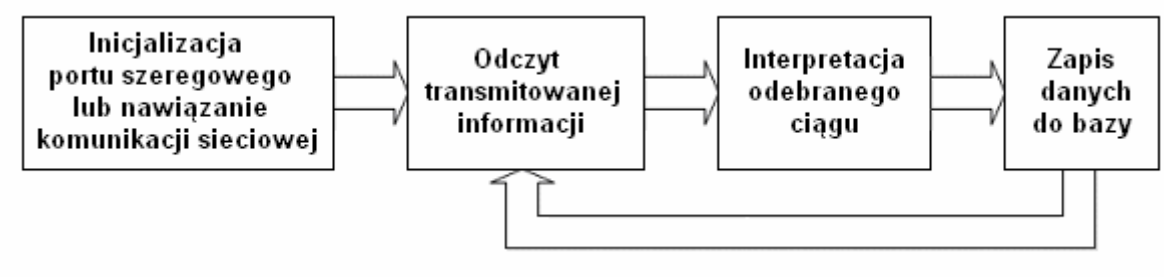
Do każdego stanowiska zbierania danych przyporządkowane jest jedno lub więcej urządzeń monitorujących. Urządzenia monitorujące wykorzystujące złącza szeregowe RS-232 podłączone są do portów szeregowych komputerów zbierających dane. Każdy taki komputer wyposażony jest w wieloportową kartę szeregową. Ilość komputerów potrzebnych do podłączenia wszystkich urządzeń medycznych zależy od wzajemnego rozmieszczenia stanowisk oraz od ilości portów szeregowych karty wieloportowej, w które komputery te są wyposażone.



Rys. 4. Schemat systemu rejestracji i predykcji

Urządzenia medyczne wykorzystujące interfejs sieciowy Ethernet LAN podłączane są za pośrednictwem urządzeń sieciowych (koncentratory i/lub przełączniki sieciowe) do interfejsów sieciowych komputerów zbierających dane. Na każdym komputerze działa specjalny zestaw programów interpretujących dane przesyłane z podłączonych urządzeń monitorujących. Ze względu na opisaną wcześniej różnorodność protokołów transmisji i formatów transmitowanych danych konieczne jest zastosowanie osobnego, specjalnie do tego celu napisanego programu dla każdego podłączonego typu urządzenia medycznego. Program taki po odczytaniu odebranych danych i ich interpretacji dokonuje wstępnego przetworzenia, a następnie przesyła je do lokalnej bazy danych. Przetworzenie tych danych polega na wyodrębnieniu z pełnego strumienia transmitowanego przez dane urządzenie interesującej informacji i odrzuceniu pozostałych informacji nieistotnych z punktu widzenia funkcjonalności systemu. Algorytm działania takiego programu przedstawiony jest na rys. 5 i zawiera następujące operacje:

1) inicjalizacja portu szeregowego (ustalenie odpowiednich parametrów transmisji) lub komunikacji sieciowej (nawiązanie połączenia sieciowego z urządzeniem),
2) zapis do tymczasowego bufora odebranej z urządzenia informacji,
3) interpretacja informacji w buforze – wydzielenie danych dotyczących monitorowanych parametrów,
4) zapis otrzymanej informacji do lokalnej bazy danych z dodaniem dokładnej informacji czasowej pomiaru,
5) powrót do punktu 2.

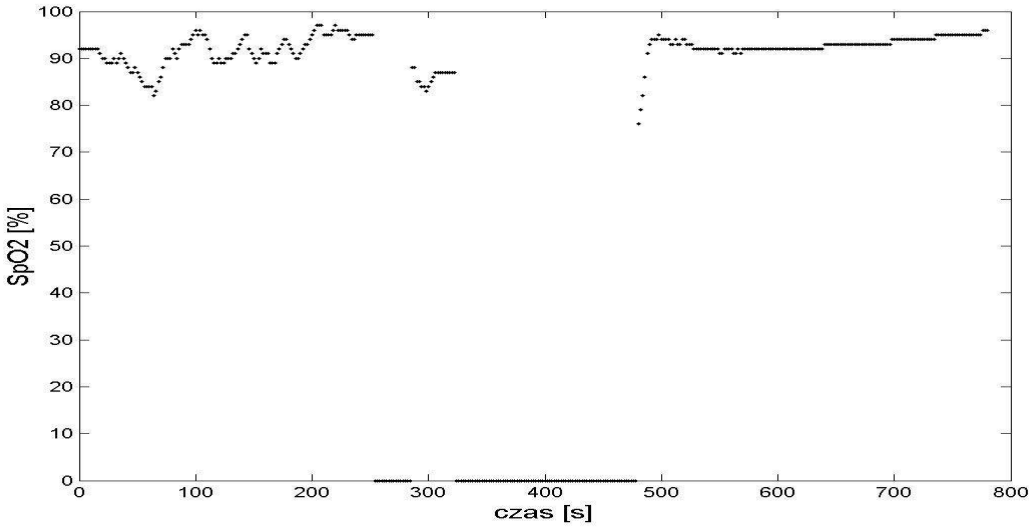


Rys. 5. Schemat blokowy programu odpowiedzialnego za odczyt danych z urządzenia monitorującego i ich zapis do lokalnej bazy danych

Zadaniem lokalnej bazy danych jest przechowywanie napływającej z komputerów przetworzonych wstępnie danych medycznych do momentu ich synchronizacji z główną bazą danych znajdującą się na komputerach laboratorium Katedry Automatyki AGH. Zastosowanie lokalnej bazy danych ma więc głównie na celu zabezpieczenie przed utratą monitorowanych danych np. w przypadku braku dostępu do Internetu (buforowanie danych). Możliwość dokonywania zmian w konfiguracji systemu (na przykład zaznaczenie chwili rozpoczęcia i zakończenia monitorowania pacjenta na stanowisku, identyfikacja monitorowanego pacjenta w systemie) została zapewniona poprzez stworzenie prostego interfejsu www, do którego dostęp ma personel medyczny na Oddziale Intensywnej Terapii Noworodka.

Spośród dostępnych na Oddziale Intensywnej Terapii Noworodka urządzeń medycznych do systemu w chwili pisania tej pracy podłączone zostały urządzenia dwóch typów : pulsoksymetry i respiratory. O ile w przypadku pulsoksymetrów pewnym ułatwieniem jest fakt, że wszystkie są jednego typu (Nellcor NPB-295), o tyle w przypadku respiratorów dużym problemem jest różnorodność dostępnych na OITN modeli. Odpowiedni program interpretujący otrzymane dane został napisany tylko dla respiratorów typu Bear CUB-750.

Z pulsoksymetrów (wyposażonych w port szeregowy RS-232) gromadzone są informacje o chwilowym wysyceniu hemoglobiny tlenem oraz ilości uderzeń serca na minutę. Dane te transmitowane są w postaci tekstowej przez pulsoksymetr co 2 sekundy i zapisywane do bazy danych. Uzyskany w ten sposób strumień informacji musi zostać poddany filtracji, zanim zostanie wykorzystany do predykcji dysplazji oskrzelowo-płucnej. Przede wszystkim konieczne jest wykrycie i usunięcie błędnych próbek, spowodowanych np. rozłączeniem się czujnika. Jest to o tyle ułatwione, że pulsoksymetr w takiej sytuacji transmituje ciąg zer (będących wartościami nieprawidłowymi) i odpowiedni status (rys. 6). Po odrzuceniu nieprawidłowych i zakłóconych próbek pozostałe dane mogą zostać użyte do obliczenia odpowiednich wartości parametrów wtórnych.



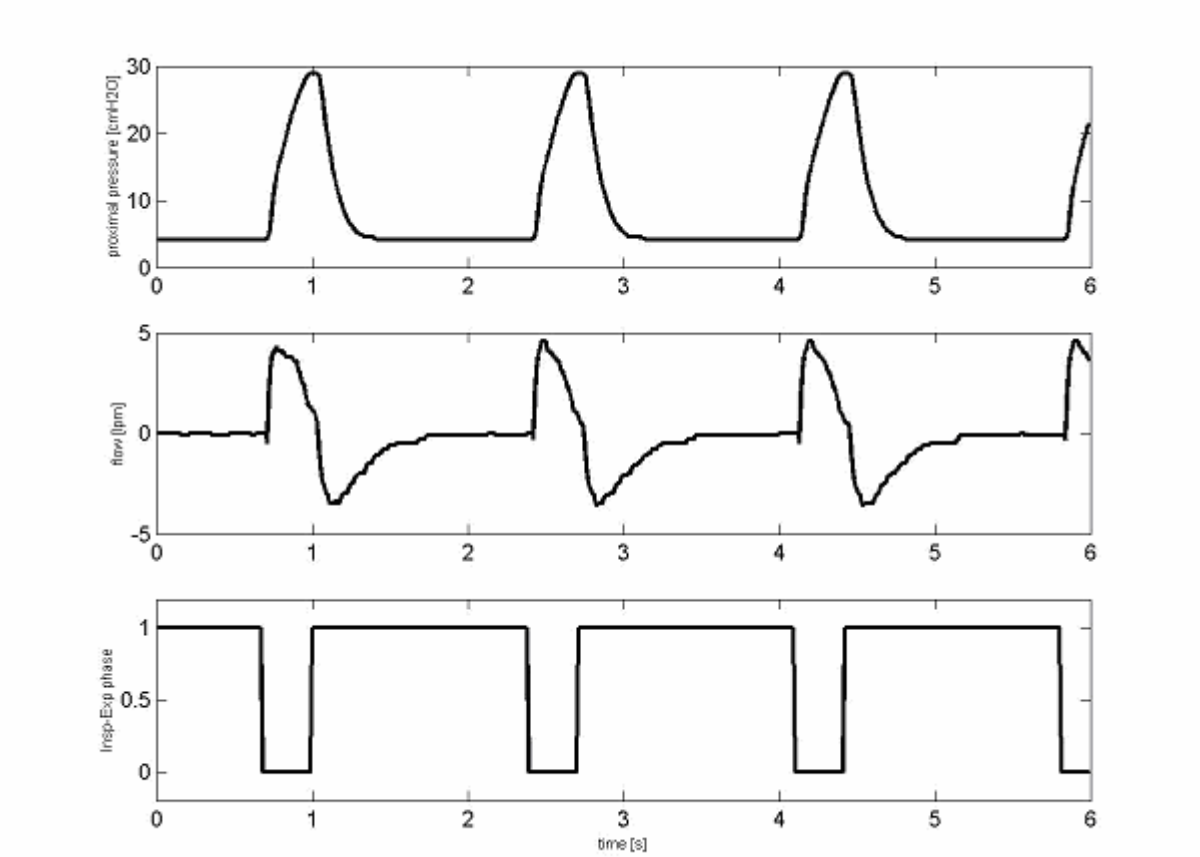
Rys. 6. Przykład przebiegu wysycenia hemoglobiny tlenem w przypadku wystąpienia zakłócenia (rozłączenia czujnika)

System umożliwia również gromadzenie danych uzyskanych z respiratorów. Dane te dotyczą zarówno ustawień, jak również wartości parametrów mechaniki oddychania. Ze względu na niewielką ilość zebranych danych w momencie pisania tej pracy nie zostały one wykorzystane do predykcji dysplazji. Do najważniejszych parametrów transmitowanych przez respirator należą : chwilowa wartość ciśnienia i przepływu mieszanki oddechowej oraz faza oddechu. Przykładowy wykres wartości powyższych parametrów przedstawiony jest na rys. 7.

Na tej podstawie można uzyskać parametry wykorzystywane przy analizie przebiegu sztucznej wentylacji, takie jak :

- MAP (ang. *mean airway pressure*) - średnie ciśnienie w drogach oddechowych,
- PIP (ang. *peak inspiratory pressure*) - szczytowe ciśnienie oddechowe,
- PEEP (ang. *positive end-expiratory pressure*) - końcowo-wydechowe ciśnienie oddechowe,
- RR (ang. *respiratory rate*) – ilość oddechów na minutę

i inne.



Rys. 7. Przykładowe przebiegi wielkości transmitowanych przez respirator Bear CUB-750 : chwilowej wartości ciśnienia i przepływu mieszanki oddechowej oraz fazy oddechu

Spośród innych urządzeń medycznych umożliwiających transmisję danych w postaci cyfrowej wspomnieć należy również o urządzeniach służących do oznaczania gazometrii krwi oraz monitorach oddechowych. Urządzenia te charakteryzują się większą komplikacją sposobu dostępu do danych. Nowsze urządzenia nierzadko wyposażane są w wewnętrzne bazy danych przechowujące wyniki kilku czy kilkunastu ostatnich badań, bazy te zazwyczaj dostępne są tylko przy użyciu specjalnych narzędzi odpowiednich dla danego urządzenia. W urządzeniach tych komunikacja za pomocą interfejsów szeregowych RS-232 coraz częściej zastępowana jest komunikacją sieciową. Urządzenia wyposażane są w interfejsy sieci Ethernet umożliwiające podłączenie takiego urządzenia do sieci komputerowej i dostęp przy użyciu protokołów TCP/IP. W takim przypadku konstrukcja systemu gromadzenia danych ulega uproszczeniu, ponieważ ograniczana zostaje ilość niezbędnych komputerów i kart szeregowych. Zamiast nich stosuje się tanie i niewielkie urządzenia sieciowe (koncentratory, przełączniki). Dodatkową zaletą wynikającą ze stosowania urządzeń wyposażonych w interfejsy sieci LAN są znacznie mniejsze ograniczenia wynikające z maksymalnej długości

kabla transmisyjnego (w zależności od prędkości transmisji od kilkunastu do kilkudziesięciu metrów dla RS-232 [10] i kilkaset metrów dla Ethernet [22]).

4.2. Charakterystyka badanej grupy pacjentów i dostępnych danych

Grupę badaną stanowili pacjenci wypisani z Oddziału Intensywnej Terapii Noworodka Kliniki Chorób Dzieci CMUJ w okresie od stycznia 2004 do grudnia 2006 roku. Dzieci te urodzone były przedwcześnie z masą ciała mniejszą lub równą 1500g i przyjęte na oddział nie później niż w drugiej dobie życia.

Wykorzystano dane tych pacjentów zgromadzone w bazie NIS i w systemie rejestracji parametrów monitorowanych w sposób ciągły. Z całkowitej ilości ponad 90 pacjentów odrzucono te przypadki, dla których nie zebrano odpowiedniej ilości danych, oraz dane dzieci zmarłych przed 28 dniem życia. Uzyskana w ten sposób grupa badana liczyła 70 przypadków.

Powyższe kryteria wyboru pacjentów do badanej grupy były jedynymi kryteriami, grupę stanowiły kolejne dzieci przyjmowane na oddział. Daje to podstawę do twierdzenia o reprezentatywności tej grupy jako podzbioru populacji.

W grupie tej u 26 noworodków rozwinęła się dysplazja, u pozostałych 44 nie, tak więc dzieci, u których dysplazja się rozwinęła, stanowią 37% wszystkich badanych pacjentów. Ponadto 24 spośród noworodków miało przetrwały przewód tętniczy (PDA), co stanowi 34% wszystkich pacjentów. Spośród wszystkich dzieci stanowiących badaną grupę u 54 zastosowana była wentylacja mechaniczna w pierwszym tygodniu życia (77% ogółu). Surfaktant w pierwszym tygodniu życia zastosowany był u 24 noworodków (34% wszystkich).

Wielkości użyte do predykcji dysplazji oskrzelowo-płucnej wybrane zostały na podstawie literatury dotyczącej dysplazji oskrzelowo-płucnej i sugestii lekarzy pracujących na Oddziale Intensywnej Terapii Noworodka Kliniki Chorób Dzieci CMUJ. Wybrano wielkości podstawowe, dobrze opisane w istniejącej bazie danych i potwierdzone w kartach historii hospitalizacji danych pacjentów. Czynniki te uważa się powszechnie za mające największy wpływ na ryzyko rozwinięcia się dysplazji oskrzelowo-płucnej.

W dostępnej literaturze brak jest informacji na temat sposobu doboru parametrów uzyskanych z systemu monitorowania danych i dotyczących wysycenia hemoglobiny tlenem. Zastosowane w tej pracy wielkości zaproponowane zostały przez lekarzy OITN na podstawie własnego doświadczenia.

Do celów analizy i predykcji dysplazji wybrane zostały następujące parametry (w nawiasie podana została skrótowa nazwa używana w dalszej części pracy) uzyskane z bazy NIS :

- fakt zaistnienia dysplazji oskrzelowo-płucnej na podstawie rozpoznania wypisowego (BPD),
- urodzeniowa masa ciała (MASAUR),
- wiek płodowy (WIEKPL),
- zastosowanie respiratora (RESPIMV),
- wskaźnik włośniczkowo-pęcherzykowy (ang. *alveolar-arterial ratio*) określany wg wzoru 4.1; wyznaczany w pierwszym badaniu po przyjęciu na oddział (AA),
- przetrwały przewód tętniczy (PDA),
- podanie surfaktantu w pierwszym tygodniu życia (SURFACT),

ponadto parametry wtórne uzyskane z pomiaru wysycenia hemoglobiny tlenem i uśrednione dla pierwszego tygodnia hospitalizacji :

- średnia wartość wysycenia hemoglobiny tlenem (SPO2MEAN),
- odchylenie standardowe wysycenia hemoglobiny tlenem (SPO2DEV),
- procent czasu, dla którego wysycenie hemoglobiny tlenem było mniejsze od 85% (LOW85),
- procent czasu, dla którego wysycenie hemoglobiny tlenem było większe od 94% (HIGH94),
- średnia ilość uderzeń serca na minutę (BPMMEAN),

oraz parametry uzyskane z pomiaru wysycenia hemoglobiny tlenem i obliczone jako stosunek wartości obliczanych dla pierwszego tygodnia hospitalizacji do wartości obliczanych dla pierwszego dnia hospitalizacji (obrazujące trend):

- stosunek wartości średniej wysycenia hemoglobiny tlenem obliczonej dla pierwszych 7 dni do analogicznej wartości średniej obliczonej dla pierwszego dnia (SPO2MEAN_TR),
- stosunek wartości odchylenia standardowego wysycenia hemoglobiny tlenem obliczonej dla pierwszych 7 dni do analogicznej wartości odchylenia standardowego wysycenia hemoglobiny tlenem obliczonej dla pierwszego dnia (SPO2DEV_TR),
- stosunek średniej ilości uderzeń serca na minutę obliczonej dla pierwszych 7 dni do analogicznej średniej ilości uderzeń serca na minutę dla pierwszego dnia (BPMMEAN_TR).

Parametry te podzielić można na dwie grupy :

1. Parametry statyczne

Parametry o charakterze statycznym nie zmieniają swoich wartości w trakcie okresu hospitalizacji, ich wartości odczytywane są podczas przyjęcia pacjenta na oddział. Do grupy tej należą następujące z wymienionych wcześniej wielkości :

- urodzeniowa masa ciała

parametr o charakterze ciągłym, wartości przyjmowane przez ten parametr zawierają się w przedziale od 600 do 1500g. Masa ciała w momencie urodzenia dziecka jest jednym z najważniejszych parametrów w całościowej ocenie stanu noworodka. Z reguły im wyższa waga, tym lepsze rokowania dla dziecka, krótszy okres jego hospitalizacji i jej łagodniejszy przebieg.

- wiek płodowy

parametr o charakterze ciągłym, wiek płodowy dziecka w momencie jego urodzin przyjmuje wartości z zakresu od 24 do 32 tygodni. Im krótszy, tym gorszy jest stan dziecka po narodzinach.

- przetrwały przewód tętniczy (PDA)

przetrwały przewód tętniczy jest pozostałym z okresu życia płodowego naczyniem łączącym tętnicę płucną z aortą [2]. Normalnie przewód tętniczy zamyka się w ciągu kilkudziesięciu godzin po urodzeniu. Im krótszy wiek płodowy, tym wyższe jest prawdopodobieństwo przetrwania drożności przewodu. Parametr o charakterze dyskretnym: 0 oznacza brak, 1 oznacza PDA.

2. Parametry dynamiczne

W trakcie hospitalizacji pacjent poddawany jest obserwacji i badaniom, w wyniku których w miarę potrzeb podejmowane są decyzje o zastosowaniu środków leczenia. Badania i stosowanie środków leczenia mogą mieć miejsce wielokrotnie w ciągu dnia, w związku z czym można powiedzieć, że informacja opisująca dany przypadek pacjenta posiada charakter dynamiczny, zmienny w czasie. W celu uproszczenia analizy większość użytych w pracy

parametrów o charakterze dynamicznym sprowadzonych zostało do jednej wartości określonej dla przedziału pierwszego tygodnia hospitalizacji noworodka. Do parametrów tych należą :

– wskaźnik włośniczkowo-pęcherzykowy

wskaźnik włośniczkowo-pęcherzykowy [15] określa stopień zaburzeń oddychania. Wyznaczany jest ze wzoru :

$$AA = \frac{pO_2}{p_{ATM} \cdot FiO_2 - pCO_2} \tag{4.1}$$

gdzie : pO_2 [mmHg] - ciśnienie parcjalne tlenu, p_{ATM}=713 mmHg - ciśnienie atmosferyczne, pCO_2 [mmHg] - ciśnienie parcjalne dwutlenku węgla, FiO_2 - stężenie tlenu w mieszaninie oddechowej. Tak więc wskaźnik ten odzwierciedla stosunek ciśnienia pO_2 tętniczego do ciśnienia tlenu w pęcherzyku płucnym i jest wyznacznikiem jakości wymiany tlenu przez barierę włośniczkowo-pęcherzykową. Wskaźnik AA jest parametrem o charakterze ciągłym, może przyjmować wartości z zakresu od 0 do 1.

– zastosowanie respiratora

parametr ten określa, czy w trakcie pierwszego tygodnia hospitalizacji pacjenta stosowana była wymuszona wentylacja mechaniczna przy użyciu respiratora (IMV - ang. *intermittent mandatory ventilation*). Zbyt intensywna wentylacja mechaniczna zwiększa ryzyko rozwoju dysplazji oskrzelowo-płucnej. Parametr o charakterze dyskretnym, może przyjmować wartość 0 w przypadku braku wentylacji lub 1 w przeciwnym przypadku.

– zastosowanie surfaktantu

surfaktant jest lekiem zawierającym związki obniżające napięcie powierzchniowe [26]. Podanie surfaktantu zwiększa podatność płuc i powoduje upowietrznienie zapadniętych pęcherzyków płucnych. Parametr określający fakt podania surfaktantu jest parametrem dyskretnym i może przyjmować wartości 1 lub 0.

Do osobnej grupy parametrów dynamicznych należy zaliczyć parametry uzyskane z systemu rejestracji danych medycznych. Gromadzenie danych z urządzeń medycznych od momentu urodzenia daje pełny obraz przebiegu tych wielkości w trakcie hospitalizacji pacjenta. Dla tego strumienia danych zastosowano to samo podejście, co w przypadku

parametrów dynamicznych uzyskanych z bazy NIS, tzn. obliczono charakterystyczne pojedyncze wartości parametrów dla pierwszego tygodnia życia noworodka. Jako podstawę obliczania parametrów wykorzystano dwie wielkości transmitowane przez pulsoksymetry NPB-295 : wysycenie hemoglobiny tlenem (SpO_2) i ilość uderzeń serca na minutę (BPM - ang. *beats per minute*). Wartość wysycenia hemoglobiny tlenem zazwyczaj oscyluje w granicach 85-90%, ale zdarzają się okresy, kiedy może spadać do niskich wartości. Ilość uderzeń serca kształtuje się w granicach 130-170 uderzeń na minutę.

Do grupy należą następujące parametry :

- średnia wartość wysycenia hemoglobiny tlenem (SPO2MEAN)

parametr ten określa średnie wysycenie hemoglobiny tlenem w pierwszym tygodniu życia pacjenta i dla badanej grupy przyjmuje wartości z zakresu od 92.8 do 99%,

- odchylenie standardowe wysycenia hemoglobiny tlenem (SPO2DEV)

parametr ten określa standardowe odchylenie wysycenia hemoglobiny tlenem i zawiera się w przedziale od 1.5 do 5.5%. Większe wartości odchylenia standardowego oznaczają większe wahania wysycenia w trakcie hospitalizacji, a więc głębsze i częstsze spadki tej wielkości,

- procent czasu, dla którego wysycenie hemoglobiny tlenem jest mniejsze od 85% (LOW85)

parametr ten obliczany jest jako procentowy stosunek czasu, w trakcie którego wysycenie hemoglobiny tlenem jest mniejsze niż 85% (okresy głębszej desaturacji), do całego okresu analizy parametru czyli jednego tygodnia. Przy założeniu, że odstęp czasu pomiędzy transmisją kolejnych próbek z pulsoksymetru jest stały (sytuacja taka ma miejsce w przypadku zastosowanego pulsoksymetru NPB-295 [40], jak również innych pulsoksymetrów [39]), parametr ten wyznaczyć można z następującego wzoru :

$$\mathrm{LOW85} = \frac{n_{85}}{n} \cdot 100\% \tag{4.2}$$

We wzorze tym n_{85} oznacza ilość zebranych w danym okresie analizy poprawnych próbek, dla których wysycenie hemoglobiny tlenem było mniejsze od 85%, zaś n oznacza ilość wszystkich zebranych w tym czasie poprawnych próbek. Parametr LOW85 dla badanej grupy zawiera się w granicach od 0 do 4.8%. Jego niższe wartości oznaczają bardziej stabilny przebieg saturacji bez dłuższych okresów spadków poniżej wartości progowej 85%,

– procent czasu, dla którego wysycenie hemoglobiny tlenem jest większe od 94% (HIGH94)

parametr obliczany podobnie jak LOW85, brany pod uwagę jest czas, w którym wartości wysycenia hemoglobiny tlenem przekraczają próg 94%. Określony wzorem :

$$\mathrm{HIGH94} = \frac{n_{94}}{n} \cdot 100\% \tag{4.3}$$

gdzie n_{94} oznacza ilość zebranych w danym okresie analizy poprawnych próbek, dla których wysycenie hemoglobiny tlenem było większe od 94%, n zaś oznacza ilość wszystkich zebranych w tym czasie poprawnych próbek. Wartości parametru HIGH94 zawierają się w przedziale 32 do 98%. Jego wyższe wartości oznaczają większą ilość i długość trwania przypadków hiperoksji (nadmiernego utlenowania),

– średnia ilość uderzeń serca na minutę (BPMMEAN)

parametr obliczany jako średnia wartość ilości uderzeń serca na minutę w pierwszym tygodniu hospitalizacji. Wartości zawierają się w przedziale 125 do 155 uderzeń na minutę.

– stosunek wartości średniej wysycenia hemoglobiny tlenem dla pierwszych 7 dni do wartości średniej dla pierwszego dnia (SPO2MEAN_TR), stosunek odchylenia standardowego wysycenia hemoglobiny tlenem dla pierwszych 7 dni do odchylenia standardowego wysycenia hemoglobiny tlenem dla 1 dnia (SPO2DEV_TR), stosunek średniej ilości uderzeń serca na minutę dla pierwszych 7 dni do średniej ilości uderzeń serca na minutę dla 1 dnia (BPMMEAN_TR)

parametry te określone zostały jako stosunek wartości obliczonych dla całego pierwszego tygodnia do wartości obliczonych dla pierwszego dnia hospitalizacji. Można je interpretować jako trend danej wielkości, na podstawie której parametr został obliczony. Wartość równa 1 oznacza brak zmiany danego parametru, wartość mniejsza od 1 oznacza spadek danej wartości w stosunku do pierwszego dnia hospitalizacji, wartość większa od 1 oznacza wzrost.

W tabeli 1 przedstawione są wartości maksymalne, minimalne i średnie podstawowych parametrów użytych do predykcji dysplazji. W tabeli 2 porównane zostały wartości średnie dla parametrów w obydwu grupach dzieci – tych, u których stwierdzono dysplazję oskrzelowo-płucną, i tych, u których jej nie stwierdzono.

Tab. 1. Charakterystyka badanej grupy

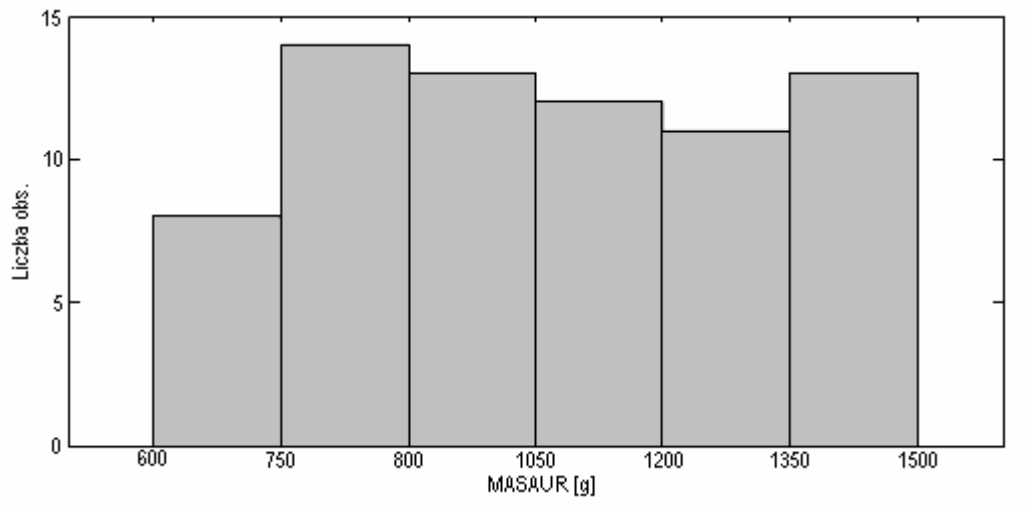
Parametr	Wartość min.	Wartość maks.	Wartość średnia
Masa urodzeniowa [g]	600	1500	1062
Wiek płodowy [tygodnie]	24	32	28.5
Wskaźnik AA	0.05	1	0.28
Średnia wartość SpO_2 [%]	92.78	98.99	95.9
Odchylenie stand. SpO_2 [%]	1.42	5.93	3.3
Parametr LOW85 [%]	0.03	4.78	1.3
Parametr HIGH94 [%]	32	98	73
Średnia ilość uderzeń serca na minutę BPMMEAN	124	156	142
Parametr SPO2MEAN_TR	0.96	1.07	1.00
Parametr SPO2DEV_TR	0.52	2.05	1.08
Parametr BPMMEAN_TR	0.81	1.19	1.03

Tab. 2. Porównanie wartości średnich parametrów w grupie dzieci z BPD i bez BPD

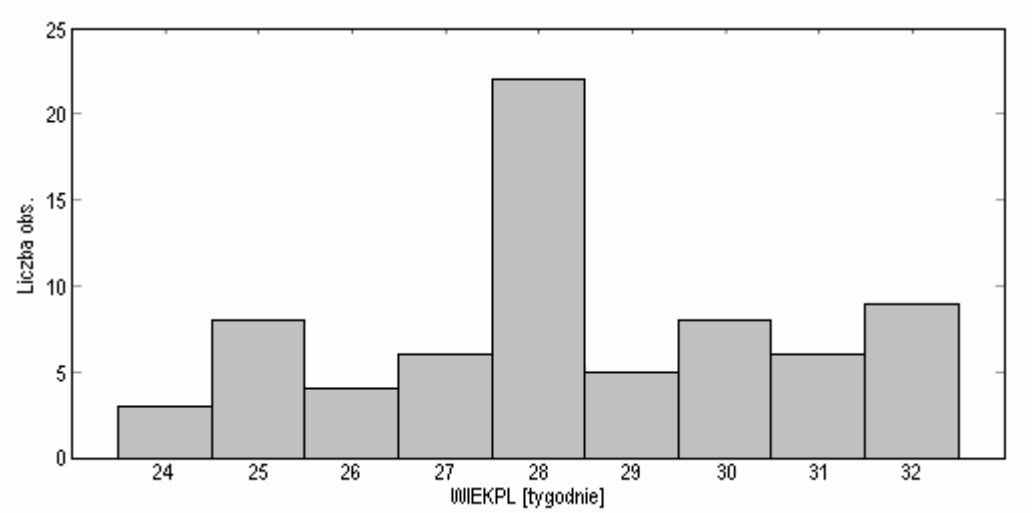
Parametr	Pacjenci bez BPD	Pacjenci z BPD
Masa urodzeniowa [g]	1140	925
Wiek płodowy [tygodnie]	29.4	27
Wskaźnik AA	0.34	0.17
Średnia wartość SpO_2 [%]	96.4	95.1
Odchylenie stand. SpO_2 [%]	3	3.7
Parametr LOW85 [%]	0.03	1.8
Parametr HIGH94 [%]	78	65
Średnia ilość uderzeń serca na minutę	141	146
Parametr SPO2MEAN_TR	1.01	1.00
Parametr SPO2DEV_TR	1.08	1.08
Parametr BPMMEAN_TR	1.02	1.04

4.3. Charakterystyka rozkładów parametrów

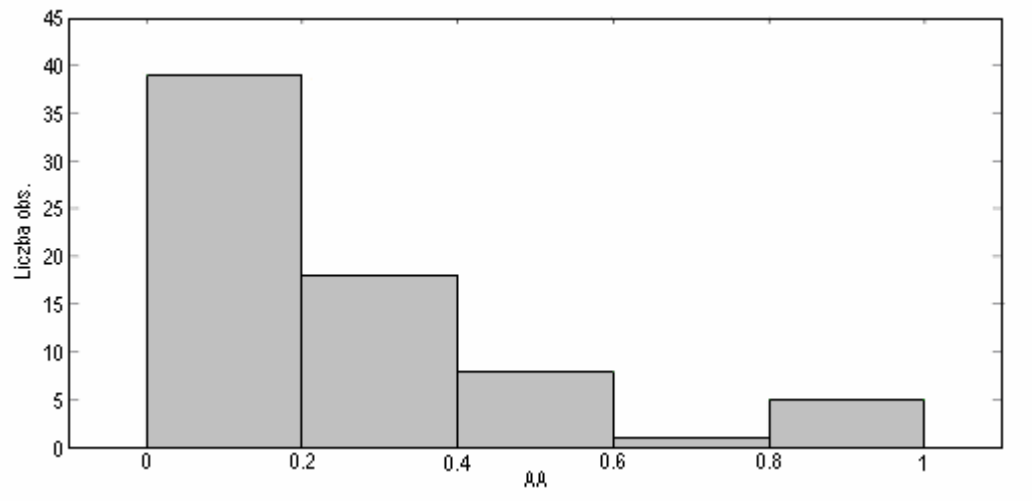
Dokładniejsza analiza obejmująca histogramy rozkładu wartości przyjmowanych przez te parametry przedstawiona jest na poniższych wykresach (rys. 8-18).



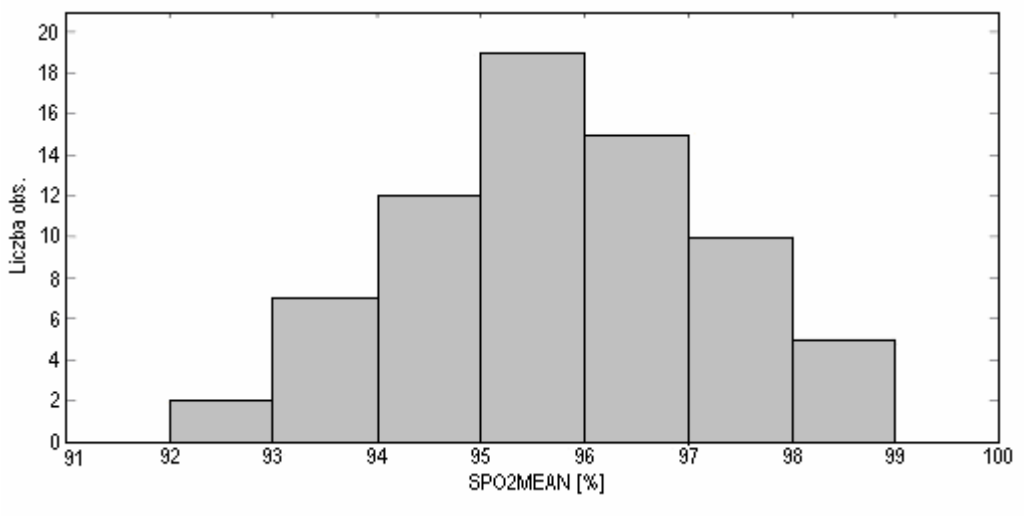
Rys. 8. Histogram parametru „urodzeniowa masa ciała”



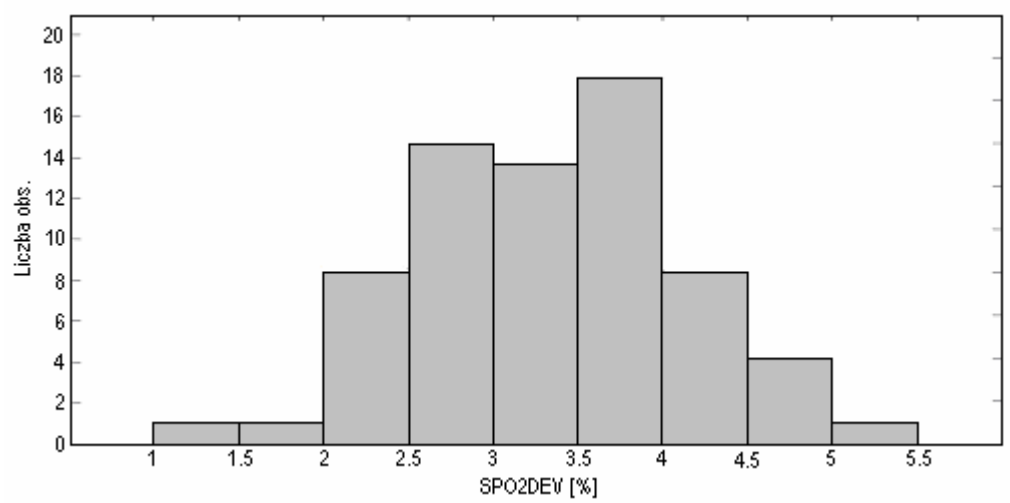
Rys. 9. Histogram parametru „wiek płodowy”



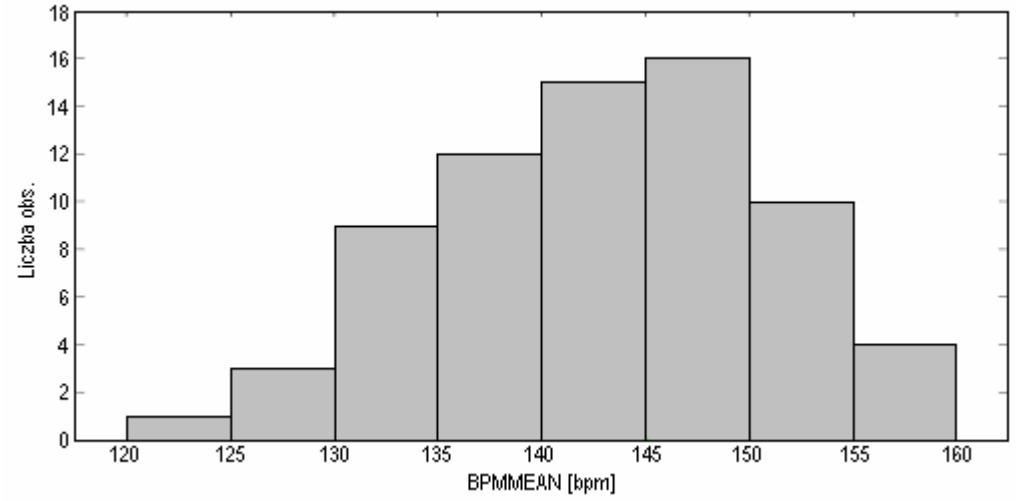
Rys. 10. Histogram parametru „AA”



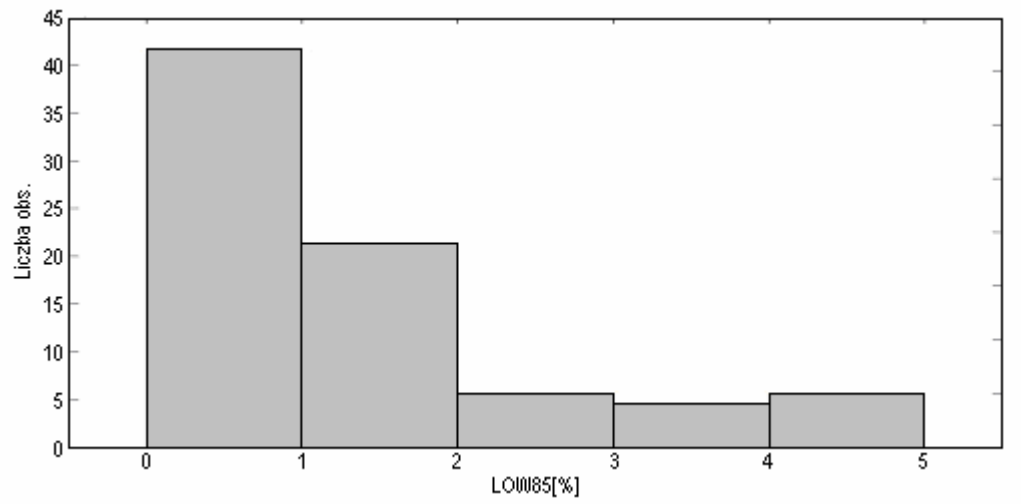
Rys. 11. Histogram parametru „średnia wartość SpO_2”



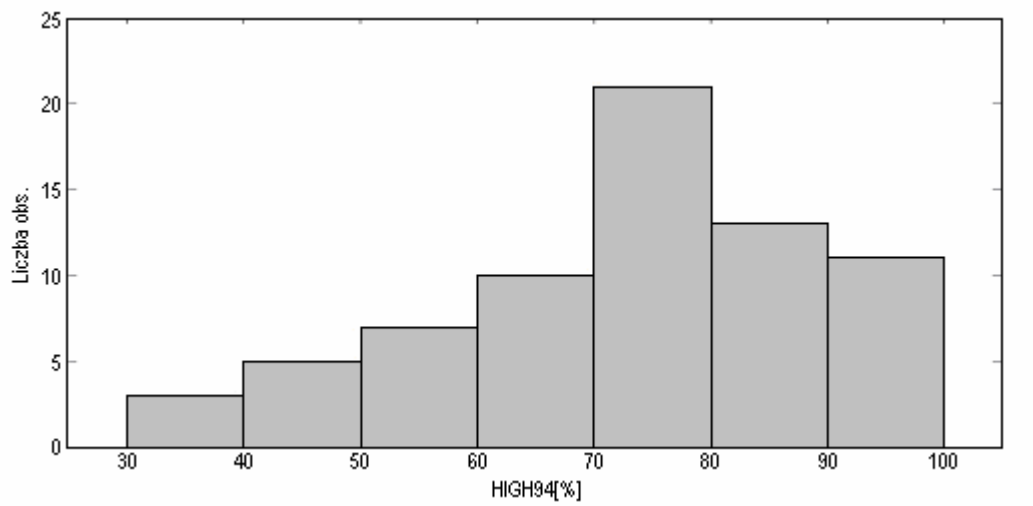
Rys. 12. Histogram parametru „odchylenie standardowe SpO_2”



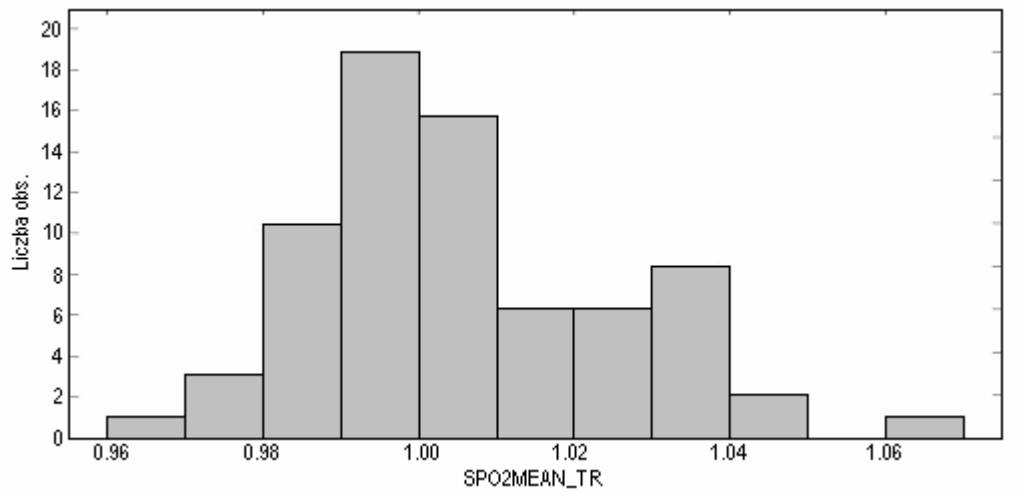
Rys. 13. Histogram parametru „średnia ilość uderzeń serca na minutę”



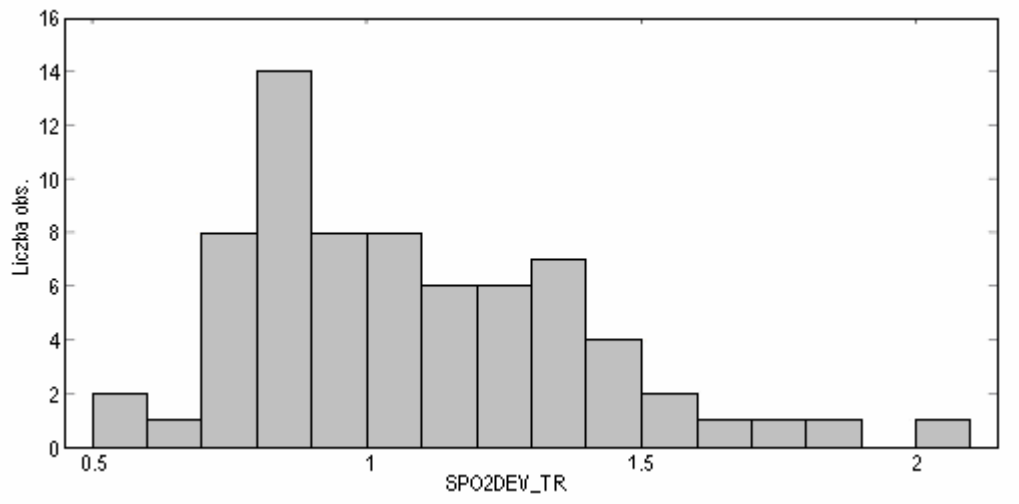
Rys. 14. Histogram parametru „LOW85”



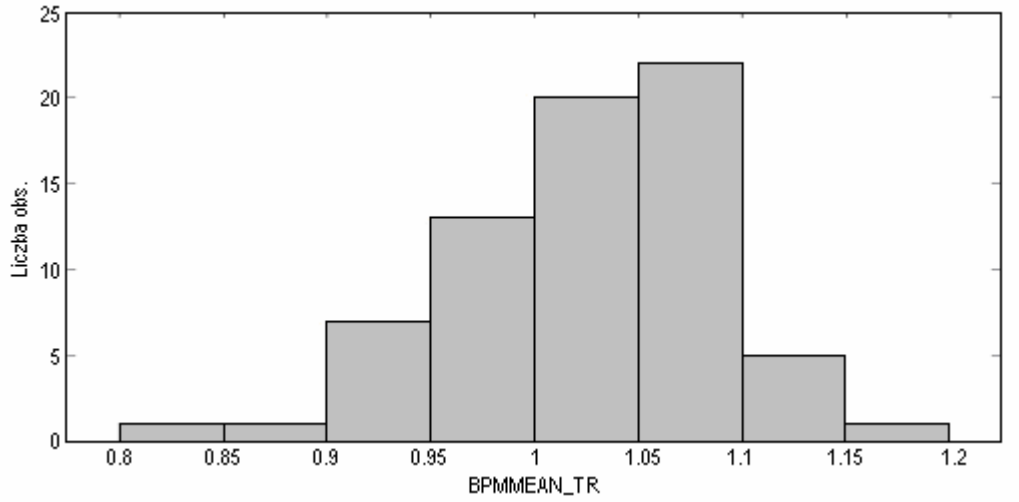
Rys. 15. Histogram parametru „HIGH94”



Rys. 16. Histogram parametru „SPO2MEAN_TR”



Rys. 17. Histogram parametru „SPO2DEV_TR”



Rys. 18. Histogram parametru „BPMMEAN_TR”

Badanie normalności rozpatrywanych rozkładów parametrów ciągłych przeprowadzono przy użyciu testu W Shapiro-Wilka [44]. Posłużono się w tym celu pakietem STATISTICA PL (wersja 6). Wyniki przedstawione są w tab. 3. Hipotezą zerową w tym teście jest hipoteza mówiąca, że próbki pochodzą z populacji o rozkładzie normalnym. Istotność statystyki W na przyjętym poziomie $p \leq 0.05$ oznacza odrzucenie hipotezy zerowej. Jak wynika z tabeli, należy odrzucić hipotezę o normalności rozkładów następujących parametrów : masa urodzeniowa, wiek płodowy, wskaźnik AA, LOW85, HIGH94,SPO2DEV_TR.

Wstępna analiza danych (tab. 2) pokazuje, że populacja dzieci, u których rozwinęła się dysplazja oskrzelowo-płucna, różni się od populacji dzieci, u których dysplazja się nie rozwinęła. Aby przekonać się o istotności statystycznej różnic rozkładu parametrów, można posłużyć się nieparametrycznym testem Manna-Whitney’a [33],[53].

Tab. 3. Wyniki testu normalności W Shapiro-Wilka

Parametr	W (S-W)	p
Masa urodzeniowa	0,96	0,02
Wiek płodowy	0,93	0,002
Wskaźnik AA	0,79	<0,001
Średnia wartość SpO_2	0,99	0,77
Odchylenie stand. SpO_2	0,98	0,27
Parametr LOW85	0,84	<0,001
Parametr HIGH94	0,96	0,02
Średnia ilość uderzeń serca na minutę BPMMEAN	0,97	0,16
Parametr SPO2MEAN_TR	0,97	0,07
Parametr SPO2DEV_TR	0,95	0,006
Parametr BPMMEAN_TR	0.98	0,36

Wyniki testu (Tab. 4) wskazują, że rozkłady wszystkich badanych parametrów oprócz parametrów SPO2DEV_TR, SPO2MEAN_TR oraz BPMMEAN_TR są silnie zróżnicowane w obydwu podgrupach.

Tab. 4. Wyniki nieparametrycznego testu U (Manna-Whitney'a) dla rozpatrywanych parametrów

parametr	U	poziom p
Masa urodzeniowa	289	<0,001
Wiek płodowy	265,5	<0,001
Wskaźnik AA	320,5	0,002
Średnia wartość SpO_2	275	<0,001
Odchylenie stand. SpO_2	302	0,001
Parametr LOW85	316,5	0,002
Parametr HIGH94	281,5	<0,001
Średnia ilość uderzeń serca na minutę BPMMEAN	294,5	<0,001
Parametr SPO2DEV_TR	563	0.91
Parametr SPO2MEAN_TR	457	0.16
Parametr BPMMEAN_TR	530	0.61

4.4. Analiza współliniowości parametrów

W przypadku regresji logistycznej, podobnie jak w regresji liniowej, duże znaczenie ma efekt liniowej zależności zmiennych niezależnych. W przypadku gdy zmienne niezależne (objaśniające) są znacząco skorelowane miedzy sobą, wyniki analizy regresji mogą być niedokładne. W takim przypadku współczynniki regresji mają duże wartości błędów standardowych, przez co znaczna liczba zmiennych niezależnych może być nieistotna. Szacowany efekt danej zmiennej X_j może zmienić wielkość, a nawet kierunek zależnie od

pozostałych zmiennych objaśniających zawartych w testowanym modelu regresji. Warunek, w którym istnieje silna liniowa zależność pomiędzy dwoma zmiennymi niezależnymi, zagrażająca trafności wyników analizy regresji, nazywany jest współliniowością (ang. *collinearity*). W przypadku, gdy jedna ze zmiennych jest liniową kombinacją kilku innych zmiennych, mamy do czynienia z wielowspółliniowością (ang. *multicollinearity*).

Analizę współliniowości w przypadku regresji logistycznej przeprowadza się analogicznie jak w przypadku regresji liniowej [36]. Zazwyczaj efekt współliniowości wyrażany jest poprzez współczynnik VIF (ang. *variance inflation factor*), który wskazuje, o ile wariancje współczynników są zawyżone z powodu zależności liniowych w testowanym modelu. VIF dla danej zmiennej niezależnej X_j zdefiniowany jest jako :

$$\mathrm{VIF_j} = \frac{1}{1 - \mathrm{R_j}^2} \tag{4.4}$$

gdzie R_j^2 jest współczynnikiem wielokrotnej determinacji dla regresji liniowej j-tej zmiennej na pozostałe zmienne objaśniające zawarte w modelu [4]. Współczynnik ten wskazuje, jaka część całkowitej zmienności zmiennej objaśnianej została wyjaśniona poprzez model regresyjny. Współczynnik ten wyznaczany jest ze wzoru [31] :

$$\mathrm{R_j}^2 = \frac{\sum_{i=1}^{N} (\hat{y}_{ij} - \bar{y}_j)^2}{\sum_{i=1}^{N} (y_{ij} - \bar{y}_j)^2} \tag{4.5}$$

gdzie :
N - ilość przypadków,
y_{ij} - i-ta wartość j-tej zmiennej,
$\bar{y}_j$ - średnia arytmetyczna wartość wszystkich przypadków j-tej zmiennej,
$\hat{y}_{ij}$ - i-ta wartość dla j-tej zmiennej wyznaczona przy użyciu regresji liniowej pozostałych zmiennych niezależnych.

Współczynnik VIF_j wskazuje, o ile wariancja szacowanego współczynnika regresji jest podwyższona z powodu współliniowości (wielowspółliniowości) danej zmiennej niezależnej z pozostałymi zmiennymi niezależnymi. Chociaż nie ma uniwersalnie określonej krytycznej wartości dla współczynnika VIF, przyjmuje się wartość $\mathrm{VIF} \geq 4$ jako wskazującą na obecność problemu współliniowości [31], gdyż oznacza to przynajmniej dwukrotne poszerzenie danego przedziału ufności z powodu zależności liniowych. Z tabeli 5 przedstawiającej wartości współczynników VIF_j dla zmiennych niezależnych wykorzystanych do predykcji dysplazji

oskrzelowo-płucnej wynika, że dla czterech zmiennych przekroczona jest wartość krytyczna współczynnika VIF (zmienne te zostały zaznaczone w tabeli).

W tabelach 6 i 7 przedstawione są wartości współczynników wzajemnej korelacji poszczególnych zmiennych niezależnych. Z tabel tych wynika, że występuje znaczna korelacja pomiędzy zmiennymi SPO2MEAN i HIGH94, wykres zależności HIGH94 od SPO2MEAN potwierdza liniową zależność tych parametrów (rys. 19).

Tab. 5. Wartości współczynników VIF dla zmiennych wykorzystywanych do predykcji dysplazji oskrzelowo-płucnej

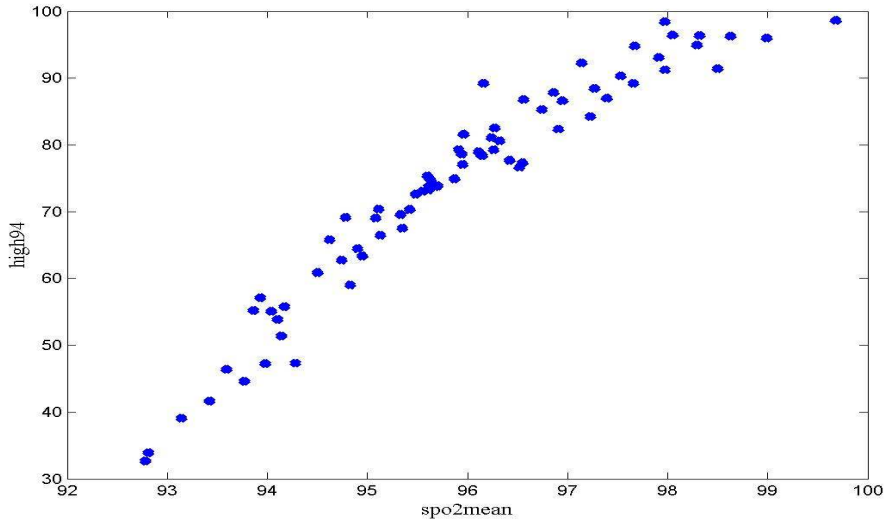
parametr	VIF
BPMMEAN	1.2428
PDA	1.2804
SURFACT	1.5149
RESPIMV	1.7952
AA	1.9848
MASAUR	2.0859
WIEKPL	2.4334
LOW85	**7.1588**
SPO2DEV	**7.6305**
HIGH94	**12.7287**
SPO2MEAN	**14.0832**

Tab. 6. Wartości współczynników korelacji dla zmiennych niezależnych

parametr	SPO2DEV	SPO2MEAN	BPMMEAN	HIGH94	LOW85
SPO2DEV	1	-0.4893	0.2210	-0.4534	0.8954
SPO2MEAN		1	-0.0834	0.9572	-0.5041
BPMMEAN			1	-0.1287	0.0705
HIGH94				1	-0.4620
LOW85					1

Tab. 7. Wartości współczynników korelacji dla zmiennych niezależnych c.d.

parametr	MASAUR	AA	PDA	RESPIMV	SURFACT	WIEKPL
SPO2DEV	-0.0539	-0.4605	0.2028	0.4423	0.4144	-0.1849
SPO2MEAN	0.2451	0.5219	-0.1127	-0.3809	-0.3444	0.2833
BPMMEAN	0.0228	-0.0269	0.0964	0.0904	0.0603	-0.1564
HIGH94	0.2221	0.4851	-0.1138	-0.3401	-0.2962	0.2789
LOW85	-0.0490	-0.3855	0.2510	0.3215	0.4029	-0.0619
MASAUR	1	0.0652	-0.1708	-0.1111	-0.0590	0.6822
AA		1	-0.1817	-0.5624	-0.4839	0.1633
PDA			1	0.2649	0.3032	-0.2177
RESPIMV				1	0.3855	-0.3098
SURFACT					1	-0.0782
WIEKPL						1



Rys. 19. Wykres zależności parametru „HIGH94" od „SPO2MEAN"

Podobna zależność liniowa występuje pomiędzy zmienną LOW85 i SPO2DEV. Usunięcie dowolnej zmiennej z pary SPO2MEAN i HIGH94 oraz LOW85 i SPO2DEV powoduje, że wartości współczynników VIF dla pozostałych zmiennych niezależnych przyjmują wartości mniejsze od przyjętej wartości VIF=4.

Tab. 8. Wartości współczynników VIF dla zmiennych niezależnych po usunięciu parametrów SPO2DEV i HIGH94

parametr	VIF
BPMMEAN	1.0731
PDA	1.2158
SURFACT	1.4970
LOW85	1.5383
RESPIMV	1.6942
SPO2MEAN	1.7887
AA	1.9296
MASAUR	2.0293
WIEKPL	2.2543

Wartości uzyskanych parametrów VIF po usunięciu przykładowo SPO2DEV i HIGH94 przedstawione są w tabeli 8. Oznacza to, że problemem jest współliniowość tych par zmiennych. Wszystkie modele regresyjne wykorzystujące równocześnie parę zmiennych SPO2MEAN i HIGH94 lub LOW85 i SPO2DEV są obciążone opisanym wcześniej błędem wynikającym ze zjawiska współliniowości.

5. Metody

Problem predykcji dysplazji oskrzelowo-płucnej zdefiniować można jako przyporządkowanie danego przypadku chorobowego do jednej z dwóch klas (klasyfikacja binarna). Dokładniej rzecz ujmując, na podstawie wektora wejściowego, definiowanego jako zbiór parametrów medycznych wyznaczanych dla pierwszego tygodnia życia noworodka, określana jest wartość prawdopodobieństwa rozwoju BPD po 28 dniu życia pacjenta. Jeśli ryzyko jest większe od pewnej ustalonej wartości progowej, przewiduje się, że u dziecka rozwinie się BPD. W przeciwnym razie przyjmuje się, że dziecko nie zapadnie na dysplazję. Stąd podstawową cechą wykorzystywanych w pracy metod regresji logistycznej oraz sztucznych sieci neuronowych jest ograniczenie wartości uzyskiwanych na wyjściu do przedziału [0,1].

Formalnie ujmując zagadnienie : jeśli przez X oznaczymy przestrzeń zbioru parametrów medycznych, na podstawie których dokonuje się predykcji dysplazji $X=\{x_1,x_2,x_3,...,x_m\}$, przez Y zbiór klas decyzyjnych $Y=\{0,1\}$, gdzie 0 oznacza brak BPD, a 1 oznacza wystąpienie BPD, to problem predykcji polega na przyporządkowaniu $h: X \rightarrow Y$, w taki sposób, że najpierw określana jest funkcja $f : X \rightarrow [0,1]$ (regresja). Następnie, w zależności od wybranego progu decyzyjnego, przypadek przyporządkowany zostaje do odpowiedniej klasy Y tak, że :

$$h(z) = \Theta(f(z)) \tag{5.1}$$

$$\Theta(x) = \begin{cases} 0 & \text{dla} \quad x < u \\ 1 & \text{dla} \quad x \geq u \end{cases} \tag{5.2}$$

gdzie u - wartość progu decyzyjnego zawierającego się w przedziale [0,1].

5.1. Regresja logistyczna

Ogólnym celem statystycznych metod regresji jest badanie związków pomiędzy wieloma zmiennymi niezależnymi (objaśniającymi) a zmienną zależną (objaśnianą). Problem obliczeniowy, jaki należy rozwiązać, polega na dopasowaniu odpowiedniego modelu regresji do zbioru punktów poprzez estymację parametrów wybranego modelu. Regresja logistyczna jest przykładem modelu regresyjnego nieliniowego. W pracy tej rozważany będzie tylko przypadek, gdy zmienna zależna przyjmuje wartość 0 lub 1, ale niekoniecznie musi to być regułą. Regresja logistyczna może być również użyta wówczas, gdy zmienna zależna przyjmuje jeden z wielu dyskretnych stanów (zmienna wielomianowa) [49]. W przypadku

regresji logistycznej nie określa się założeń co do charakteru i rozkładu zmiennych niezależnych. Mogą być one dowolnego rodzaju i nie muszą mieć rozkładu normalnego lub równej wariancji w każdej z grup. Ta właściwość jest jedną z bardzo ważnych zalet regresji logistycznej. W przypadku predykcji dysplazji oskrzelowo-płucnej właśnie warunek normalności rozkładów zmiennych uniemożliwia zastosowanie innych narzędzi, np. analizy dyskryminacyjnej [49].

Relacja pomiędzy zmiennymi niezależnymi a zmienną objaśnianą ma postać funkcji logistycznej:

$$\Theta(x) = \frac{\exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_N \cdot x_N)}{1 + \exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_N \cdot x_N)} \tag{5.3}$$

gdzie : α – wyraz wolny (stała równania), β - współczynniki przy zmiennych niezależnych. Alternatywna formą równania regresji logistycznej jest :

$$\text{logit}[\Theta(x)] = \ln\left[\frac{\Theta(x)}{1-\Theta(x)}\right] = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_N \cdot x_N \tag{5.4}$$

Funkcja logistyczna przyjmuje wartości z zakresu od 0 do 1. Model może więc opisywać wartości prawdopodobieństwa przyjęcia przez zmienną objaśnianą odpowiedniego stanu.

Ze względu na to, że zmienne niezależne posiadają różne wariancje, do oszacowania parametrów modelu nie można użyć stosowanej w regresji liniowej metody najmniejszych kwadratów, zamiast niej używa się metody największej wiarygodności. Metoda ta polega na maksymalizacji funkcji wiarygodności lub minimalizacji ujemnego logarytmu funkcji wiarygodności. Funkcję wiarygodności dla modelu logistycznego określa się jako :

$$L = \prod_{yi=1} p_i \prod_{yi=0} (1 - p_i) \tag{5.5}$$

gdzie : y_i - wartości obserwowane dla i-tego przypadku, p_i – oczekiwane (przewidywane) prawdopodobieństwo dla i-tego przypadku.

Logarytm funkcji wiarygodności przyjmuje postać :

$$\ln(L) = \sum_{i=1}^{N} \left[y_i \cdot \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i) \right] \tag{5.6}$$

Logarytm naturalny funkcji wiarygodności często mnoży się przez współczynnik -2, otrzymując w rezultacie wielkość zwaną dewiancją (ang. *deviance*). Przyjmuje ona wartości dodatnie i jest tym mniejsza, im lepsze dopasowanie modelu.

$$\mathrm{DEV} = -2 \cdot \ln(\mathrm{L}) \tag{5.7}$$

Dla modelu zerowego (L_0), czyli takiego, który zawiera tylko wyraz wolny (pozostałe współczynniki regresji równe są 0), logarytm wiarygodności modelu zerowego oblicza się jako :

$$\ln(\mathrm{L}_0) = \mathrm{n}_0 \cdot \ln\frac{\mathrm{n}_0}{\mathrm{n}} + \mathrm{n}_1 \cdot \ln\frac{\mathrm{n}_1}{\mathrm{n}} \tag{5.8}$$

gdzie : n_0 jest liczbą obserwacji o wartości 0, n_1 jest liczbą obserwacji o wartości 1, n jest całkowitą liczbą obserwacji.

W celu określenia statystycznej istotności zaobserwowanej różnicy pomiędzy dwoma modelami wykorzystuje się statystykę χ^2. Typowym podejściem w tym przypadku jest wykonanie testu statystycznego. Jako hipotezę zerową wybiera się następującą :

$$\mathrm{H}_0 : \mathrm{logit}\left[\Theta_0(\mathrm{x})\right] = \exp\left(\alpha + \beta_1 \cdot \mathrm{x}_1 + \ldots + \beta_\mathrm{m} \cdot \mathrm{x}_\mathrm{m}\right) \tag{5.9}$$

Hipotezą alternatywną jest :

$$\mathrm{H}_\mathrm{a} : \mathrm{logit}\left[\Theta_\mathrm{a}(\mathrm{x})\right] = \exp\left(\alpha + \beta_1 \cdot \mathrm{x}_1 + \ldots + \beta_\mathrm{m} \cdot \mathrm{x}_\mathrm{m} + \beta_{\mathrm{m}+1} \cdot \mathrm{x}_{\mathrm{m}+1} + \ldots + \beta_\mathrm{k} \cdot \mathrm{x}_\mathrm{k}\right) \tag{5.10}$$

Statystyką testową jest test ilorazu wiarygodności (ang. *likelihood test ratio*) :

$$\Lambda = \frac{\mathrm{L}(\mathrm{H}_0)}{\mathrm{L}(\mathrm{H}_\mathrm{a})} \tag{5.11}$$

Statystyka ta może być zmodyfikowana w ten sposób, aby miała rozkład χ^2 :

$$-2\ln(\Lambda) = -2\left[\ln\left(\mathrm{L}(\mathrm{H}_0)\right) - \ln\left(\mathrm{L}(\mathrm{H}_\mathrm{a})\right)\right] = -2\ln\left(\mathrm{L}(\mathrm{H}_0)\right) - \left[-2\ln\left(\mathrm{L}(\mathrm{H}_\mathrm{a})\right)\right] \tag{5.12}$$

We wzorze tym wyraz $-2\ln\left(\mathrm{L}(\mathrm{H}_0)\right)$ oznacza dewiancję dla hipotezy zerowej, a $-2\ln\left(\mathrm{L}(\mathrm{H}_\mathrm{a})\right)$ dewiancję dla hipotezy alternatywnej. Statystyka taka ma rozkład χ^2 o ilości stopni swobody

równej k-m [23]. Używając tego testu można określić istotność statystyczną spadku dewiancji spowodowanego zastosowaniem parametrów x_{m+1} do x_k.

5.2. Sztuczne sieci neuronowe

Sztuczne sieci neuronowe (SSN) powstały w wyniku badań prowadzonych w dziedzinie sztucznej inteligencji. Szczególne znaczenie miały tutaj te prace, które dotyczyły budowy modeli podstawowych struktur występujących w mózgu. SSN stanowią próbę wykorzystania zjawisk zachodzących w systemach nerwowych do rozwiązywania złożonych zadań, w tym również problemów przemysłowych.

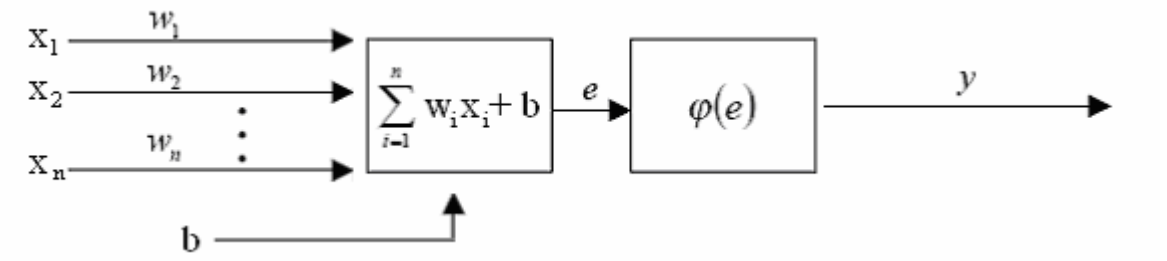
Podstawowym elementem budowy sztucznych sieci neuronowych jest sztuczny neuron, który można rozpatrywać jako przetwornik sygnałowy. Sygnał wyjściowy y związany jest z sygnałami wejściowymi x_i, i=1,2,..,n poprzez funkcję aktywacji φ określającą działanie neuronu [50],[51] :

$$y = \varphi\left(\sum w_i x_i + b\right) = \varphi(e) \tag{5.13}$$

gdzie : w_i – współczynniki zwane wagami synaptycznymi,

b - składnik stały zwany biasem.

Współczynniki te mogą podlegać modyfikacjom w trakcie procesu uczenia sieci neuronowej. Schemat obrazujący zasadę działania sztucznego neuronu przedstawiono na rys. 20 [50].



Rys. 20. Schemat budowy sztucznego neuronu

Do predykcji dysplazji oskrzelowo-płucnej wykorzystano dwa typy sieci neuronowych : dwu- i trójwarstwowe sieci jednokierunkowe zbudowane z neuronów o sigmoidalnej funkcji aktywacji oraz sieci radialne.

Sigmoidalna funkcja aktywacji neuronu opisana jest wzorem :

$$y = \varphi(e) = \frac{1}{1+\exp(-e)} \tag{5.14}$$

Sieci neuronowe złożone z trzech warstw nieliniowych neuronów są w stanie zrealizować dowolne odwzorowanie wiążące w dowolny sposób sygnały wejściowe z wyjściowymi [51].

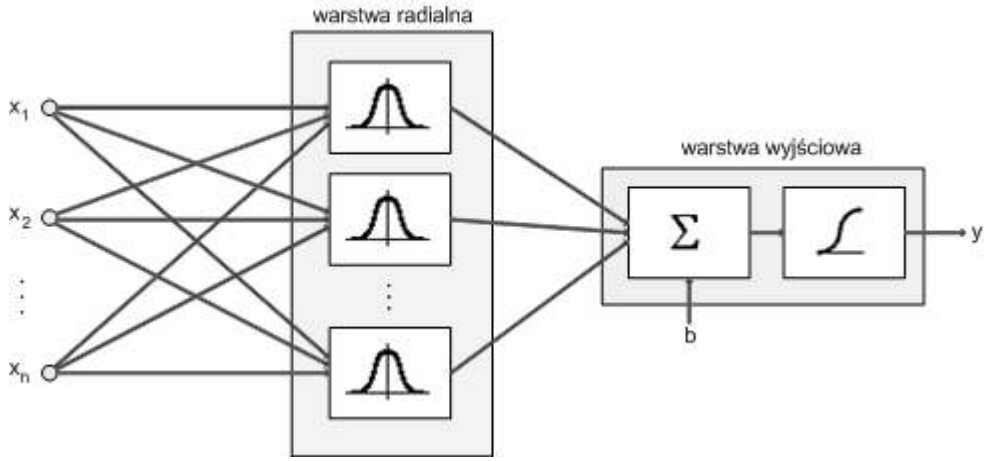
Wykorzystane w pracy sieci jednokierunkowe wielowarstwowe uczone były metodą „z nauczycielem". W metodzie tej prezentowane są zarówno sygnały wejściowe, jak i wyjściowe. Podstawą kryterium optymalizacyjnego jest minimalizacja różnicy pomiędzy wyuczonym a oczekiwanym zachowaniem sieci.

Do uczenia jednokierunkowej wielowarstwowej sieci neuronowej wykorzystano wariant metody wstecznej propagacji błędów z członem momentum. Metoda polega na modyfikacji wag i biasów sieci neuronowej w kolejnych epokach procesu uczenia według następującego wzoru [42] :

$$w_{ij}^{(k)}(t+1) = w_{ij}^{(k)}(t) + 2\eta\delta_i^{(k)}x_j^{(k)}(t) + \alpha\left[w_{ij}^{(k)}(t) - w_{ij}^{(k)}(t-1)\right] \tag{5.15}$$

We wzorze tym $w_{ij}^{(k)}(t+1)$ jest wagą i-tego neuronu, warstwy k, łączącą ten neuron z j-tym sygnałem wejściowym $x_j^{(k)}(t)$. Współczynnik uczenia oznaczony jest jako η, zaś błąd jako $\delta_i^{(k)}$. Współczynnik α, zwany momentum, uzależnia wartość wagi w kroku następnym (t+1) nie tylko od jej wartości w kroku obecnym (t), ale również w kroku poprzednim (t-1). Jego wprowadzenie pozwala przyspieszyć działanie algorytmu na płaskich odcinkach funkcji celu oraz pozwala na opuszczenie obszaru lokalnego minimum tej funkcji.

Drugim rodzajem sieci neuronowych wykorzystanych w pracy były sieci radialne. Sieci te składają się z dwóch warstw neuronów. Warstwa ukryta złożona jest z neuronów o radialnej funkcji aktywacji RBF (ang. *Radial Basis Function*). Warstwę wyjściową stanowi jeden neuron o sigmoidalnej funkcji aktywacji. Schemat sieci radialnej przedstawiony został na rys.21



Rys. 21. Schemat zastosowanej w pracy sieci radialnej

Dla wykorzystanych w pracy sieci neuronowych neuron o radialnej funkcji aktywacji realizuje odwzorowanie [42] :

$$y = \exp\left(-\frac{\|x - c\|^2}{r^2} \right) \tag{5.16}$$

gdzie $\| \cdot \|$ oznacza normę euklidesową, x - wektor wejściowy, c - wagi neuronów warstwy radialnej, zaś parametr r związany jest z szerokością krzywej radialnej. Do określenia szerokości krzywej radialnej w dalszych symulacjach posłużono się parametrem SPREAD określonym wg definicji podanej w [8] : dla wektora wejściowego oddalonego od wektora wag neuronu radialnego o wartość SPREAD wartość przyjmowana na wyjściu neuronu radialnego równa jest 0.5. Parametry SPREAD i r wiąże następująca zależność :

$$r = \frac{SPREAD}{\sqrt{-\log 0.5}} = \frac{SPREAD}{0.8326} \tag{5.17}$$

Ze względu na zastosowanie w warstwie wyjściowej neuronu o sigmoidalnej funkcji aktywacji dobranie macierzy wag warstwy wyjściowej poprzez stosowaną w przypadku liniowego neuronu wyjściowego pseudoinwersję macierzy Greena nie jest możliwe. Zamiast tego zastosowano inne podejście : wagi warstwy radialnej dobrane zostały z wykorzystaniem procesu samoorganizacji [19]. W procesie tym przestrzeń wejściowa podzielona została na grupy przy pomocy algorytmu k-means [34]. Parametr SPREAD określający szerokość krzywej dzwonowej dobierany był eksperymentalnie w taki sposób, aby był mniejszy od rozpiętości danych wejściowych i jednocześnie większy od najmniejszej odległości między

wzorcami uczącymi. Po określeniu centrów i szerokości funkcji radialnych neuronów warstwy ukrytej do wyznaczenia wartości wag neuronu warstwy wyjściowej zastosowano metodę propagacji wstecznej.

5.3. Miary oceny zdolności klasyfikacyjnej modelu

Jednym z podstawowych problemów pojawiających się przy próbie oceny zdolności predykcyjnych zbudowanych modeli jest wybór miary, którą ocenia się tą zdolność. W pracy użyto często stosowaną metodę, jaką jest tzw. macierz pomyłek (ang. *confusion matrix*).

Macierz pomyłek jest w ogólnym przypadku macierzą kwadratową o wymiarach $k \times k$ (gdzie k - ilość klas decyzyjnych), w której wiersze odpowiadają poprawnym klasom decyzyjnym, a kolumny decyzjom przewidywanym przez klasyfikator. Na przecięciu wiersza i oraz kolumny j umieszcza się liczbę przykładów testowych n_{ij} należących do klasy i-tej, a zaliczonych przez klasyfikator do klasy j-tej. Macierz pomyłek dla dwóch klas decyzyjnych przedstawiono w tabeli 9.

Tab. 9. Macierz pomyłek - ogólna postać w przypadku dwóch klas decyzyjnych

		obserwowane klasy rzeczywiste	
		pozytywna	negatywna
przewidywane klasy decyzyjne	pozytywna	Prawdziwie Pozytywne (TP – *True Positives*)	Fałszywie Pozytywne (FP – *False Positives*)
	negatywna	Fałszywie Negatywne (FN – *False Negatives*)	Prawdziwie Negatywne (TN – *True Negatives*)

Nazewnictwo poszczególnych pól w tej macierzy jest inspirowane technikami oceny testów medycznych. W polu oznaczonym jako TP (ang. *true positives*) określa się liczbę poprawnie sklasyfikowanych przykładów z rzeczywistej pozytywnej klasy, FN (ang. *false negative*) oznacza liczbę błędnie sklasyfikowanych przykładów z tej klasy. Dla rzeczywistej klasy negatywnej określa się pole FP (ang. *false negatives*) oznaczające liczbę błędnie sklasyfikowanych przypadków z tej klasy i pole TN (ang. *true negatives*) oznaczające liczbę poprawnie sklasyfikowanych przypadków z tej klasy.

Macierz pomyłek stosuje się generalnie wtedy, gdy rozróżnienie pomiędzy błędnymi klasyfikacjami jest istotne. Przykładowo w medycynie zaliczenie chorego pacjenta do grupy zdrowych jest znacznie bardziej niebezpieczne niż sytuacja odwrotna. W sytuacji, gdy błędy

niesłusznego zaklasyfikowania przypadku do innej klasy mają równe znaczenie, można zastosować często inną, prostszą miarę : trafność, parametr określany w literaturze również jako „miara trafności klasyfikowania" (ang. *accuracy*). Współczynnik ten obliczany jest jako stosunek ilości poprawnie sklasyfikowanych obserwacji do ilości wszystkich obserwacji dla ustalonego progu decyzyjnego u (wzór 5.2). Stosując wcześniejsze oznaczenia :

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5.18}$$

Czasami zamiennie używany jest współczynnik będący uzupełnieniem do jedynki trafności klasyfikowania i nazywany łącznym błędem klasyfikowania (ang. *overall error rate*). Definiowany jest jako :

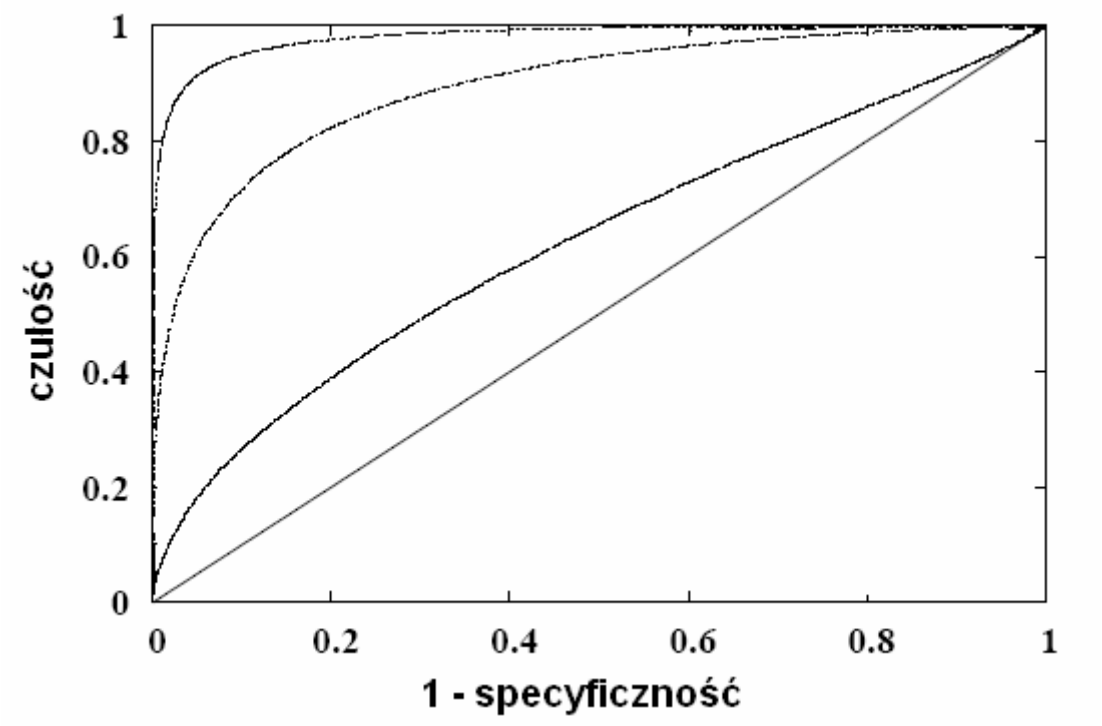
$$\mathrm{ERR} = 1 - \mathrm{ACC} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5.19}$$

Miary te alternatywnie wyraża się procentowo. Im większa wartość trafności, tym skuteczniejszy jest klasyfikator.

Inną bardzo często używaną metodą określania zdolności dyskryminacyjnej modelu klasyfikującego jest analiza ROC (ang. *Receiver Operating Characteristcs*). Polega ona na wyznaczeniu krzywej ROC w ten sposób, że dla określonej wartości progu decyzyjnego u (wzór 5.2) zlicza się liczbę przypadków prawdziwie (TP) i fałszywie pozytywnych (FP), liczbę przypadków prawdziwie (TN) i fałszywie negatywnych (FN), a następnie na podstawie tych wielkości wyznacza się czułość (ang. *sensitivity*) i specyficzność (ang. *specifity*) metody klasyfikacyjnej opartej na badanym parametrze [13]. Czułość i specyficzność metody określa się jako :

- $\text{czułość} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$ (5.20)
- $\text{specyficzność} = \dfrac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$ (5.21)

Zmieniając wartość progu decyzyjnego w całym możliwym zakresie [0,1] i wyznaczając powyższe wielkości można je wykreślić uzyskując krzywą ROC. Przykładowe przebiegi krzywych ROC przedstawione są na rys. 22.



Rys. 22. Przykładowe krzywe ROC

Krzywa pokrywająca się z przekątną oznacza brak zdolności klasyfikujących (czysto losowa klasyfikacja do obydwu grup). Im bardziej krzywa zbliża się do lewej górnej strony wykresu, tym lepsze zdolności dyskryminacyjne wykazuje badany model.

Liczbową miarą tej zdolności jest pole powierzchni pod krzywą ROC, oznaczane symbolem AUC (ang. *Area Under ROC curve*). Wielkość tego parametru zawsze zawiera się w przedziale [0,1]. Dla idealnego klasyfikatora krzywa ROC pokrywa się z lewą i górną osią wykresu, więc pole powierzchni pod krzywą ROC jest równe 1. Krzywa ROC pokrywająca się z przekątną (a więc AUC=0.5) oznacza, że model klasyfikacyjny nie posiada żadnych zdolności dyskryminacyjnych.

Standardowy błąd dla pola powierzchni pod krzywą ROC można oszacować w sposób zaproponowany przez Hanley'a i McNeil [20], korzystając ze wzoru :

$$SE = \sqrt{\frac{AUC(1-AUC) + (n_A - 1)(Q_1 - AUC^2) + (n_N - 1)(Q_2 - AUC^2)}{n_A n_N}} \qquad (5.22)$$

gdzie :

$n_A = TP + FN$ - ilość przypadków rzeczywistej klasy pozytywnej,

$n_N = FP + TN$ - ilość przypadków rzeczywistej klasy negatywnej,

wartości Q_1 i Q_2 mogą być oszacowane z zależności [20] :

$$Q_1 = \frac{AUC}{2 - AUC} \tag{5.23}$$

oraz

$$Q_2 = \frac{2AUC^2}{1 + AUC} \tag{5.24}$$

Należy zauważyć, że wartości pojawiające się w macierzy pomyłek, jak również wartości trafności i łącznego błędu klasyfikowania, zależą od wyboru wartości progu decyzyjnego u, natomiast wyniki uzyskane przy zastosowaniu analizy ROC od tego progu nie zależą.

5.4. Metody wyboru optymalnego podzbioru zmiennych niezależnych

Problem poszukiwania optymalnego podzbioru zmiennych niezależnych modelu, zwany też selekcją cech (ang. *feature selection*, *variable selection*), jest jednym z podstawowych problemów procesu zgłębiania danych (ang. *data mining*). Konieczność redukcji ilości zmiennych związana jest z problemem wymiarowości polegającym na konieczności dostarczenia dostatecznej liczby wzorców uczących, aby wypełnić nimi w odpowiednim stopniu przestrzeń sygnałów wejściowych (tzw. „klątwa wymiarowości" - ang. *curse of dimensionality*) [3],[30]. Szacuje się, że dla przestrzeni zawierającej N cech, zbiór uczący powinien zawierać około 2^N elementów [9]. Oznacza to, że przy ustalonej liczbie wzorców uczących należy wyeliminować część cech, aby uzyskać poprawną reprezentatywność zbioru uczącego. Ponadto, jak wynika z analizy zawartej w rozdziale 4.4 dotyczącym współliniowości parametrów, nie wszystkie zmienne mogą zostać użyte w modelu regresyjnym, gdyż skutkuje to zwiększeniem się błędu wynikowego. Niektóre ze zmiennych charakteryzują się lepszą zdolnością predykcyjną, inne zaś gorszą. Optymalny podzbiór jest zatem podzbiorem zawierającym jak najmniejszą ilość zmiennych tak dobranych, aby model predykcyjny charakteryzował się jak największą jakością klasyfikacji.

Najprostszą możliwą metodą wyboru optymalnego podzbioru zmiennych jest przeszukanie wszystkich możliwych kombinacji zmiennych. Jak łatwo zauważyć, ilość koniecznych do zbudowania, nauczenia i przetestowania modeli równa jest :

$$n = 2^w \tag{5.25}$$

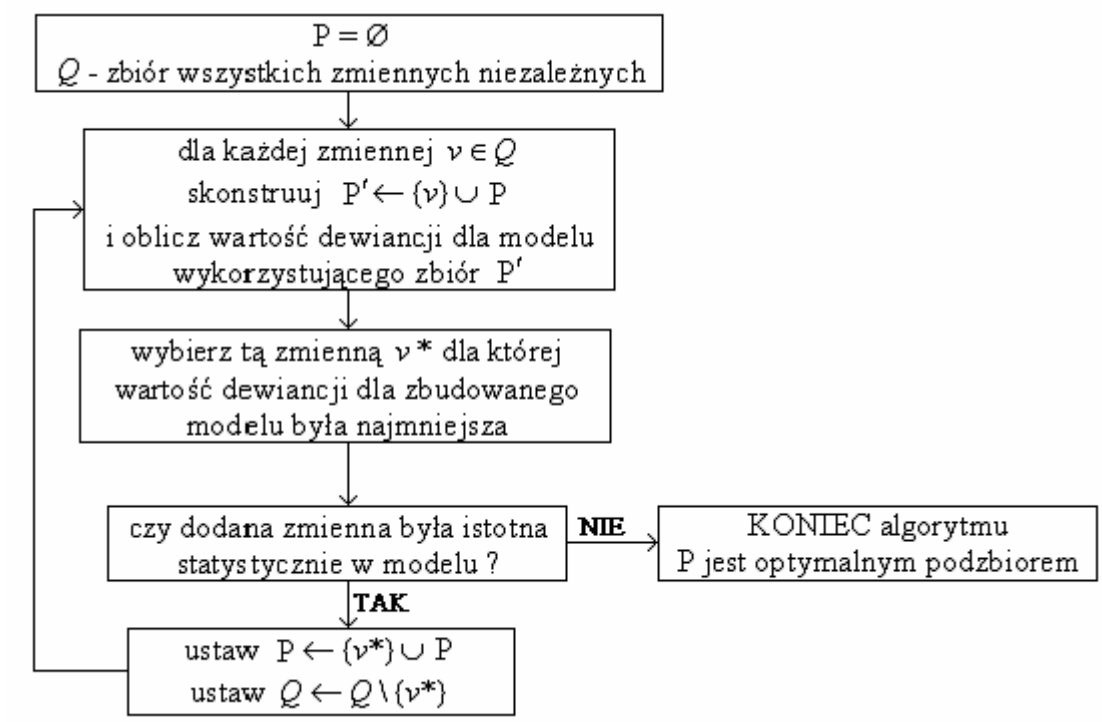
gdzie w – ilość zmiennych objaśniających. Ilość niezbędnych do wykonania operacji i wiążąca się z tym czasochłonność całej metody rośnie wykładniczo wraz ze wzrostem ilości parametrów, dlatego można ją zastosować jedynie w przypadku stosunkowo niewielkiej

liczby zmiennych. W pracy tej ilość cech wybranych wstępnie do predykcji dysplazji oskrzelowo-płucnej w=14, więc ilość wszystkich niezbędnych do przeszukania modeli n=16384. Ilość ta jest stosunkowo niewielka, co pozwala na wykorzystanie metody przeszukania wszystkich możliwych kombinacji i dla metody regresji logistycznej uzyskanie wyników w odpowiednim do oczekiwań czasie.

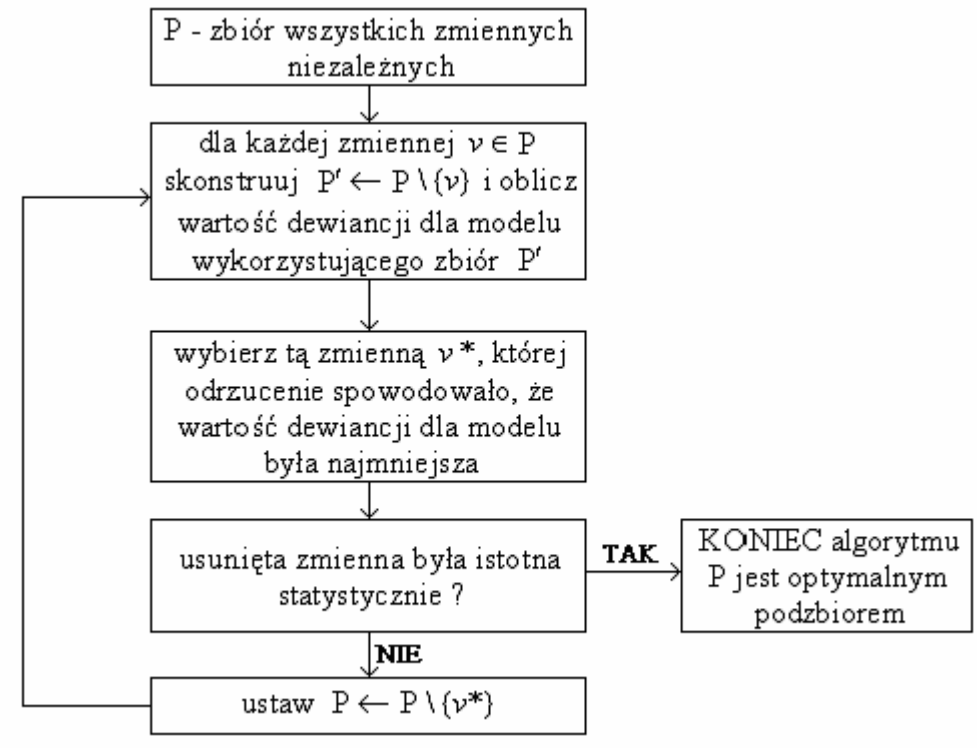
Rozważając możliwość uzyskania większej ilości danych zarówno z systemu rejestracji danych medycznych, jak i z bazy danych NIS, należy przewidzieć konieczność zastosowania innych metod poszukiwania optymalnego podzbioru tych zmiennych. Można w tym celu wykorzystać algorytmy heurystyczne. Algorytmem heurystycznym lub heurystyką określa się metodę twórczego rozwiązywania problemów zarówno logicznych, jak i matematycznych przez eksperyment, metodą prób i błędów bądź odwołaniem do analogii [42]. Stosuje się ją w sytuacji, gdy metoda rozwiązania nie jest znana lub jest złożona i czasochłonna. Podstawową cechą heurystyki jest możliwość zmniejszenia kosztów obliczeniowych i przyspieszenia znalezienia rozwiązania. Metody heurystyczne należą do podstawowych narzędzi sztucznej inteligencji. Do najprostszych heurystyk służących do wyszukiwania optymalnego modelu należą metody selekcji postępującej (ang. *forward selection*) i eliminacji wstecznej (ang. *backward elimination*) [60].

W procedurze selekcji postępującej rozpoczyna się od pustego zbioru zmiennych niezależnych. W każdym kolejnym kroku metody dodając jedną zmienną budowany jest kolejny model i oceniana jego zdolność klasyfikacyjna. Jeśli w danym kroku zbudowany model odpowiada założonym oczekiwaniom, przerywa się procedurę. Uzyskany zbiór zmiennych jest zbiorem optymalnym. W przypadku metody eliminacji wstecznej procedurę rozpoczyna się od pełnego zbioru zmiennych. W kolejnych iteracjach usuwa się z tego zbioru jedną zmienną i ocenia zdolność klasyfikacyjną modelu. Jeśli dany model odpowiada założonym oczekiwaniom, algorytm jest przerywany, a uzyskany zbiór zmiennych jest zbiorem optymalnym. Algorytm selekcji postępującej z racji tego, że rozpoczynany jest od najmniej licznego zbioru zmiennych, jest szybszy od algorytmu eliminacji wstecznej [24].

Do wyboru optymalnego podzbioru zmiennych zastosowanych do predykcji dysplazji oskrzelowo-płucnej procedurę selekcji postępującej użyto w ten sposób, że rozpoczynając od jednej zmiennej niezależnej dodawano po kolei zmienne, dla których spadek dewiancji był największy (rys. 23). Algorytm ten zatrzymuje się, gdy poprawa dewiancji jest zbyt mała (oznaczająca dodanie zmiennej nieistotnej statystycznie).



Rys. 23. Algorytm selekcji postępującej



Rys. 24. Algorytm eliminacji wstecznej

Metodę eliminacji wstecznej (rys. 24) zastosowano podobnie, z tym że proces rozpoczyna się od pełnego modelu, zawierającego wszystkie zmienne niezależne, po czym usuwane są z modelu kolejno te zmienne, których brak nie prowadzi do większego od ustalonego wzrostu

dewiancji (czyli oznacza to usuwanie zmiennych nieistotnych statystycznie). Metody te są często używane (i zaimplementowane w większości programów statystycznych) ze względu na szybkość ich działania i niewielką ilość iteracji potrzebnych do wykonania. Niestety okupione to jest uzyskiwaniem wyników w pewnych przypadkach dalekich od rozwiązań optymalnych. Ponadto każda z tych metod może w rezultacie prowadzić do uzyskania różniących się między sobą rozwiązań [25].

W przypadku opisanym w pracy do predykcji dysplazji oskrzelowo-płucnej zastosowano 14 zmiennych niezależnych. Taka ich liczba umożliwia przeszukanie w zadowalającym przedziale czasu wszystkich możliwych modeli. W przyszłości możliwe będzie zgromadzenie większej ilości danych oraz zastosowanie większej ilości zmiennych objaśniających. To spowoduje jednak, że operacja przeszukania wszystkich modeli będzie zbyt czasochłonna. Dlatego do rozwiązania tego problemu godnym uwagi algorytmem heurystycznym jest algorytm genetyczny. Algorytm genetyczny oparty jest na mechanizmach doboru naturalnego oraz dziedziczności. Przykład wykorzystania tego algorytmu do wyszukiwania optymalnych podzbiorów modelu wykorzystującego regresję logistyczną można znaleźć w pracy S. Vinterbo i L. Ohno-Machado [56]. Natomiast w pracy J. Yanga i V. Honavara [61] można znaleźć opis zastosowania algorytmu genetycznego do klasyfikacji przy użyciu sieci neuronowych.

Algorytm genetyczny jest metodą poszukiwania optymalnego zbioru zmiennych objaśniających. Umożliwia znalezienie w zadanym czasie rozwiązania (rozwiązań) ze znacznie większym prawdopodobieństwem uzyskania wyniku optymalnego (lub bardzo zbliżonego do optymalnego) w porównaniu do metod selekcji postępującej czy eliminacji wstecznej [37].

W przypadku stosowania algorytmu genetycznego do wyszukiwania optymalnego podzbioru zmiennych niezależnych, pierwszym krokiem jest odpowiednia reprezentacja rzeczywistych danych w postaci chromosomu. Zbiór zmiennych objaśniających wykorzystanych do predykcji dysplazji oskrzelowo-płucnej zakodowany został jako binarny ciąg 14-elementowy. Każdej zmiennej niezależnej odpowiada jeden bit tego ciągu (tab. 10). Wystąpienie jedynki na danej pozycji w chromosomie oznacza, że odpowiednia zmienna objaśniająca jest reprezentowana w podzbiorze uczącym. Natomiast obecność zera eliminuje tę zmienną z podzbioru.

Tab. 10. Reprezentacja zmiennych niezależnych w postaci chromosomu

numer bitu w chromosomie	zmienna niezależna
1	SPO2DEV_TR
2	SPO2MEAN_TR
3	BPMMEAN_TR
4	SPO2DEV
5	SPO2MEAN
6	BPMMEAN
7	HIGH94
8	LOW85
9	MASAUR
10	AA
11	PDA
12	RESPIMV
13	SURFACT
14	WIEKPL

Istota algorytmu genetycznego polega na realizowaniu procesu podobnego do naturalnej selekcji zachodzącej w przyrodzie podczas ewolucji. Schemat blokowy algorytmu genetycznego przedstawiony jest na rys.25 [16].

Wybór początkowej populacji chromosomów dokonywany jest poprzez losowanie ze zbioru wszystkich 16384 możliwych przypadków. Do oceny przystosowania chromosomów w populacji i ich selekcji zastosowano kryteria posiadające własność faworyzowania modeli o mniejszej ilości zmiennych niezależnych, charakteryzujących się równocześnie maksymalną zdolnością predykcyjną. Kryteriami tymi były: kryterium informacyjne Akaike (AIC - ang. *Akaike Information Criterion*) oraz kryterium informacyjne Schwarza (BIC - ang. *Bayesian Information Criterion*). Kryterium informacyjne AIC [1] dla modelu regresji logistycznej określone jest wzorem :

$$AIC = DEV + 2w \tag{5.26}$$

gdzie :

DEV – dewiancja modelu (obliczana wg wzoru 5.7), w – ilość parametrów modelu.

W przypadku, gdy liczba parametrów przekracza n/40 (n – liczba przypadków), stosuje się zmodyfikowany wzór na AIC [55] :

$$AIC_0 = DEV + \frac{2nw}{n - w - 1} \tag{5.27}$$

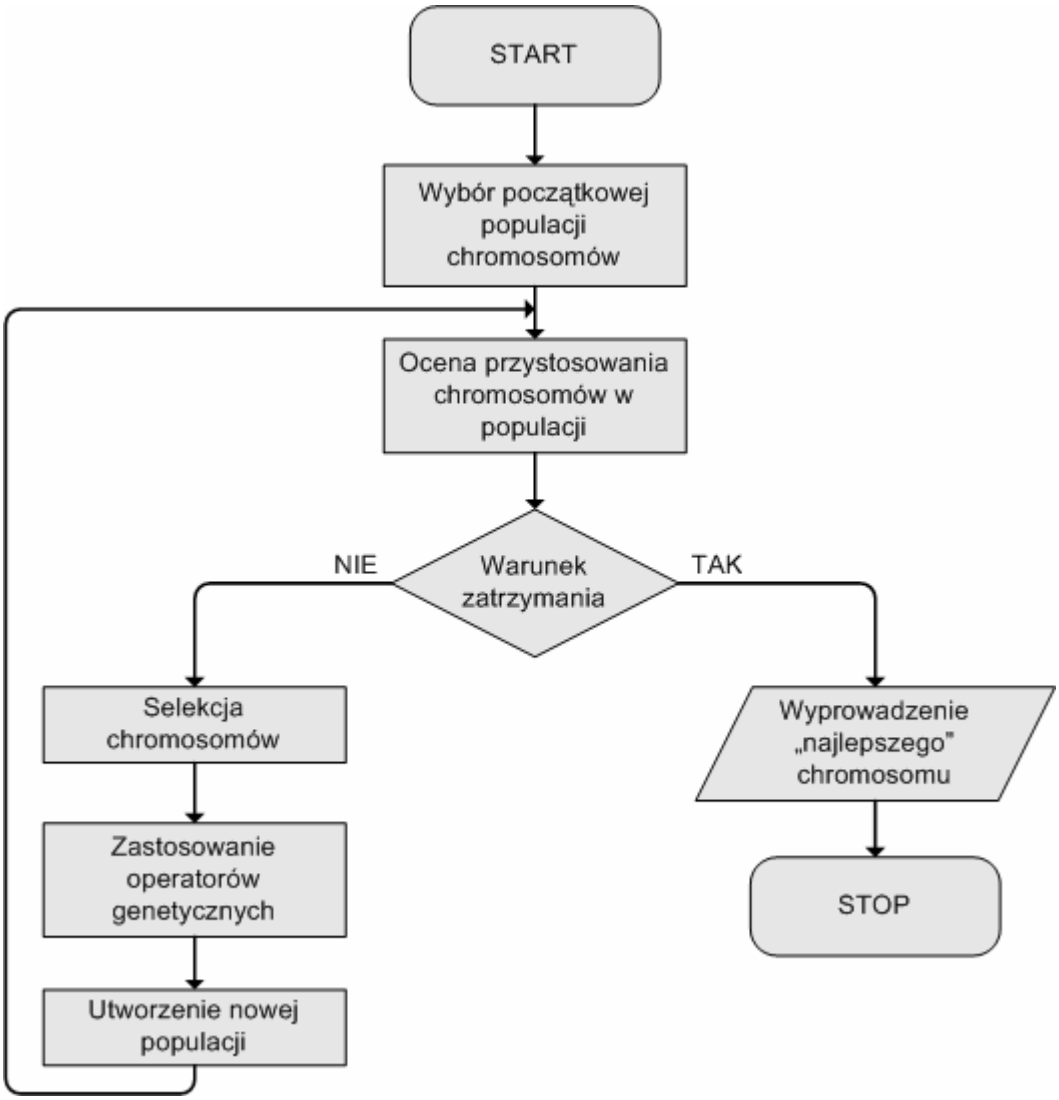
gdzie n jest ilością przypadków (obserwacji).

Model regresyjny dla optymalnego podzbioru zmiennych objaśniających powinien mieć możliwie jak najmniejszą wartość kryterium AIC. Charakteryzuje się on wówczas niewielką wartością dewiancji przy równocześnie niewielkiej ilości zmiennych niezależnych.

W przypadku kryterium informacyjnego BIC, w porównaniu do poprzedniego kryterium dodany został czynnik uzależniający jego wartość od ilości obserwacji (przypadków). Wzór określający BIC [43] (oznaczenia jak we wcześniejszych wzorach) :

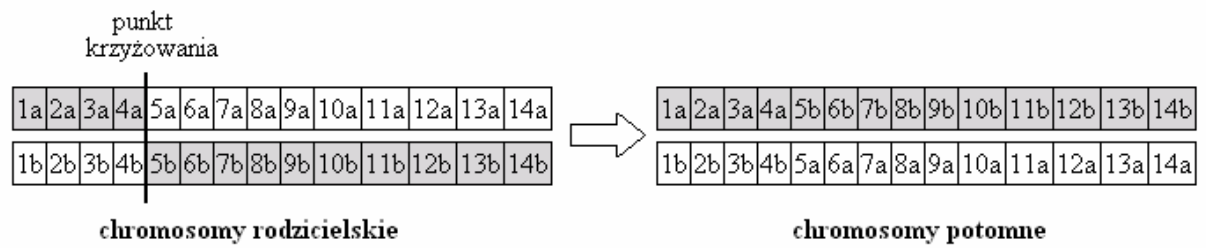
$$BIC = DEV + p \cdot \log(n) \tag{5.28}$$

Podobnie jak w przypadku AIC, również stosując kryterium BIC dąży się do minimalizacji jego wartości przy wyborze optymalnego podzbioru zmiennych.

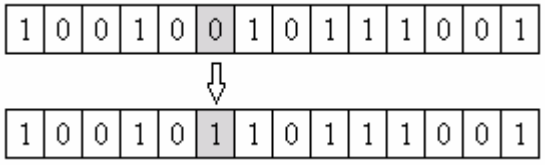


Rys. 25. Schemat blokowy algorytmu genetycznego

W celu utworzenia nowej populacji do chromosomów wybranych w procesie selekcji zastosowano operatory genetyczne : krzyżowania (ang. *crossover*) oraz mutacji (ang. *mutation*). Krzyżowanie polega na kojarzeniu chromosomów z wybranej populacji rodzicielskiej w pary w sposób losowy, zgodnie z prawdopodobieństwem krzyżowania p_k. Następnie dla każdej pary wylosowano punkt rozcięcia i skonstruowano dwa chromosomy potomne wg reguły przedstawionej na rys. 26. Na rysunku tym literami a i b oznaczono chromosomy rodzicielskie, zaś cyframi 1 do 14 wartości kolejnych genów w tych chromosomach. Mutacja, zgodnie z prawdopodobieństwem mutacji p_m, dokonuje zmiany wartości genu w chromosomie na przeciwną (tzn. z 0 na 1 lub z 1 na 0) - rys.27.



Rys. 26. Zasada działania operatora krzyżowania



Rys. 27. Zasada działania operatora mutacji

Prawdopodobieństwo p_m jest znacznie niższe niż p_k. Wybór genu podlegającego mutacji dokonywany jest poprzez wylosowanie liczby z przedziału [0,1] dla każdego genu i wybraniu do mutacji tych genów, dla których wylosowana liczba jest mniejsza lub równa prawdopodobieństwu p_m [42].

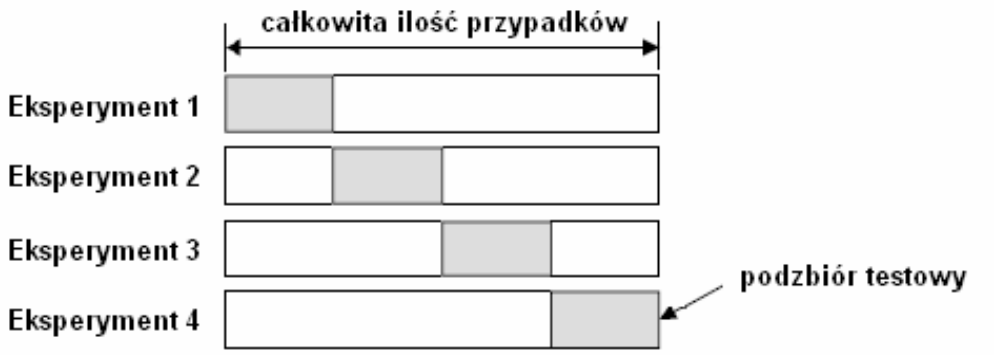
Warunkiem zatrzymania algorytmu jest przekroczenie określonej liczby iteracji lub stwierdzenie braku poprawy wartości kryterium dla najlepszego chromosomu w populacji w ciągu kilku kolejnych iteracji algorytmu.

5.5. Metody walidacji modelu predykcyjnego

W przypadku rzeczywistych problemów predykcyjnych zazwyczaj nie dysponuje się zbiorem wszystkich możliwych przypadków. Dostępny zbiór o ograniczonych rozmiarach

należy traktować jako próbę losową pobraną z populacji wszystkich możliwych przykładów. Taki ograniczony liczbowo zbiór przykładów musi być użyty do estymacji miar oceny klasyfikatora. W przypadku stosowania tych samych zestawów danych do dopasowania modelu (treningu) i testowania jego jakości klasyfikacji, otrzymane wyniki obarczone są błędem związanym ze zbytnim dopasowaniem modelu do danych, czyli przeuczeniem (ang. *overfitting*). W celu określenia nieobciążonego błędu klasyfikacji często stosuje się metodę walidacji krzyżowej [21].

W pracy zastosowano metodę k-krotnej walidacji krzyżowej (ang. *k-fold cross-validation*). Polega ona na losowym podziale zbioru na k rozłącznych podzbiorów (rys. 28). Jeden z tych podzbiorów używany jest jako zbiór testowy, zaś pozostałe k-1 podzbiorów w sumie stanowią zbiór treningowy. Procedura powtarzana jest dla każdego k. Tym sposobem każdy z elementów zbioru jest wykorzystywany zarówno w podzbiorze testowym, jak i w podzbiorach treningowych (w innej iteracji). Idea walidacji krzyżowej pojawiła się już w pracach Lachenbrucha i Mickeya [29] oraz Mostellera i Tukeya [38] (1968 r.), a następnie rozwijana była w wielu innych pracach [11],[14],[47],[57],[58].



Rys. 28. Walidacja krzyżowa k-krotna (dla k=4)

Wybór wartości parametru k, czyli ilości podzbiorów, jest w praktyce wynikiem kompromisu pomiędzy z jednej strony wzrostem wariancji, a z drugiej wzrostem tendencji błędu (ang. *bias-variance tradeoff problem*) [5],[17],[32]. Jako optymalną wartość przyjmuje się ok. 10, w pracy przyjęto k=14 ze względu na możliwość prostego podziału zbioru 70 przypadków na podzbiory uczące i testowe.

6. Wyniki

Do celów obliczeniowych wykorzystane zostało środowisko MATLAB 6.5 (R13) zainstalowane na serwerach obliczeniowych Katedry Automatyki AGH. Do predykcji dysplazji oskrzelowo-płucnej przy użyciu regresji logistycznej użyto funkcji „glmfit", znajdującej się w zestawie standardowych narzędzi statystycznych pakietu MATLAB („*Statistics Toolbox*"). Funkcje umożliwiające konstrukcję, uczenie i testowanie sieci neuronowych, wykorzystane w tej pracy, znajdują się w zestawie narzędzi „sieci neuronowe" („*Neural Network Toolbox*") tego środowiska. Wszystkie pozostałe użyte w pracy funkcje i implementacje algorytmów, takich jak m.in. walidacja krzyżowa, algorytm genetyczny, analiza ROC, zostały napisane przez autora.

6.1. Predykcja dysplazji przy użyciu regresji logistycznej

W pierwszym kroku przeanalizowany został pełny model, zawierający wszystkie zmienne niezależne. Chociaż, jak pokazano w rozdziale 4.4, model ten obarczony jest błędem wynikającym ze zjawiska współliniowości, jednak można dla porównania wyznaczyć wartości podstawowych parametrów określających jakość klasyfikacji takiego modelu. Następnie przedstawione zostały wyniki poszukiwania optymalnego podzbioru zmiennych objaśniających (niezależnych). Zastosowane zostały opisane wcześniej proste metody heurystyczne: selekcji postępującej i eliminacji wstecznej. Ze względu na stosunkowo niewielką ilość wszystkich możliwych do skonstruowania kombinacji zmiennych niezależnych, dla porównania została wykorzystana również metoda przeszukania wszystkich możliwych modeli. W dalszym kroku przy użyciu walidacji krzyżowej przebadana została zdolność predykcyjna uzyskanych optymalnych modeli regresyjnych.

6.1.1. Analiza pełnego modelu z użyciem wszystkich parametrów

Pełny model zawiera wszystkie dostępne zmienne niezależne, ilość przypadków (pacjentów) n = 70. Dopasowanie modelu i testowanie jakości predykcji dokonywane było na tych samych danych – wszystkich zmiennych objaśniających i wszystkich pacjentach.

Dewiancję dla modelu bez parametrów (zawierającego tylko wyraz wolny) wyznaczyć można ze wzorów 5.7 i 5.8 i równa jest $DEV_0 = 92.36$

Wartości podstawowych parametrów określających jakość predykcji przedstawione są w tabeli 11.

Tab. 11. Wartości podstawowych parametrów określających jakość predykcji dla pełnego modelu (z wykorzystaniem wszystkich dostępnych zmiennych niezależnych)

Parametr	Wartość
AUC (pole powierzchni pod krzywą ROC)	$0.99 \pm 0.01^{*}$
DEV (dewiancja)	20.86
ACC (trafność klasyfikowania) [%]	97.1
Kryterium AIC	59.75
Kryterium BIC	84.59

* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

W tabeli 12 przedstawione są wartości poziomu istotności poszczególnych zmiennych niezależnych. Pogrubioną czcionką zaznaczono parametry, które są istotne przy przyjętym poziomie p=0.05 (5%).

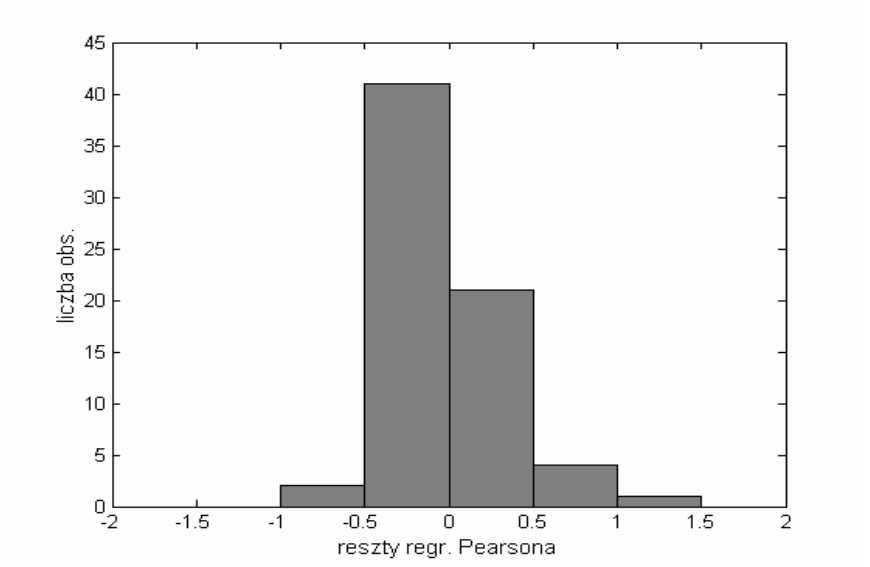
Na rys. 29 przedstawiony jest histogram reszt regresyjnych Pearsona wyznaczanych ze wzoru :

$$e_i = \frac{t_i - y_i}{\sqrt{y_i(1-y_i)}} \tag{6.1}$$

gdzie t_i - wartości rzeczywiste zmiennej zależnej, y_i - wartości przewidywane przez model regresyjny. W dobrze dopasowanym modelu wartość absolutna reszt nie przekracza wartości 1.96 [36].

Tab. 12. Wartości poziomu istotności dla wyrazu wolnego i poszczególnych zmiennych niezależnych dla pełnego modelu

Parametr	poziom istotności
wyraz wolny	0.232279
Parametr SPO2DEV_TR	**0.041200**
Parametr SPO2MEAN_TR	**0.033023**
Parametr BPMMEAN_TR	0.405792
Odchylenie stand. SpO_2 [%] (SPO2DEV)	0.140122
Średnia wartość SpO_2 [%] (SPO2MEAN)	0.136931
Średnia ilość uderzeń serca na minutę (BPMMEAN)	**0.014172**
Parametr HIGH94 [%] (HIGH94)	0.129490
Parametr LOW85 [%] (LOW85)	0.336904
Masa urodzeniowa [g] (MASAUR)	0.415483
Wskaźnik AA	0.864381
PDA	0.158891
Użycie respiratora (RESPIMV)	**0.031502**
Podanie surfaktantu (SURFACT)	0.141321
Wiek płodowy [tyg] (WIEKPL)	**0.043422**



Rys. 29. Histogram reszt regresyjnych Pearsona dla pełnego modelu

Macierz pomyłek dla analizowanego modelu przedstawiona jest w tab. 13.

Tab. 13. Macierz pomyłek dla pełnego modelu

	Przewidywany brak BPD	Przewidywane BPD		
Obserwowany brak BPD	43	1	specyficzność	97.7 %
Obserwowane BPD	1	25	czułość	96.2 %

Model, w którym większość zmiennych niezależnych nie jest istotna statystycznie, nie jest modelem optymalnym pomimo tego, że charakteryzuje się dużymi wartościami trafności klasyfikowania, czułości, specyficzności, pola powierzchni pod krzywą ROC oraz niską wartością dewiancji. Wskaźniki te bowiem wyznaczane zostały przy użyciu tego samego zestawu danych zarówno przy dopasowywaniu modelu, jak i ocenie jego zdolności predykcyjnej. Ponadto jak to zostało opisane w rozdziale 4.4, wyniki zwracane przez ten model obarczone są błędami wynikającymi ze współliniowości.

6.1.2. Metoda selekcji postępującej

Zgodnie z algorytmem podanym w rozdziale 5.4, rozpoczynając od parametru mającego najlepszą zdolność dyskryminacyjną względem BPD konstruuje się model regresji logistycznej, następnie dodaje po jednym parametrze tak, aby spadek dewiancji był jak największy. Za każdym razem należy sprawdzić, czy wzrost dewiancji jest znaczący statystycznie (różnica dewiancji ma rozkład χ^2 o jednym stopniu swobody, ponieważ kolejne modele różnią się jednym parametrem), zakładając 5% poziom istotności. Optymalnym

modelem wybieranym przez tą metodę jest taki model, w którym dodanie następnego parametru powodowałoby brak istotnego statystycznie wzrostu dewiancji. W kolejnych wierszach tabeli 14 przedstawione zostały wartości dewiancji oraz istotności statystycznej zmiennych wprowadzanych do modelu w kolejnych krokach metody selekcji postępującej. W pierwszym wierszu przedstawiona jest dewiancja dla modelu zerowego (zawierającego tylko wyraz wolny).

Tab. 14. Zastosowanie metody selekcji postępującej

Krok	dodawany do modelu parametr	dewiancja modelu	poziom istotności różnicy pomiędzy modelami
-	tylko wyraz wolny	92.35	-
1	**WIEKPL**	**74.47**	**<0.001**
2	**SPO2DEV**	**62.58**	**<0.001**
3	**PDA**	**55.60**	**0.008**
4	**BPMMEAN**	**47.07**	**0.003**
5	**SURFACT**	**40.19**	**0.008**
6	**SPO2MEAN_TR**	**36.22**	**0.047**
7	**RESPIMV**	**30.18**	**0.014**
8	**SPO2DEV_TR**	**24.38**	**0.016**
9	HIGH94	24.10	0.597
10	SPO2MEAN	23.13	0.325
11	LOW85	22.42	0.399
12	BPMMEAN_TR	21.78	0.426
13	MASAUR	20.88	0.343
14	AA	20.86	0.867

Wprowadzenie wszystkich możliwych parametrów do modelu skutkuje uzyskaniem maksymalnej dewiancji równej 20.86 (model pełny - rozdz. 6.1.1), jednak jak wynika z tabeli ostatnie 6 parametrów nie wprowadza istotnej różnicy. Znaczącymi statystycznie zmiennymi (przy założonym poziomie istotności równym 5%) są :

- WIEKPL,
- SPO2DEV,
- PDA,
- BPMMEAN,
- SURFACT,
- SPO2MEAN_TR,
- SPO2DEV_TR,

– RESPIMV.

Wartości podstawowych parametrów określających jakość predykcji przedstawione są w tabeli 15.

Tab. 15. Wartości podstawowych parametrów określających jakość predykcji dla modelu uzyskanego przy pomocy metody selekcji postępującej

Parametr	Wartość
AUC (pole powierzchni pod krzywą ROC)	$0.985 \pm 0.01^{*}$
DEV (dewiancja)	24.38
ACC (trafność klasyfikowania) [%]	95.7
Kryterium AIC	45.38
Kryterium BIC	62.61

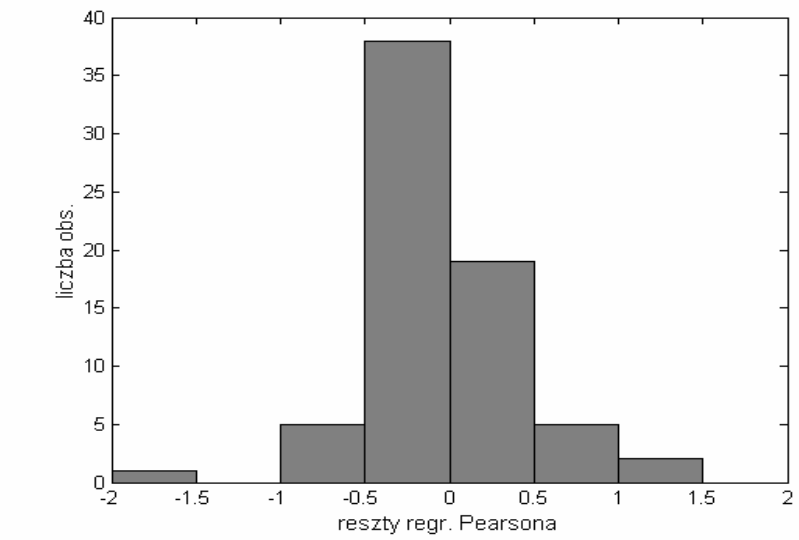
* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

Macierz pomyłek dla analizowanego modelu przedstawiona jest w tabeli 16.

Tab. 16. Macierz pomyłek dla modelu uzyskanego przy pomocy metody selekcji postępującej

	Przewidywany brak BPD	Przewidywana BPD		
Obserwowany brak BPD	43	1	specyficzność	97.7 %
Obserwowane BPD	2	24	czułość	92.3 %

Histogram reszt regresyjnych Pearsona dla uzyskanego modelu przedstawiony jest na rys. 30.

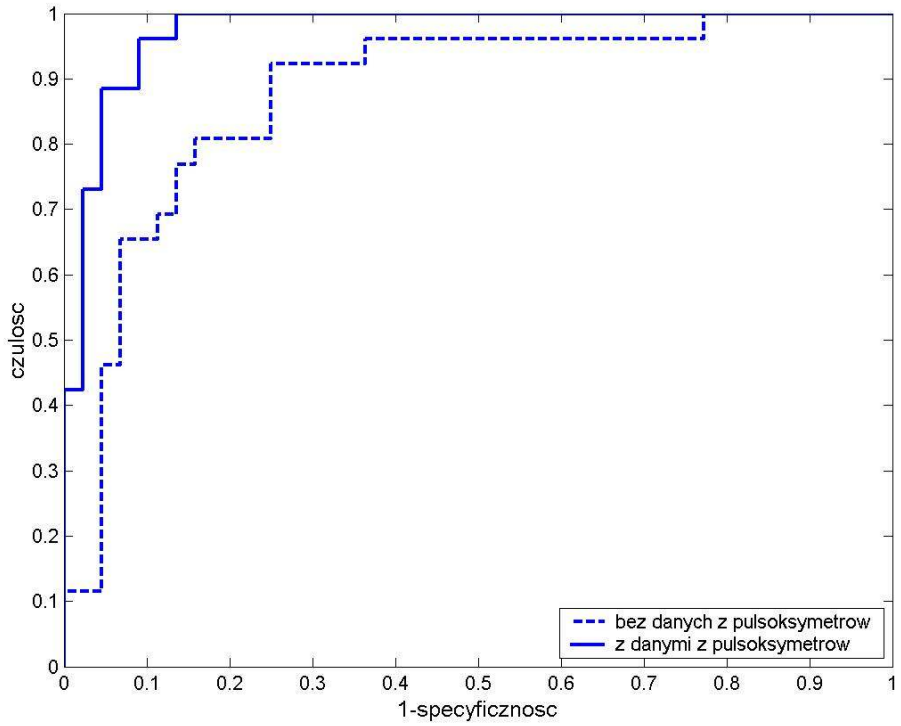


Rys. 30. Histogram reszt regresyjnych Pearsona dla modelu uzyskanego przy pomocy metody selekcji postępującej

Niskie wartości współczynników VIF przedstawione w tabeli 17 oznaczają wiarygodność otrzymanych współczynników równania regresji logistycznej dla uzyskanego modelu.

Tab. 17. Wartości współczynników VIF dla poszczególnych zmiennych wchodzących w skład modelu wybranego przy użyciu metody selekcji postępującej

zmienna	VIF
BPMMEAN	1.10
SPO2DEV_TR	1.22
PDA	1.25
WIEKPL	1.25
SPO2DEV	1.34
SPO2MEAN_TR	1.34
SURFACT	1.47
RESPIMV	1.50



Rys. 31. Porównanie krzywych ROC dla modeli regresji logistycznej z użyciem zbioru zmiennych niezależnych zawierających dane z systemu rejestracji danych medycznych i z użyciem zbioru nie zawierającego tych danych

W uzyskanym zestawie zmiennych objaśniających znalazły się cztery zmienne uzyskane z systemu gromadzenia danych medycznych. Są to :

- SPO2DEV,
- BPMMEAN,
- SPO2MEAN_TR,
- SPO2DEV_TR.

Korzystając z testu χ^2 można oszacować ich istotność statystyczną w uzyskanym modelu. Usunięcie tych zmiennych powoduje wzrost dewiancji do wartości 57.05, co daje różnicę 27.18. Przy 4 stopniach swobody (usunięcie czterech parametrów) poziom istotności wyznaczony przy pomocy statystyki χ^2 jest mniejszy od 0.000002, co oznacza że wpływ tych parametrów jest silnie istotny statystycznie w rozpatrywanym modelu. Krzywe ROC dla obydwu modeli (z danymi z systemu gromadzenia danych medycznych oraz bez tych danych) przedstawione są na rys. 31.

6.1.3. Metoda eliminacji wstecznej

Zgodnie z algorytmem podanym w rozdziale 5.4, rozpoczyna się od pełnego modelu zawierającego wszystkie zmienne niezależne i usuwa po kolei te zmienne, których odrzucenie nie powoduje znaczącego statystycznie spadku dewiancji. Podobnie jak w przypadku metody selekcji postępującej, istotność różnicy pomiędzy modelami sprawdza się poprzez obliczenie różnicy pomiędzy dewiancjami dla obydwu modeli i porównaniu ich ze statystyką χ^2 o jednym stopniu swobody (rozdział 5.1). Procedurę powtarza się do momentu uzyskania optymalnego modelu, w którym wszystkie parametry są istotne statystycznie. W tabeli 18 przedstawione są wartości dewiancji i poziomu istotności dla kolejnych modeli uzyskiwanych w wyniku opisywanej powyżej procedury.

Jak wynika z tabeli, optymalny model uzyskany w wyniku przeprowadzenia tej metody różni się nieco od modelu uzyskanego przy pomocy metody selekcji postępującej. Optymalnym modelem w tym przypadku jest model zawierający następujące zmienne objaśniające :

- WIEKPL,
- SPO2DEV,
- PDA,
- BPMMEAN,
- SPO2MEAN_TR,
- SPO2DEV_TR,
- RESPIMV.

Tab. 18. Zastosowanie metody eliminacji wstecznej

Krok	usuwany z modelu parametr	dewiancja modelu	poziom istotności różnicy pomiędzy modelami
-	pełny model	20.86	-
1	AA	20.89	0.87
2	MASAUR	21.79	0.34
3	BPMMEAN_TR	22.42	0.42
4	LOW85	23.13	0.40
5	SPO2MEAN	24.10	0.33
6	HIGH94	24.38	0.60
7	SURFACTANT	27.31	0.087
8	**SPO2DEV_TR**	**33.38**	**0.014**
9	**SPO2MEAN_TR**	**40.80**	**0.006**
10	**PDA**	**46.44**	**0.018**
11	**SPO2DEV**	**55.52**	**0.003**
12	**RESPIMV**	**65.77**	**0.001**
13	**BPMMEAN**	**74.47**	**0.003**
14	**WIEKPL**	**92.35**	**<0.001**

Jak łatwo zauważyć, odrzucenie zmiennej oznaczającej podanie surfaktantu w pierwszym tygodniu życia w tym modelu spowodowane było przekroczeniem założonego 5% poziomu istotności dla tej zmiennej. Zakładając większy poziom istotności (np. 10%), zmienna ta znalazłaby się w modelu i w rezultacie modele uzyskane obydwoma metodami byłyby takie same. Wartości podstawowych parametrów określających jakość predykcji otrzymanego modelu przedstawione zostały w tabeli 19. Macierz pomyłek dla analizowanego modelu przedstawiona jest w tabeli 20.

Tab. 19. Wartości podstawowych parametrów określających jakość predykcji dla modelu uzyskanego przy pomocy metody eliminacji wstecznej

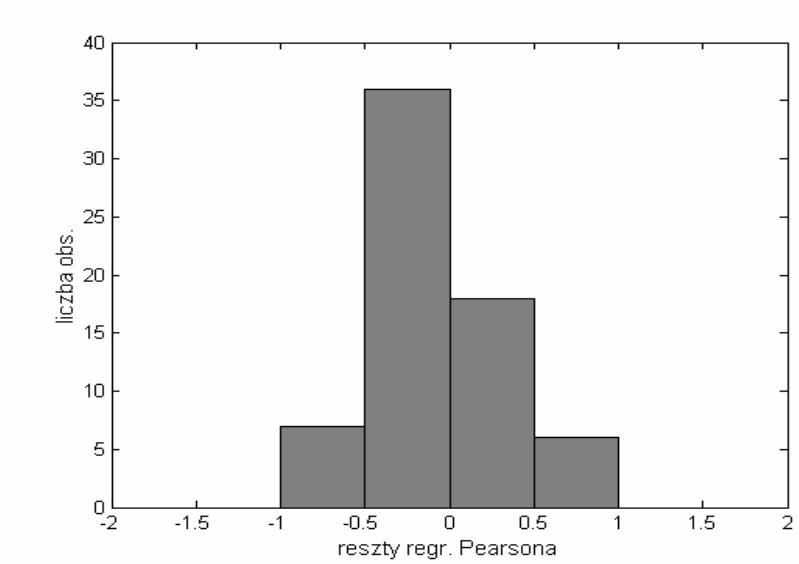
Parametr	Wartość
AUC (pole powierzchni pod krzywą ROC)	$0.976 \pm 0.02^{*}$
DEV (dewiancja)	27.31
ACC (trafność klasyfikowania) [%]	95.7
Kryterium AIC	45.67
Kryterium BIC	61.30

* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

Tab. 20. Macierz pomyłek dla modelu uzyskanego przy pomocy metody eliminacji wstecznej

	Przewidywany brak BPD	Przewidywane BPD		
Obserwowany brak BPD	43	1	specyficzność	97.7 %
Obserwowane BPD	2	24	czułość	92.3 %

Histogram reszt regresyjnych Pearsona dla uzyskanego modelu przedstawiony jest na rys. 32.



Rys. 32. Histogram reszt regresyjnych Pearsona dla modelu uzyskanego przy pomocy metody eliminacji wstecznej

Wszystkie uzyskane zmienne niezależne charakteryzują się niskimi wartościami współczynników VIF oznaczającymi wiarygodność otrzymanych współczynników równania regresji logistycznej (tab. 21).

Tab. 21. Wartości współczynników VIF dla poszczególnych zmiennych wchodzących w skład modelu wybranego przy użyciu metody eliminacji wstecznej

zmienna	VIF
BPMMEAN	1.10
PDA	1.16
SPO2DEV_TR	1.21
WIEKPL	1.23
SPO2DEV	1.25
SPO2MEAN_TR	1.32
RESPIMV	1.35

Podobnie jak dla modelu wybranego w metodzie selekcji postępującej, również w tym przypadku sprawdzić można istotność statystyczną parametrów uzyskanych z systemu gromadzenia danych medycznych. Wyznaczona podobnie jak poprzednio różnica dewiancji obydwu modeli równa jest 32.92. Przy 4 stopniach swobody (usunięcie czterech parametrów)

poziom istotności różnicy wyznaczony przy pomocy statystyki χ^2 jest mniejszy od 0.000002 co oznacza, że wpływ tych parametrów jest silnie istotny statystycznie w modelu.

6.1.4. Przeszukanie wszystkich możliwych modeli

Analiza wszystkich możliwych modeli została wykonana w celu oceny jakości wyboru dokonanego przez wymienione wcześniej metody eliminacji wstecznej i selekcji postępującej. W przedstawionym przykładzie stwarza to konieczność przeanalizowania 16384 kombinacji podzbiorów zmiennych niezależnych. Moc obliczeniowa zastosowanych serwerów wystarcza, żeby przeprowadzić taką operację w czasie kilku minut, ponadto prostym sposobem można obliczenia wykonywać równolegle, dzieląc je pomiędzy kolejne serwery obliczeniowe.

Wyszukiwanie podzbiorów zmiennych niezależnych według wartości poziomów istotności tych zmiennych w sposób opisany poprzednio skutkowałoby uzyskaniem również podzbiorów zawierających minimalną ilość zmiennych (np. jedną zmienną silnie istotną statystycznie). Takie modele charakteryzują się niską zdolnością predykcyjną (mała wartość pola powierzchni pod krzywą ROC, duża dewiancja, niska trafność klasyfikacji). W związku z tym odpowiednimi kryteriami służącymi do wyboru podzbiorów są opisane w rozdziale 5.4 kryteria AIC i BIC, faworyzujące modele zawierające jak największą ilość zmiennych niezależnych równocześnie najbardziej istotnych statystycznie. Dodatkowo dla porównania przedstawiono wyniki poszukiwania, zakładając jako kryterium maksimum pola powierzchni pod krzywą ROC oraz minimum dewiancji. Wartości poszczególnych wskaźników dla tych wybranych modeli przedstawione są w tabeli poniżej.

- Model o największym polu powierzchni pod krzywą ROC

Modelem o największym polu powierzchni pod krzywą ROC jest model zawierający następujący podzbiór zmiennych : SPO2DEV_TR, SPO2MEAN_TR, BPMMEAN_TR, SPO2DEV, SPO2MEAN, BPMMEAN, HIGH94, MASAUR, PDA, RESPIMV, SURFACT, WIEKPL. Wartości poziomów istotności poszczególnych parametrów w tym modelu oraz wartości podstawowych współczynników określających zdolność predykcyjną przedstawiają tabele 22 i 23.

Tab. 22. Wartości poziomów istotności parametrów modelu o maksymalnej wartości pola powierzchni pod krzywą ROC

parametr	poziom istotności parametru
wyraz wolny	0.420473
SPO2DEV_TR	**0.031587**
SPO2MEAN_TR	**0.028637**
BPMMEAN_TR	0.342446
SPO2DEV	**0.029734**
SPO2MEAN	0.205267
BPMMEAN	**0.011222**
HIGH94	0.185920
MASAUR	0.390886
PDA	**0.031655**
RESPIMV	**0.032126**
SURFACT	0.183218
WIEKPL	**0.045948**

Tab. 23. Wartości podstawowych parametrów określających jakość predykcji dla modelu charakteryzującego się maksymalną wartością pola powierzchni pod krzywą ROC

Parametr	Wartość
AUC (pole powierzchni pod krzywą ROC)	0.99 ± 0.01*
DEV (dewiancja)	21.83
ACC (trafność klasyfikowania) [%]	98.6
Kryterium AIC	54.32
Kryterium BIC	77.06

* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

Jak wynika z tych tabel, w modelu, dla którego pole powierzchni pod krzywą ROC jest największe, nie wszystkie zmienne są istotne statystycznie.

- Model o najmniejszej dewiancji

Modelem o minimalnej wartości dewiancji jest model zawierający wszystkie zmienne. Jego wartości parametrów predykcyjnych przedstawione zostały w rozdziale 6.1.1. Jak wynika z tabeli 11, znaczna część zmiennych nie jest istotna statystycznie.

- Model o minimalnej wartości kryterium AIC

Modelem o najmniejszej wartości kryterium AIC jest model zawierający następujące zmienne: SPO2DEV_TR, SPO2MEAN_TR, SPO2DEV, BPMMEAN, PDA, RESPIMV,

SURFACT, WIEKPL. Jak widać jest to ten sam podzbiór zmiennych, który został wybrany przy użyciu metody selekcji postępującej (rozdział 6.1.2). W modelu tym wszystkie zmienne niezależne są istotne statystycznie.

- Model charakteryzujący się minimalną wartością kryterium BIC

Modelem o najmniejszej wartości kryterium BIC jest model zawierający następujące zmienne: SPO2DEV_TR, SPO2MEAN_TR, SPO2DEV, BPMMEAN, PDA, RESPIMV, WIEKPL. Jest to ten sam podzbiór zmiennych, który został uzyskany w wyniku zastosowania algorytmu eliminacji wstecznej (rozdział 6.1.3).

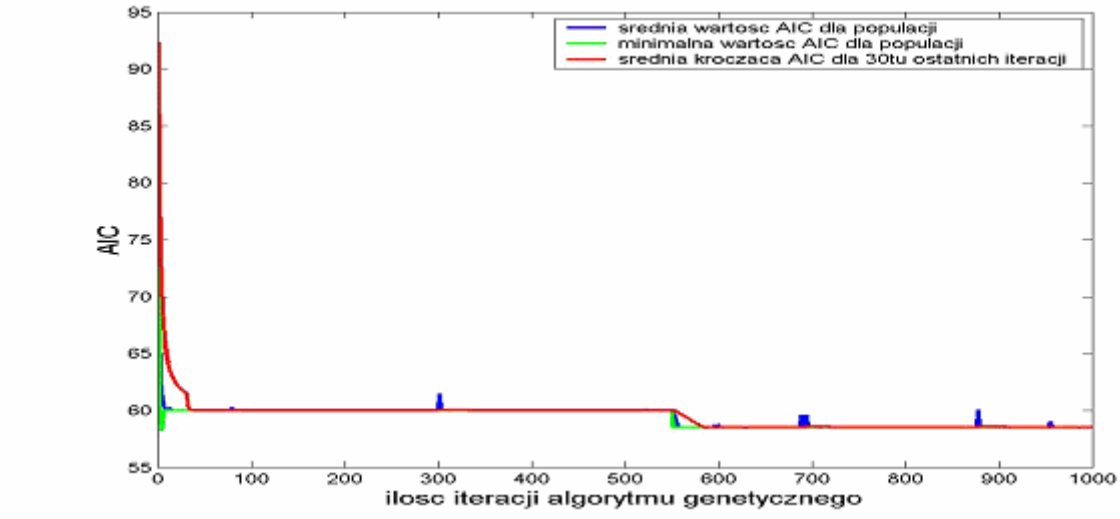
W wyniku analizy modeli dla wszystkich możliwych kombinacji podzbiorów zmiennych niezależnych zostały znalezione modele charakteryzujące się maksymalną wartością pola powierzchni pod krzywą ROC, minimalną wartością dewiancji oraz minimalnymi wartościami kryteriów AIC i BIC. Podzbiory uzyskane przy pomocy pierwszych dwóch kryteriów zawierają zmienne nieistotne statystycznie. Kryteria te nie są więc dobrymi kryteriami do wyboru optymalnego podzbioru zmiennych niezależnych. Lepsze wyniki uzyskuje się korzystając z kryteriów AIC i BIC. Dla modeli charakteryzujących się minimalnymi wartościami tych kryteriów wszystkie zmienne niezależne są istotne statystycznie, ponadto okazało się, że są to modele uzyskane w wyniku zastosowania algorytmów selekcji postępującej i eliminacji wstecznej. Ogólnie jednak nie musi to być regułą. Modele otrzymane przy użyciu kryterium BIC charakteryzują się ogólnie mniejszą ilością parametrów, większą dewiancją i mniejszym polem powierzchni pod krzywą ROC niż modele uzyskane przy użyciu kryterium AIC.

6.1.5. Zastosowanie algorytmu genetycznego do poszukiwania optymalnego podzbioru

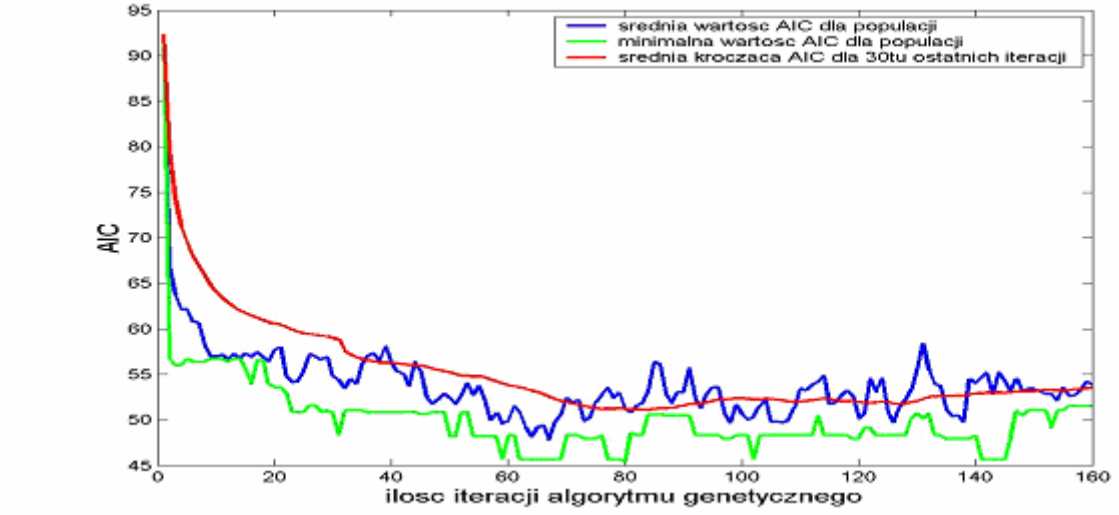
W rozwiązaniu problemu poszukiwania optymalnego podzbioru zmiennych niezależnych przydatnym narzędziem jest algorytm genetyczny opisany w rozdziale 5.4. Przeprowadzone symulacje pokazują, że algorytm ten przy prawidłowo dobranych parametrach umożliwia znalezienie optymalnego podzbioru w kilkudziesięciu iteracjach. Wielkość populacji ustalono na 10, zaś prawdopodobieństwo krzyżowania p_k=0.75. Duże znaczenie ma odpowiedni dobór prawdopodobieństwa mutacji. Jak pokazują wykresy na rysunkach 33-35, zbyt duża wartość przyjętego prawdopodobieństwa mutacji powoduje nadmierne zróżnicowanie populacji w kolejnych iteracjach, co może prowadzić do braku zbieżności algorytmu i wymusza

przebadanie zbyt dużej ilości modeli. Z kolei zbyt mała wartość tego prawdopodobieństwa prowadzi do zbyt małego zróżnicowania populacji i w rezultacie może skutkować zatrzymaniem algorytmu w lokalnym minimum. Jako optymalną wartość przyjęto p_m=0.05. Dla tak przyjętych wartości powyższych parametrów optymalny wynik uzyskiwany był po 100-150 iteracjach algorytmu, co przekładało się na konieczność przebadania ok. 60-70 modeli regresji logistycznej dla różnych zmiennych niezależnych. Znalezionym modelem był model o minimalnej wartości AIC=45.38 zawierający zmienne niezależne : SPO2DEV_TR, SPO2MEAN_TR, SPO2DEV, BPMMEAN, PDA, RESPIMV, SURFACT, WIEKPL.

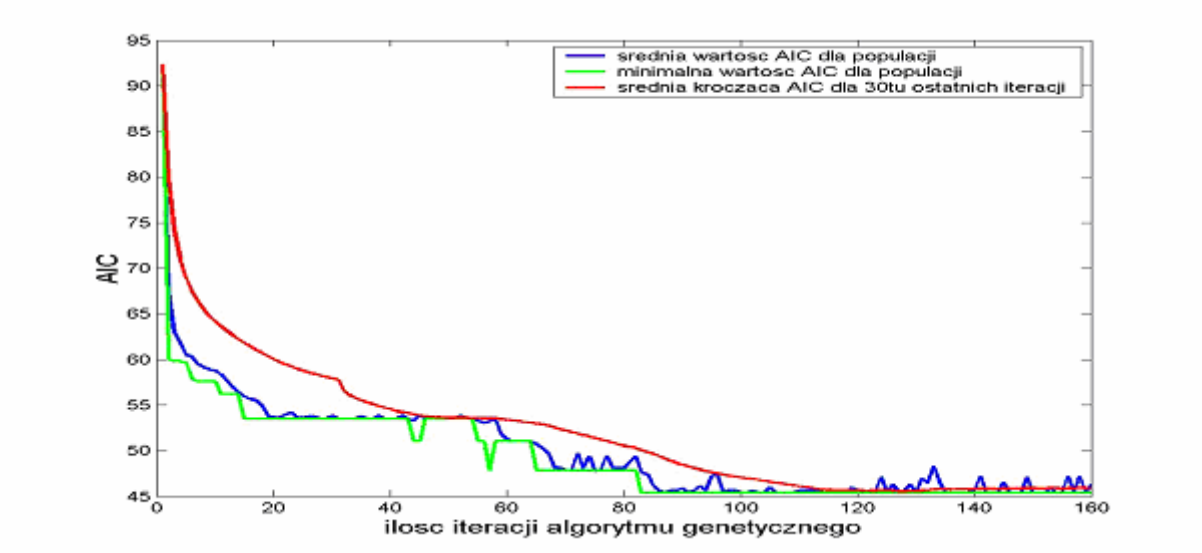
Zysk czasowy tej metody w stosunku do podejścia polegającego na przeszukaniu wszystkich możliwych kombinacji jest znaczny. W dodatku B znajduje się listing programu wykorzystanego w symulacjach.



Rys. 33. Zależność AIC od ilości iteracji algorytmu genetycznego dla p_m=0.001



Rys. 34. Zależność AIC od ilości iteracji algorytmu genetycznego dla p_m=0.2



Rys. 35. Zależność AIC od ilości iteracji algorytmu genetycznego dla p_m=0.05

6.1.6. Oszacowanie zdolności predykcyjnej uzyskanych modeli przy użyciu walidacji krzyżowej

Do oszacowania zdolności predykcyjnej metody regresji logistycznej dla nowych danych zastosowana została walidacja krzyżowa 14-krotna (opisana w rozdziale 5.5). Dla optymalnych modeli wybranych w poprzednim kroku wyznaczone zostały ponownie podstawowe wskaźniki jakości predykcji, tym razem jednak wyznaczone po podziale zbioru danych na podzbiory testowe i treningowe.

Wyniki otrzymane dla modelu uzyskanego przy pomocy metody selekcji postępującej i zawierającego następujące parametry :

- SPO2DEV_TR,
- SPO2MEAN_TR,
- SPO2DEV,
- BPMMEAN,
- PDA,
- RESPIMV,
- SURFACT,
- WIEKPL

przedstawione są w tabelach 24 i 25.

Tab. 24. Wartość AUC i trafności predykcji dla modelu uzyskanego przy użyciu metody selekcji postępującej i zastosowaniu walidacji krzyżowej 14-krotnej

Parametr	Wartość
AUC (pole powierzchni pod krzywą ROC)	$0.91 \pm 0.03^{*}$
ACC (trafność klasyfikowania) [%]	88.6

* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

Tab. 25. Macierz pomyłek dla modelu uzyskanego przy użyciu metody selekcji postępującej i zastosowaniu walidacji krzyżowej 14-krotnej

	Przewidywany brak BPD	Przewidywane BPD		
Obserwowany brak BPD	41	3	specyficzność	93.2 %
Obserwowane BPD	5	21	czułość	80.8 %

Wyniki otrzymane dla modelu uzyskanego przy pomocy metody eliminacji wstecznej i zawierającego następujące parametry :

- WIEKPL,
- SPO2DEV,
- PDA,
- BPMMEAN,
- SPO2MEAN_TR,
- SPO2DEV_TR,
- RESPIMV.

przedstawione są w tabelach 26 i 27.

Tab. 26. Wartość AUC i trafności predykcji dla modelu uzyskanego przy użyciu metody eliminacji wstecznej i zastosowaniu walidacji krzyżowej 14-krotnej

Parametr	Wartość
AUC (pole powierzchni pod krzywą ROC)	$0.90 \pm 0.03^{*}$
ACC (trafność klasyfikowania) [%]	87.1

* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

Tab. 27. Macierz pomyłek i trafność klasyfikacji dla modelu uzyskanego przy użyciu metody eliminacji wstecznej i zastosowaniu walidacji krzyżowej 14-krotnej

	Przewidywany brak BPD	Przewidywane BPD		
Obserwowany brak BPD	40	4	specyficzność	90.9 %
Obserwowane BPD	5	21	czułość	80.8 %

W celu porównania przeprowadzono za pomocą tej samej metody walidację pozostałych modeli z rozdziału 6.1.4. Dla modelu charakteryzującego się maksymalną wartością pola powierzchni pod krzywą ROC (AUC=0.99), obliczaną bez podziału na zbiory uczące i testowe, podzbiór zmiennych ograniczony był do następujących parametrów :

- SPO2DEV_TR,
- SPO2MEAN_TR,
- BPMMEAN_TR,
- SPO2DEV,
- SPO2MEAN,
- BPMMEAN,
- HIGH94 ,
- MASAUR,
- PDA,
- RESPIMV,
- SURFACT,
- WIEKPL.

Po zastosowaniu walidacji krzyżowej wartość pola powierzchni pod krzywą ROC wynosi 0.84±0.03, zaś trafność prognozy ok. 73%. Wynika stąd, że model ten ma w rzeczywistości znacznie słabszą zdolność predykcyjną dla nowych przypadków, niż modele charakteryzujące się minimalnymi wartościami kryteriów AIC i BIC. Podobnie przedstawia się sytuacja dla pełnego modelu, zbudowanego z wykorzystaniem wszystkich zmiennych niezależnych : po zastosowaniu walidacji krzyżowej 14-krotnej pole powierzchni pod krzywą ROC równe jest 0.82±0.03, trafność prognozy ok. 70%.

W tabeli 28 przedstawiono wartości współczynników równania logistycznego i ich błędów standardowych (oznaczenia wg wzoru 5.3) dla optymalnego modelu wybranego przy użyciu metody selekcji postępującej, obliczone bez użycia walidacji krzyżowej.

W wyniku zastosowania k-krotnej walidacji krzyżowej uzyskuje się k modeli regresyjnych. Każdy z tych modeli dopasowany został dla innego zbioru przypadków, tak więc współczynniki α oraz β_i równania logistycznego są różne dla każdego z tych k zbiorów.

Tab. 28. Wartości parametrów równania logistycznego dla modelu wybranego przy pomocy selekcji postępującej i bez użycia walidacji krzyżowej

Parametr	β_i	std(β_i)
wyraz wolny α	119	58
SPO2DEV_TR	-5.5	2.7
SPO2MEAN_TR	-142	60
SPO2DEV	3.5	1.4
BPMMEAN	0.35	0.12
PDA	3.5	1.5
RESPIMV	6.1	3.0
SURFACT	2.3	1.6
WIEKPL	-1.5	0.5

Ponadto można zauważyć, że istnieje :

$$C = \frac{n!}{(n-m)!\,m!} \tag{6.2}$$

sposobów podziału n-elementowego zbioru wszystkich przypadków na m-elementowe zbiory testowe (i n-m elementowe zbiory treningowe). We wzorze tym :

$$m = \frac{n}{k} \tag{6.3}$$

dla k-krotnej walidacji krzyżowej o tak dobranej wartości k, aby była dzielnikiem liczby przypadków n. Dla przyjętych w pracy wartości n=70, k=14 oraz m=5 ilość sposobów podziału przekracza 12 milionów.

Tab. 29. Wartości średnie i odchylenia standardowe parametrów równań logistycznych dla modeli tworzonych w wyniku walidacji krzyżowej

Parametr	$\bar{\beta}$	std(β)
wyraz wolny α	125	60
SPO2DEV_TR	-6.2	2.7
SPO2MEAN_TR	-153	61
SPO2DEV	3.9	1.5
BPMMEAN	0.39	0.12
PDA	3.9	1.5
RESPIMV	8.4	3
SURFACT	2.6	1.7
WIEKPL	-1.6	0.5

W tabeli 29 przedstawiono uśrednione arytmetycznie wartości współczynników równań logistycznych $\bar{\beta}$ oraz ich odchylenia standardowe std(β). Współczynniki te otrzymane zostały w wyniku 100-krotnego powtórzenia walidacji krzyżowej. Za każdym razem losowano rozkład przypadków w zbiorze dzielonym następnie na przypadki testowe i

treningowe. Porównując wartości w tabelach 28 i 29 można zauważyć, że w granicach błędu obydwa modele mają zbliżone wartości parametrów równania logistycznego.

6.1.7. Wnioski dotyczące predykcji dysplazji przy pomocy regresji logistycznej

Wykorzystanie do predykcji dysplazji oskrzelowo-płucnej wszystkich dostępnych zmiennych objaśniających skutkuje co prawda uzyskaniem modelu o bardzo niskiej dewiancji i dużej wartości pola powierzchni pod krzywą ROC, ale tylko gdy dopasowanie i testowanie odbywa się na tym samym zbiorze danych. Duża część zmiennych niezależnych jest w takim modelu nieistotna statystycznie, uwydatnia się również opisany wcześniej wpływ współliniowości tych zmiennych. Zastosowanie walidacji krzyżowej w celu oceny zdolności predykcyjnej dla nowych przypadków, nie uwzględnionych w zbiorze wykorzystanym do budowy modelu, pokazuje wyraźnie znaczne pogorszenie współczynników określających zdolność predykcyjną (pole powierzchni pod krzywą ROC, dewiancja, macierz pomyłek).

W celu wyznaczenia optymalnego podzbioru zmiennych niezależnych, czyli takiego, dla którego zdolność predykcyjna modelu jest maksymalna, posłużono się kilkoma metodami. W wyniku zastosowania najprostszych z nich, czyli metody eliminacji wstecznej i selekcji postępującej, uzyskano dwa różne podzbiory, w obydwu wszystkie zmienne były istotne statystycznie. Z uwagi na względnie niewielką ilość wszystkich możliwych do skonstruowania kombinacji zmiennych objaśniających, możliwym staje się przebadanie wszystkich podzbiorów. Jako kryteria oceny modeli wybrano AIC i BIC, ponieważ minimalizują one ilość zmiennych wchodzących w skład podzbioru, przy jednoczesnym zapewnieniu maksymalnego dopasowania uzyskanego modelu. W wyniku działania tej procedury okazało się, że przy pomocy metod eliminacji wstecznej i selekcji postępującej w rezultacie uzyskano jedne z najlepszych modeli. Niestety nie zawsze jest to regułą, często bywa tak, że zwracane przez te metody podzbiory są dalekie od optimum. Metod tych można używać w celu przybliżonego określenia podzbioru zmiennych predykcyjnych. Bardziej wiarygodne wyniki uzyskuje się przy zastosowaniu bardziej skomplikowanych metod heurystycznych, takich jak np. algorytm genetyczny. Odpowiednio dobrane współczynniki takiego algorytmu pozwalają znaleźć globalne optimum przy stosunkowo niewielkiej ilości koniecznych do wykonania obliczeń.

Model zbudowany przy użyciu optymalnego podzbioru zmiennych niezależnych (uzyskanego w wyniku zastosowania kryteriów AIC i BIC) charakteryzuje się bardzo dobrą zdolnością predykcyjną dla nowych przypadków. Potwierdzają to wyniki walidacji krzyżowej

przeprowadzone w rozdziale 6.1.6. Pole powierzchni pod krzywą ROC w takim przypadku osiąga wartość maksymalną, podobnie jak trafność klasyfikacji przypadków.

Obydwa przypadki podzbiorów znalezionych przy pomocy prostych metod heurystycznych charakteryzujące się maksymalną zdolnością predykcyjną (określoną wartością pola powierzchni pod krzywą ROC) zawierały zmienne pochodzące ze strumienia danych zgromadzonych z pulsoksymetrów. Jak wykazano przy pomocy testu ilorazu wiarygodności, wpływ tych zmiennych jest silnie istotny statystycznie. Usunięcie tych zmiennych powoduje znaczne pogorszenie zdolności predykcyjnych modelu.

6.2. Predykcja dysplazji przy użyciu sztucznych sieci neuronowych

Opisany w rozdziale 5.1 model regresji logistycznej odpowiada funkcji realizowanej przez neuron o sigmoidalnej funkcji aktywacji. Do predykcji dysplazji można również zastosować sieci neuronowe składające się z większej ilości neuronów. Poniżej przedstawione zostały wyniki zastosowania do tego celu jednokierunkowych wielowarstwowych sieci neuronowych oraz sieci radialnych. W celu porównania wyników otrzymywanych przy pomocy opisanej wcześniej regresji logistycznej, wykorzystany zostanie ten sam co wyznaczony poprzednio optymalny zbiór parametrów :

- SPO2DEV_TR,
- SPO2MEAN_TR,
- SPO2DEV,
- BPMMEAN,
- PDA,
- RESPIMV,
- SURFACT,
- WIEKPL.

Model ten charakteryzował się maksymalną wartością pola powierzchni pod krzywą ROC dla modelu regresji logistycznej przy zastosowaniu walidacji krzyżowej użytej do określenia zdolności uogólniających modelu. Przed wprowadzeniem na wejście sieci neuronowej dane zostały poddane normalizacji, w ten sposób, żeby ich wartość średnia była równa 0, natomiast odchylenie standardowe równe 1 [9]. Wektor wyjściowy stanowiła zmienna określająca wystąpienie dysplazji BPD (0 – nie, 1 – tak).

6.2.1. Predykcja przy użyciu jednokierunkowych sieci wielowarstwowych

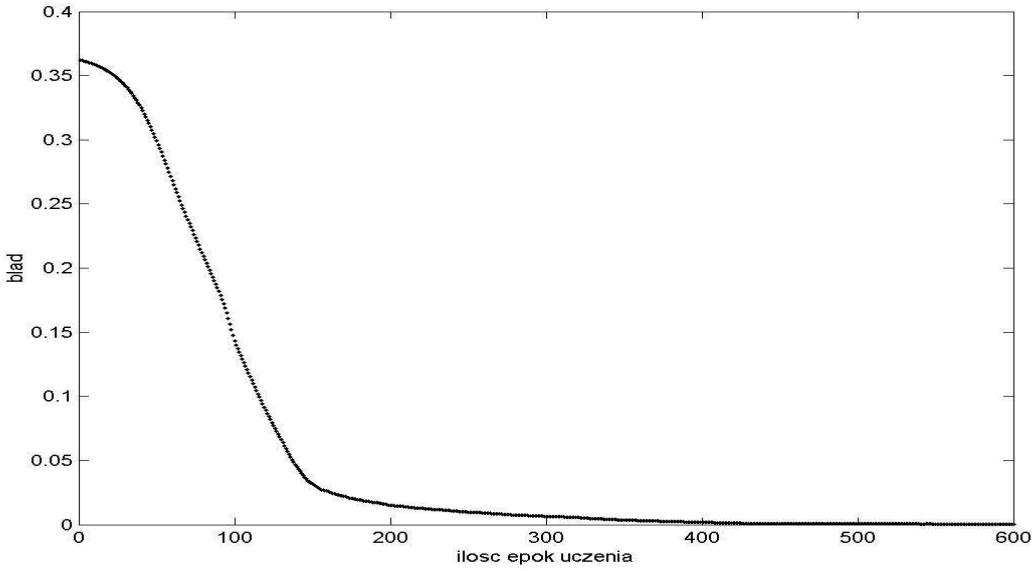
Wykorzystywane w pracy jednokierunkowe sieci wielowarstwowe zbudowane zostały z dwóch oraz trzech warstw neuronów o sigmoidalnej funkcji aktywacji. Liczna neuronów w warstwach ukrytych dobierana była eksperymentalnie. Liczba neuronów w warstwie wyjściowej jest zdeterminowana przez liczbę sygnałów generowanych przez sieć. Sieć wykorzystana została do predykcji jednego parametru, w związku z tym warstwa wyjściowa sieci neuronowej zawiera pojedynczy neuron. Jego funkcja aktywacji ma charakter sigmoidalny ze względu na dychotomiczny charakter przewidywanej zmiennej (dysplazji BPD). Liczba neuronów w poszczególnych warstwach i związana z tym ilość parametrów sieci (wag i biasów) decyduje o zdolnościach generalizacyjnych sieci neuronowej. Jeśli ilość ta jest wystarczająco duża, sieć jest w stanie odwzorować dokładnie każdą zależność pomiędzy danymi wejściowymi a wyjściowymi. Niestety sytuacja ta wiąże się ze słabą zdolnością nauczonej sieci do uogólniania. Objawia się to tym, że niskiemu poziomowi błędu na zbiorze uczącym odpowiada wysoki poziom błędu dla nowych, nie prezentowanych w procesie uczenia sieci danych (czyli dla zbioru testowego).

Na rys. 36 przedstawiony został wykres zależności błędu średniokwadratowego obliczanego dla zbioru uczącego od liczby epok uczenia sieci neuronowej. Błąd średniokwadratowy określony jest wzorem :

$$e = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2 \tag{6.4}$$

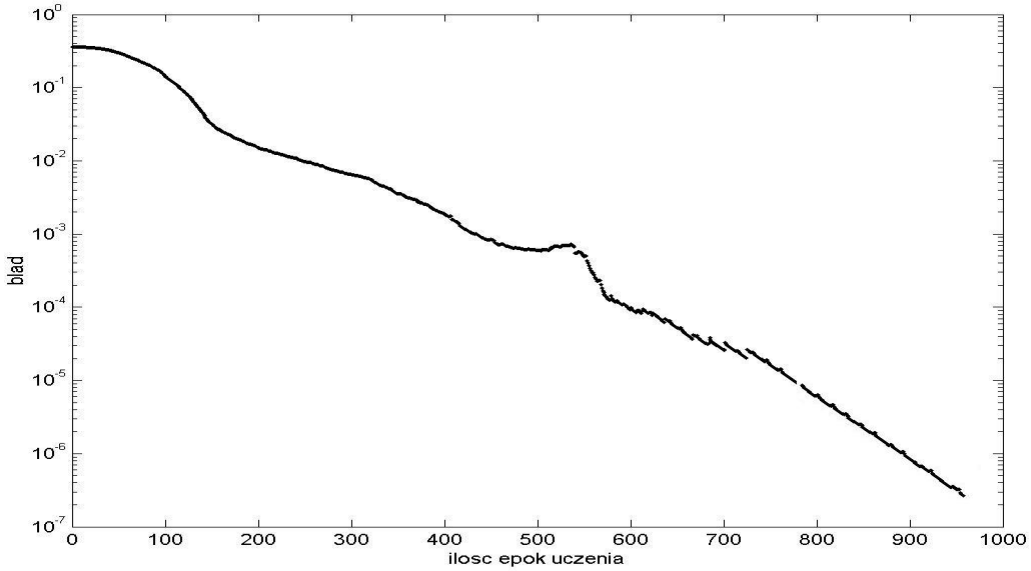
gdzie N - ilość wzorców uczących, y_i - odpowiedź sieci dla i-go wzorca uczącego, x_i - rzeczywista wartość wyjściowa dla i-go wzorca uczącego.

Konfiguracja sieci neuronowej oznaczona została jako [NL_1-NL_2-NL_3], gdzie NL_i, dla i=1,2,3 oznacza liczbę neuronów w i-tej warstwie. Sieć użyta tutaj miała strukturę [8-8-1], czyli składała się z 8-miu neuronów w każdej warstwie ukrytej i pojedynczego neuronu wyjściowego. W zbiorze treningowym znalazły się dane wszystkich dostępnych pacjentów. Wykres ten uzyskany został poprzez uśrednienie wyników otrzymanych z kilkudziesięciu niezależnych procesów uczenia tej samej sieci neuronowej. Ze względu na różnice w początkowych wartościach wag i biasów (zależnych od zastosowanej procedury inicjalizacyjnej) wartości błędu mają duże odchylenie we wstępnej fazie uczenia sieci (przy małej liczbie epok).



Rys. 36. Wykres zależności błędu średniokwadratowego wyznaczonego dla zbioru treningowego od liczby epok uczenia

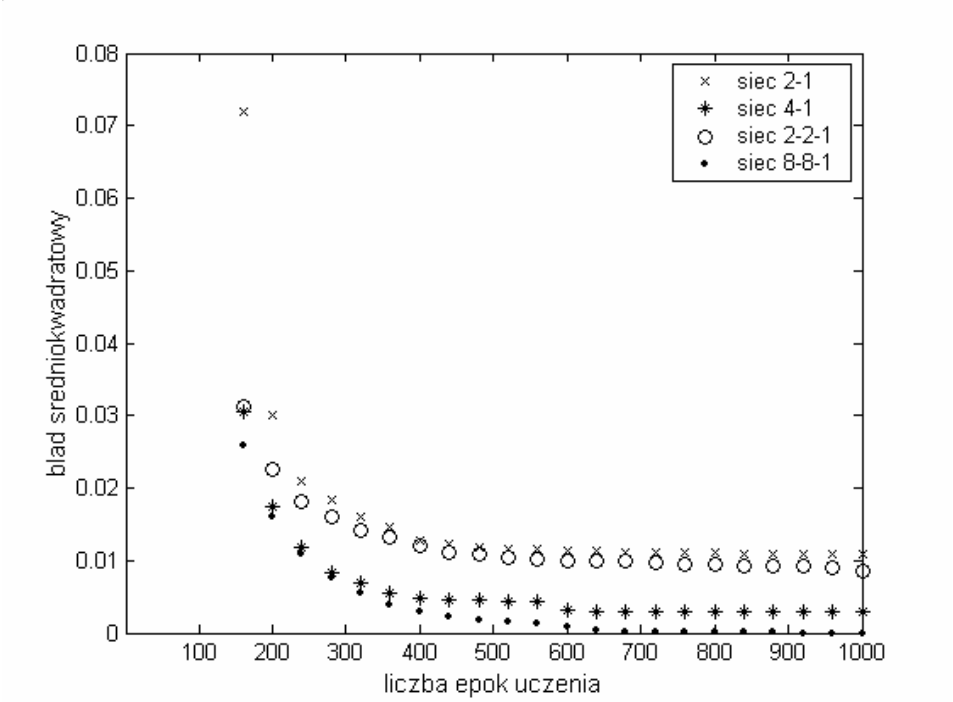
Sieć o takiej strukturze jest w stanie zminimalizować błąd dla danych w zbiorze treningowym do bardzo małych wartości, co lepiej pokazuje następny wykres (rys. 37), gdzie wartość błędu odwzorowano w skali logarytmicznej. Zdolność ta wynika z dużej ilości wag i biasów tej sieci neuronowej w porównaniu z ilością wektorów w zbiorze uczącym.



Rys. 37. Wykres logarytmiczny zależności błędu średniokwadratowego wyznaczonego dla zbioru treningowego od liczby epok uczenia

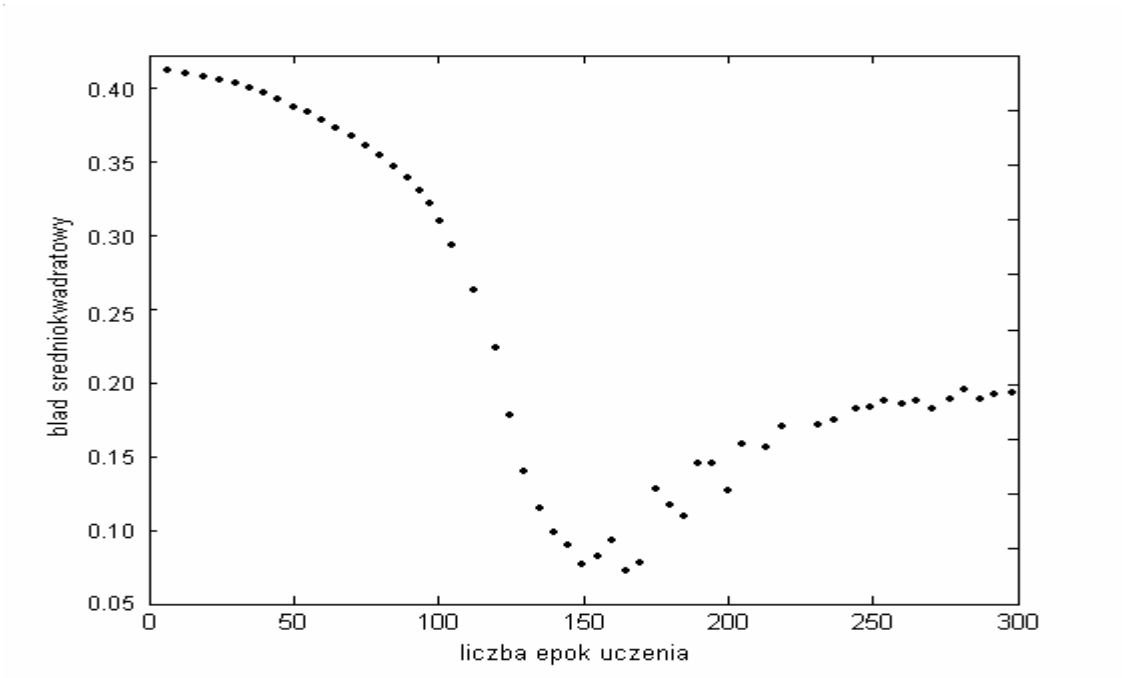
Zmniejszając wielkość sieci neuronowej (poprzez ograniczanie liczby neuronów w warstwach ukrytych) zmniejszeniu ulega również ilość wag i biasów sieci neuronowej. Powoduje to, że

sieć nie jest już w stanie dobrze zminimalizować błędu liczonego dla zbioru uczącego. Drugim, znacznie ważniejszym skutkiem jest zwiększenie zdolności uogólniających sieci neuronowej. Z drugiej strony zbyt duże ograniczanie liczby neuronów w sieci nie jest korzystne ze względu na utratę zdolności sieci do uczenia się. Wykresy wielkości błędu na zbiorze uczącym od liczby epok uczenia dla różnych struktur sieci neuronowych przedstawione są na rys.38.



Rys. 38. Wykres zależności błędu na zbiorze uczącym od liczby epok uczenia dla różnych struktur sieci neuronowych

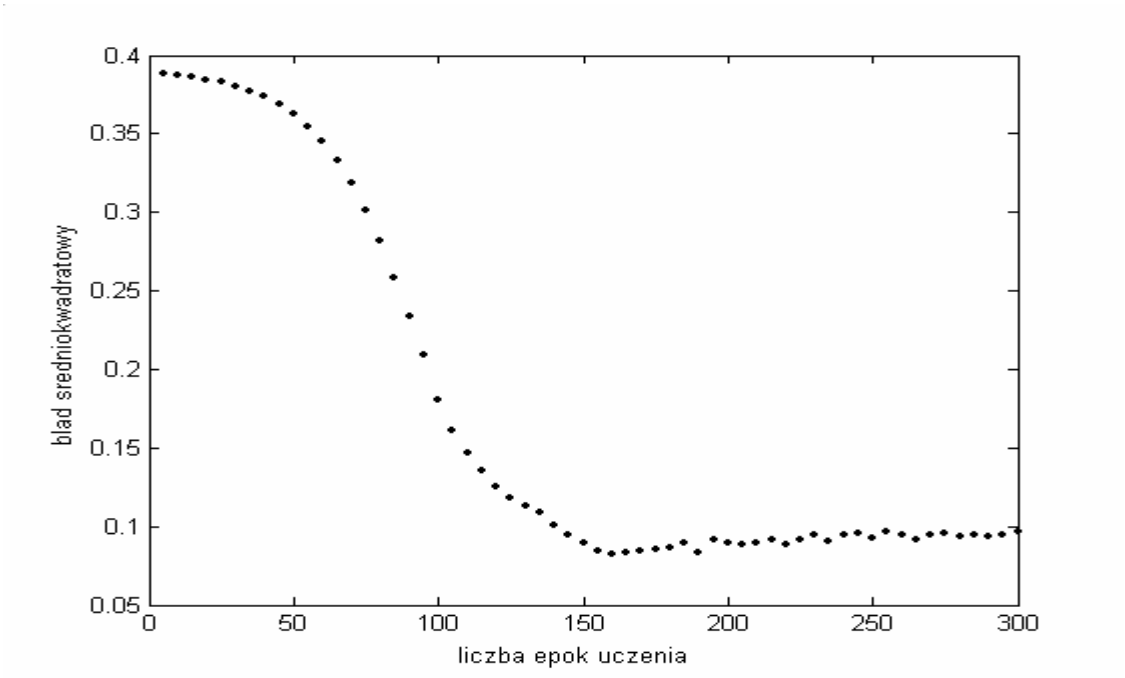
Zupełnie inaczej kształtuje się błąd dla zbioru testowego utworzonego z danych, które nie były prezentowane w procesie uczenia sieci. Wykres błędu na zbiorze testowym w zależności od liczby epok uczenia przedstawiony jest na rys. 39. Zbiór testowy utworzony został losowo z 5-ciu pacjentów, pozostała grupa przypadków została użyta do treningu sieci neuronowej. Procedura losowania przypadków wchodzących w skład zbioru testowego powtarzana była wielokrotnie, a wyniki błędu dla każdej epoki uczenia uśredniane arytmetycznie. Sieć neuronowa zastosowana w tym przypadku miała strukturę [8-8-1].



Rys. 39. Wykres zależności błędu wyznaczonego dla zbioru testowego od liczby epok uczenia sieci neuronowej dla sieci 8-8-1

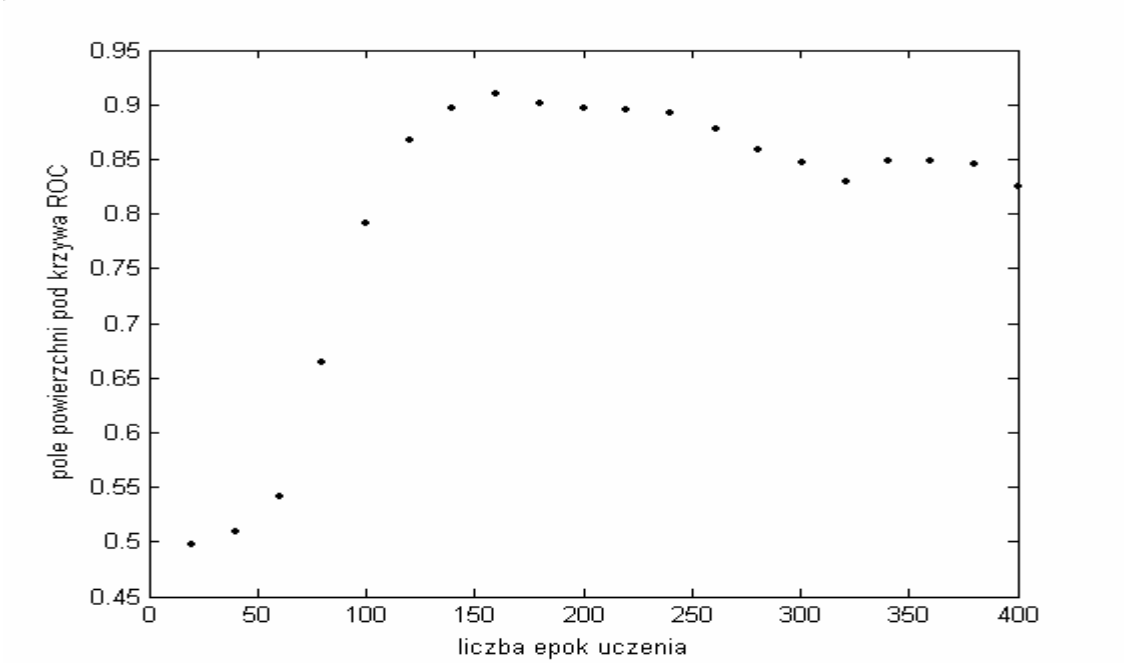
Z wykresu wynika, że przy zwiększaniu liczby epok uczenia zmniejszeniu ulega błąd dla zbioru testowego. Oznacza to, że sieć uczy się rozpoznawać zależności pomiędzy danymi wejściowymi a wyjściowymi i w rezultacie prognozować zadany parametr (dysplazję BPD) dla nowych przypadków. Dzieje się tak jednak tylko do określonego momentu, w którym następuje optimum - dalsze zwiększanie liczby epok uczenia powoduje zwiększenie błędu na zbiorze testowym. Wynika to z faktu przeuczenia sieci, której działanie polega bardziej na zapamiętaniu wszystkich kombinacji danych wejściowych i odpowiadających im danych wyjściowych zaprezentowanych w zbiorze treningowym, niż na wykrywaniu ogólniejszych zależności pomiędzy danymi.

Dla sieci neuronowej o mniejszej liczbie neuronów (prostszej strukturze) wykres ten wygląda podobnie (rys. 40), z jedną różnicą - błąd na zbiorze testowym dla dużej ilości epok uczenia nie rośnie lub rośnie znacznie wolniej, niż w przypadku sieci bardziej złożonej. Jak wynika z wcześniejszego wykresu (rys.38), sieć [2-1] nie była w stanie zminimalizować błędu na zbiorze uczącym poniżej pewnej wartości pomimo zwiększania liczby epok uczenia. Z tego samego powodu nie następuje wzrost błędu na zbiorze testowym.

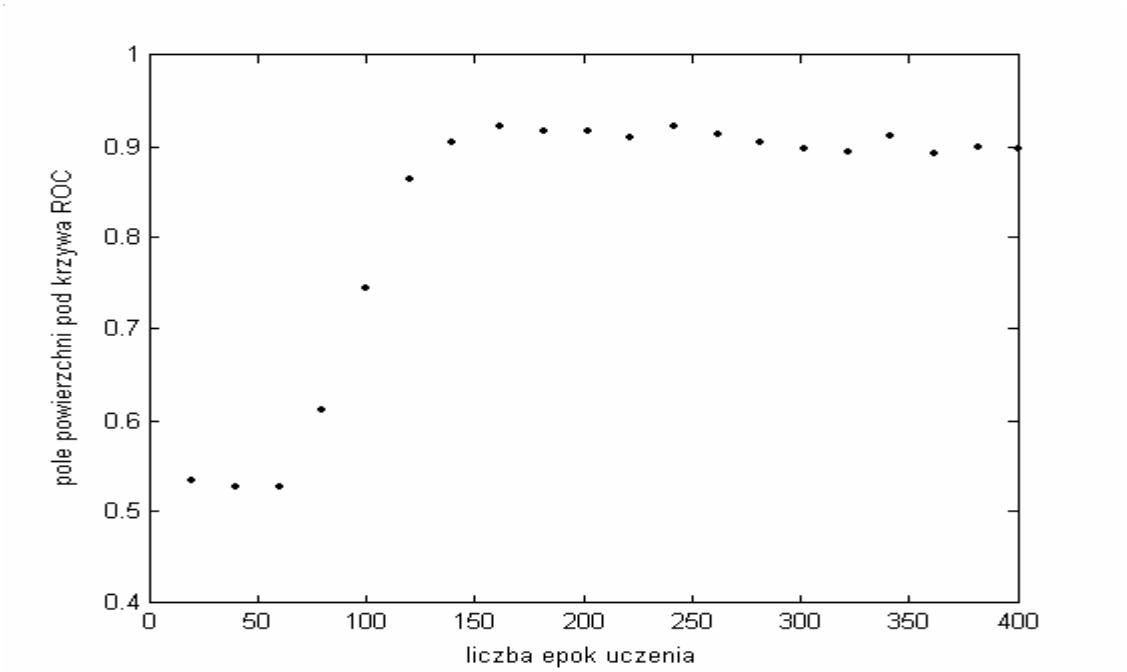


Rys. 40. Wykres zależności błędu wyznaczonego dla zbioru testowego od liczby epok uczenia sieci neuronowej dla sieci [2-1]

Do wykreślenia powyższych zależności błędu na zbiorze testowym zastosowano losowy dobór zbioru testowego. Aby porównać wyniki predykcji dysplazji oskrzelowo-płucnej i zdolności uogólniających sieci neuronowych i regresji logistycznej, podobnie jak poprzednio zastosowano walidację krzyżową 14-krotną. Jak wynika z wykresu na rys. 39, dla sieci [8-8-1] istnieje pewna wartość liczby epok uczenia, dla której błąd na zbiorze testowym osiąga minimum. Zatem w trakcie uczenia takiej sieci konieczne jest zatrzymanie procesu uczenia sieci po osiągnięciu tej właśnie liczby epok. Procedura taka nazywana jest algorytmem wczesnego stopu lub zatrzymanym uczeniem (ang. *early stopping*). Wraz ze zmianą wartości błędu na zbiorze testowym zmienia się również wartość pola powierzchni pod krzywą ROC. Wykres zależności AUC od liczby epok uczenia dla sieci [8-8-1] przedstawiony jest na rys. 41. Analogiczny wykres dla sieci o strukturze [2-1] przedstawiony jest na rys. 42. W takim przypadku sieć może być uczona dłużej niż poprzednia sieć [8-8-1] ze względu na bardziej płaski charakter zależności błędu na zbiorze testowym od liczby epok uczenia i mniejszą skłonność sieci do przeuczenia. Jak wynika z rys. 42, również pole powierzchni pod krzywą ROC nie maleje przy dłuższym uczeniu sieci. W takim przypadku liczba epok uczenia może i powinna być większa w porównaniu do sieci z przypadku poprzedniego.



Rys. 41. Wykres zależności pola powierzchni pod krzywą ROC od liczby epok uczenia sieci neuronowej [8-8-1]



Rys. 42. Wykres zależności pola powierzchni pod krzywą ROC od liczby epok uczenia sieci neuronowej [2-1]

Tabela 30 przedstawia wartości pola powierzchni pod krzywą ROC oraz błędu na zbiorze testowym dla sieci o różnych strukturach (różnej liczby neuronów w warstwach ukrytych) przy założeniu, że liczba epok procesu uczenia dobrana została optymalnie dla danej struktury sieci.

Tab. 30. Wartość pola powierzchni pod krzywą ROC, błędu średniokwadratowego obliczanego dla zbioru testowego oraz trafności prognozy dla różnych struktur sieci neuronowej

Struktura sieci neuronowej	AUC	błąd średniokwadratowy na zbiorze testowym	procentowa trafność prognozy
[2-1]	0.92 ± 0.03*	0.07	88.6
[3-1]	0.91 ± 0.03*	0.09	87.1
[4-1]	0.91 ± 0.03*	0.09	87.1
[2-2-1]	0.92 ± 0.03*	0.07	88.6
[3-3-1]	0.91 ± 0.03*	0.09	87.1
[4-4-1]	0.91 ± 0.03*	0.08	85.7
[8-8-1]	0.91 ± 0.03*	0.08	85.7

* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

Obserwując wyniki uzyskane przy użyciu różnych konfiguracji sieci neuronowych można zauważyć, że różnice w zdolności predykcyjnej są niewielkie. Dla sieci wszystkich przedstawionych w tabeli sieci oprócz struktur [2-1] i [2-2-1] konieczne było zastosowanie algorytmu wczesnego stopu po mniej więcej 170 epokach uczenia sieci. Dalszy trening tych sieci prowadził bowiem do spadku zdolności predykcyjnej.

Macierz pomyłek dla modelu wykorzystującego sieć neuronową o strukturze [2-1] przedstawiona jest w tabeli 31.

Tab. 31. Macierz pomyłek i trafność klasyfikacji dla modelu opartego o sieć [2-1]

	Przewidywany brak BPD	Przewidywane BPD		
Obserwowany brak BPD	41	3	specyficzność	93.1 %
Obserwowane BPD	5	21	czułość	80.8 %

6.2.2. Predykcja przy użyciu sieci radialnych

Wykorzystywane w pracy sieci radialne składają się z dwóch warstw - warstwy radialnej oraz warstwy wyjściowej, którą stanowi jeden neuron. W zależności od ilości neuronów w warstwie radialnej można wyróżnić dwa przypadki :

1. liczba neuronów równa jest liczbie wzorców uczących,
2. liczba neuronów jest mniejsza od liczby wzorców uczących.

W pierwszym przypadku sieć jest w stanie zminimalizować do zera błąd średniokwadratowy obliczany dla wektora uczącego. Oczywiście wiąże się z tym również utrata zdolności

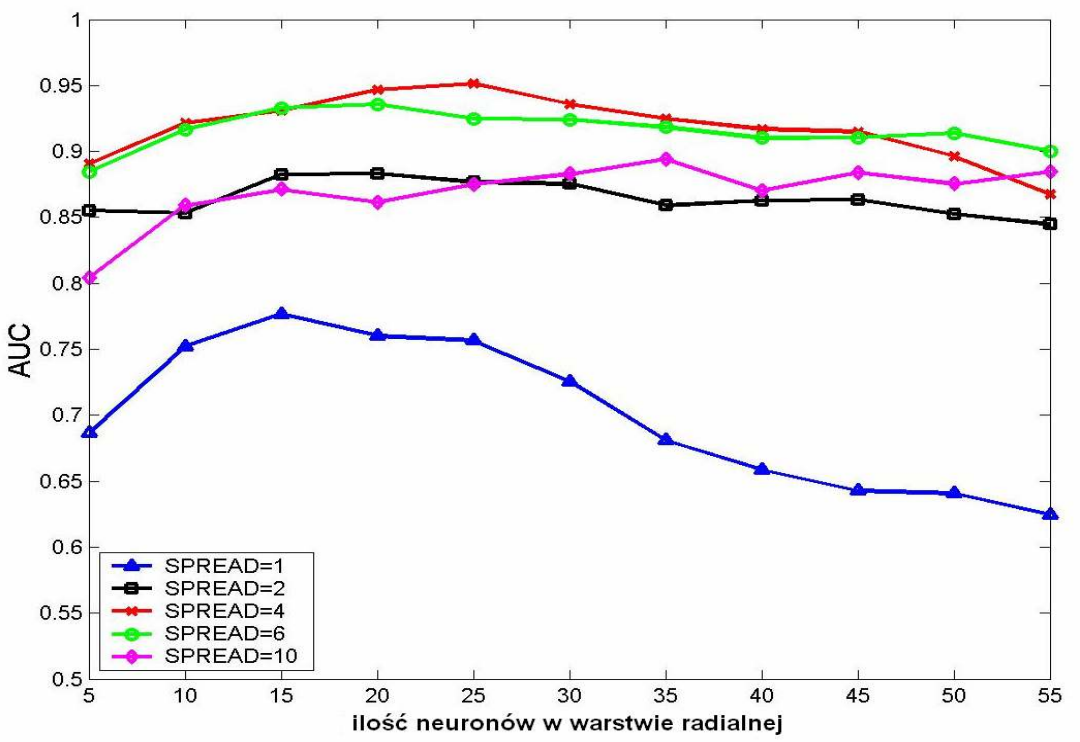
uogólniania, a ponadto przy dużej liczbie sygnałów wzorcowych struktura sieci rozrasta się do ogromnych rozmiarów. Dlatego też w dalszych symulacjach zastosowano strukturę określoną w przypadku drugim, dobierając eksperymentalnie liczbę neuronów warstwy radialnej oraz wartość parametru SPREAD (wzór 5.17) określającego szerokość funkcji radialnej. W tabeli 32 przedstawione zostały wartości maksymalnej, minimalnej i średniej arytmetycznej odległości (rozumianej w sensie euklidesowym) pomiędzy znormalizowanymi wektorami wejściowymi. Zakres zmienności parametru SPREAD zawierał się pomiędzy minimalną a maksymalną wartością tej odległości.

Tab. 32. Minimalna, maksymalna i średnia wartość odległości pomiędzy znormalizowanymi wektorami wejściowymi

maksymalna odległość	7.36
minimalna odległość	0.67
średnia odległość	3.85

Warstwa wyjściowa sieci neuronowej zawiera pojedynczy neuron, ponieważ sieć wykorzystana została do predykcji jednego parametru. Jego funkcja aktywacji ma kształt sigmoidalny ze względu na dychotomiczny charakter przewidywanej zmiennej (dysplazji BPD). Do uczenia sieci wykorzystano opisaną dokładniej w rozdziale 5.2 metodę podziału wektorów wejściowych na grupy przy użyciu algorytmu k-means oraz metodę propagacji wstecznej błędów do treningu warstwy wyjściowej. Podobnie jak poprzednio zastosowano walidację krzyżową 14-krotną.

Na rysunku 43 przedstawione zostały wykresy zależności pola powierzchni pod krzywą ROC od liczby neuronów w warstwie radialnej dla różnych wartości parametru SPREAD. Jak wynika z tego wykresu, pole powierzchni pod krzywą ROC osiąga maksimum dla sieci neuronowej zbudowanej z około 25 neuronów w warstwie radialnej i dla wartości parametru SPREAD równej około 4. Kolejne tabele 33 i 34 przedstawiają oprócz pola powierzchni pod krzywą ROC również wartość trafności predykcji oraz macierz pomyłek dla tego przypadku.



Rys. 43. Zależność pola powierzchni pod krzywą ROC od ilości neuronów w warstwie radialnej dla różnych szerokości krzywej radialnej

Tab. 33. Wartość pola powierzchni pod krzywą ROC i trafności predykcji dla sieci radialnej o maksymalnej zdolności predykcyjnej

Parametr	Wartość
AUC (pole powierzchni pod krzywą ROC)	$0.95 \pm 0.03^{*}$
ACC (trafność klasyfikowania) [%]	92.8

* wartość standardowego błędu dla AUC została określona wg wzoru 5.22

Tab. 34. Macierz pomyłek dla sieci radialnej o maksymalnej zdolności predykcyjnej

	Przewidywany brak BPD	Przewidywane BPD		
Obserwowany brak BPD	41	3	specyficzność	93.2 %
Obserwowane BPD	2	24	czułość	92.3 %

6.2.3. Wnioski dotyczące predykcji dysplazji przy użyciu sztucznych sieci neuronowych

W celu porównania zdolności predykcyjnych sieci neuronowych zastosowano otrzymany wcześniej metodą eliminacji wstecznej optymalny podzbiór zmiennych wejściowych. Zdolności predykcyjne sieci neuronowych jednokierunkowych wielowarstwowych zależą zarówno od ilości neuronów w warstwach ukrytych, jak również od sposobu uczenia sieci.

Sieci o wystarczająco rozbudowanej strukturze i uczone odpowiednio dużą liczbę epok są w stanie zminimalizować błąd średniokwadratowy dla zbioru treningowego do dowolnie małej wielkości. Wiąże się to jednak na ogół z brakiem właściwości generalizacji sieci, co objawia się dużymi wartościami błędu średniokwadratowego dla nowych, nie prezentowanych w procesie treningu sieci przypadków. Aby temu zapobiec, należy zapewnić odpowiednią strukturę sieci neuronowej albo przerwać proces uczenia w momencie, gdy błąd średniokwadratowy wyznaczany dla zbioru testowego osiągnie minimum. Pierwszy przypadek dotyczy sieci neuronowych o najprostszej strukturze ([2-1] i [2-2-1]), dla których błąd średniokwadratowy rośnie w bardzo nieznacznym stopniu wraz ze wzrostem liczby epok uczenia (rys. 40). W drugim przypadku konieczne jest przerwanie treningu, aby nie dopuścić do przeuczenia sieci. Jak wynika z wykresu przedstawionego na rys. 39, minimalna wartość błędu średniokwadratowego dla zbioru testowego osiągana jest po mniej więcej 150 epokach uczenia. Porównując wartości podstawowych wskaźników określających zdolność predykcyjną dla sieci o różnych strukturach można zauważyć, że różnica pomiędzy nimi jest niewielka. Dla sieci o bardziej złożonej strukturze trudniej jest znaleźć optymalną ilość epok uczenia. Sieci neuronowe jednowarstwowe są w stanie przewidzieć dysplazję oskrzelowo-płucną z trafnością porównywalną do modelu regresji logistycznej. Dla modeli tych wartość pola powierzchni pod krzywa ROC równa jest ok. 0.92, zaś trafność predykcji ok. 88%.

Nieco lepsze wyniki osiąga się przy zastosowaniu sieci radialnych. Dobierając liczbę neuronów w warstwie radialnej oraz wartość parametru SPREAD określającego szerokość krzywej dzwonowej funkcji aktywacji neuronów radialnych, można uzyskać model charakteryzujący się wartością pola powierzchni pod krzywą ROC równą w przybliżeniu 0.95 i trafnością predykcji ok. 92%. Jak wynika z rys. 43 optymalny model zbudowany jest z 25-ciu neuronów, zaś wartość parametru SPREAD równa jest w przybliżeniu średniej arytmetycznej wartości odległości pomiędzy znormalizowanymi wektorami uczącymi (tab. 31). Model wykorzystujący sieci radialne charakteryzuje się większą wartością czułości (por. tab. 30 i 33).

7. Podsumowanie

W niniejszej pracy przedstawiony został problem zastosowania metod regresji logistycznej oraz sztucznych sieci neuronowych do predykcji dysplazji oskrzelowo-płucnej u noworodków. Wykazano prawdziwość obu postawionych tez. Wykorzystując dane pobrane z bazy NIS oraz zgromadzone przy użyciu systemu rejestracji danych medycznych można skonstruować model umożliwiający predykcję dysplazji oskrzelowo-płucnej. Jak wynika z tabel 24,26,30 i 33, trafność prognozy, określana dla najlepszych uzyskanych modeli, wynosi ok. 90%, a pole powierzchni pod krzywą ROC ok. 0.95. Wartości te wyznaczone zostały przy zastosowaniu walidacji krzyżowej, zatem przy zapewnieniu rozdzielenia danych użytych do uczenia modelu i testowania jego zdolności predykcyjnej. Trafność predykcji i pole powierzchni pod krzywą ROC przyjmują zbliżone wartości zarówno dla modelu regresji logistycznej, jak również jednokierunkowych wielowarstwowych sieci neuronowych. Ponadto, jak pokazują dane zawarte w tabeli 30, zastosowanie sieci trójwarstwowych, jak również zwiększanie ilości neuronów w warstwach ukrytych, nie powoduje poprawy zdolności predykcyjnych. Szacuje się, zgodnie ze znanym twierdzeniem Vapnika-Chervonenkisa, że liczba wzorców uczących powinna być minimum 20 razy większa od liczby wag zastosowanej sieci neuronowej [9]. Warunek ten jest trudny do spełnienia przy dostępnej ilości danych, które można na obecnym etapie wykorzystać do predykcji dysplazji oskrzelowo-płucnej. Stąd też podobne rezultaty uzyskano stosując modele regresyjne.

Nieco lepsze wyniki uzyskano stosując sieci radialne. Uczenie takich sieci trwa krócej niż sieci trójwarstwowych, jednak sieci te na ogół składają się z większej ilości neuronów. Trafność prognozy i pole powierzchni pod krzywą ROC - dwa podstawowe wskaźniki zastosowane w pracy do określania zdolności predykcyjnej modeli - przyjmują nieco większe wartości (odpowiednio ok. 93% i 0.95 - tab.33). Należy tutaj jednak zauważyć względnie duże wartości standardowego błędu dla tych wielkości. Jak wynika z macierzy pomyłek (tabele 31 i 34), różnice w predykcji dotyczą tylko 3 pacjentów (z całkowitej liczby 70).

Generalnie, oceniając sieci neuronowe i regresję logistyczną można dojść do wniosku, że narzędzia te charakteryzują się porównywalną zdolnością predykcyjną dysplazji oskrzelowo-płucnej. Niewątpliwą przewagą regresji logistycznej nad sieciami neuronowymi jest prostota modelu i związana z tym faktem niewielka złożoność obliczeniowa algorytmu. Do dodatkowych zalet zaliczyć należy możliwość łatwej interpretacji otrzymanych zależności i wyników. Sieci neuronowe są strukturalnie bardziej skomplikowane, a współczynniki wagowe nauczonej sieci nie mają prostej interpretacji w analizowanym modelu. Techniki

modelowania dostępne dla regresji logistycznej, takie jak wybór optymalnego modelu, szacowanie dobroci dopasowania, czy testowanie statystycznej istotności zmiennych wejściowych nie mają jeszcze swoich ogólnie uznanych odpowiedników w przypadku sieci neuronowych. W rezultacie budowanie modelu sieci neuronowych w dużym stopniu zależy od intuicji, czynnika niemierzalnego i subiektywnego. Trening sieci neuronowej zajmuje znacznie więcej czasu obliczeniowego (niekiedy kilkanaście a nawet kilkaset razy), co przy braku znaczących zalet powoduje, że preferowanym modelem jest model regresji logistycznej.

Porównując uzyskane wyniki z pracami wspomnianymi we wstępie należy zauważyć, że w większości przypadków ich autorzy skupiają się głównie na poszukiwaniu czynników wpływających na rozwój dysplazji oskrzelowo-płucnej i określeniu ich siły wpływu na ryzyko wystąpienia BPD. Spośród wymienionych prac jedynie Rozycki i Narla [41] zbudowali model predykcyjny i oszacowali jego zdolności predykcyjne. Stosując regresję logistyczną wykorzystali oni do dopasowania modelu dane 116 pacjentów, a do testowania oddzielny zbiór danych zebranych w późniejszym czasie - 61 pacjentów. Uzyskali model charakteryzujący się 83% trafnością predykcji przy 82% czułości i 89% specyficzności. Do predykcji wykorzystali przy tym podstawowe wielkości dla pierwszych 14 dni hospitalizacji. Ponadto użyli również danych dla pierwszych ośmiu godzin, ale uzyskany model charakteryzował się niską trafnością prognozy. Warto zauważyć, że specyficzność użytej przez nich metody była większa od czułości. Porównując wartości czułości i specyficzności przedstawione w tabelach 25, 27 i 31 niniejszej pracy zauważyć można podobną zależność pomiędzy czułością i specyficznością. Większa jakość predykcji dla przypadków, u których nie rozwinęła się dysplazja, może wynikać z większej ich liczebności w całkowitej grupie badanych pacjentów. Z drugiej jednak strony, jak pokazują dane zawarte w tab. 34, zastosowanie sieci radialnych skutkuje uzyskaniem zbliżonych wartości czułości i specyficzności (odpowiednio 92% i 93%). Fakt też może sugerować lepsze zdolności predykcyjne tego narzędzia.

W literaturze medycznej nie udało się znaleźć żadnych informacji na temat sposobu wykorzystywania danych gromadzonych poprzez ciągły zapis informacji transmitowanych przez pulsoksymetry. Opierając się na doświadczeniu lekarzy Oddziału Intensywnej Terapii Noworodka określono więc sposób obliczania z uzyskanego strumienia danych wybranych wielkości, które następnie zastosowane zostały do predykcji dysplazji oskrzelowo-płucnej. Analiza optymalnych modeli regresji logistycznej wskazuje, że parametry uzyskane przy zastosowaniu systemu gromadzenia danych medycznych znajdują się w optymalnych

podzbiorach zmiennych niezależnych (tab. 14 i 18). Porównując krzywe ROC na rys. 31 i badając przy pomocy statystyki χ^2 istotność statystyczną parametrów uzyskanych za pomocą pulsoksymetrów można zauważyć, że zmienne te w badanych modelach są silnie znaczące statystycznie (rozdz. 6.1.2 i 6.1.3). W ten sposób wykazano prawdziwość drugiej tezy pracy.

W przyszłości rozbudowa systemu gromadzenia danych medycznych pozwoli wzbogacić model predykcyjny o nowe wielkości. Zebranie większej ilości przypadków w systemie spowoduje polepszenie zdolności predykcyjnej prezentowanych modeli. Możliwe będzie wykorzystanie sieci neuronowych o większej ilości neuronów, co może pokazać ich przewagę nad metodą regresji logistycznej. Duże znaczenie może mieć również zgromadzenie danych z innych urządzeń diagnostyczno - monitorujących, na przykład respiratorów transmitujących cyfrowo dane dotyczące składu mieszaniny oddechowej, jak również sposobu, przebiegu i intensywności wentylacji mechanicznej. W pracach dotyczących predykcji dysplazji oskrzelowo-płucnej (m.in. [6],[7],[46]) wykazano, że wielkości te mają duży wpływ na ryzyko rozwoju BPD. Ich zastosowanie może znacznie poprawić zdolność predykcyjną modelu. W takim przypadku zastosowana w pracy metoda przeszukania wszystkich możliwych podzbiorów zmiennych niezależnych stanie się operacją zbyt czasochłonną. Oczywiście można posłużyć się prostymi metodami heurystycznymi, takimi jak selekcja postępująca czy eliminacja wsteczna, jednak te metody często dostarczają wynik daleki od optymalnego rozkładu. Zaprezentowany w pracy algorytm genetyczny umożliwia efektywne znalezienie zbliżonego do optymalnego modelu użytego do predykcji dysplazji.

8. Bibliografia

[1] Akaike H., *Information theory as an extension of the maximum likelihood principle*, Second International Symposium on Information Theory, 1974, 267-281.

[2] Archer N., *Cardiovascular disease*, Roberton's Textbook of Neonatology, Rennie J.M, London, 2005, 4th ed., 634-636.

[3] Bellman R., *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.

[4] Belsley D.A., Kuh E., Welsch R.E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York, 1980.

[5] Breiman L., Spector P., *Submodel Selection and Evaluation in Regression. The X-Random Case.*, International Statistical Review, Vol. 60, No. 3, 1992, 291-319.

[6] Cunha G.S., Mezzacappa-Filho F., Ribeiro J. D., *Maternal and neonatal factors affecting the incidence of bronchopulmonary dysplasia in very low birth weight newborns*, J Pediatr (Rio J), 2003, 79(6), 550-556.

[7] Cunha G.S., Mezzacappa-Filho F., Ribeiro J. D., *Risk Factors for Bronchopulmonary Dysplasia in very Low Birth Weight Newborns Treated with Mechanical Ventilation in the First Week of Life*, Journal of Tropical Pediatrics, 2005, 51(6), 334-340.

[8] Demuth H., Beale M., *Neural Network Toolbox – for use with Matlab*, The Math Works Inc., 1998.

[9] Duch W., Korbicz J., Rutkowski L., Tadeusiewicz R., *Biocybernetyka i inżynieria biomedyczna, tom 6, Sieci neuronowe*, Exit, Warszawa, 2000.

[10] Electronics Industries Association: *EIA Standard RS-232-C Interface Between Data Terminal Equipment and Data Communication Equipment Employing Serial Data Interchange*, 1969.

[11] Eubank R.L., *Nonparametric Regression and Spline Smoothing*, CRC Press, 1999.

[12] Farstad T., Bratlid D., *Incidence and prediction of bronchopulmonary dysplasia in a cohort of premature infants*, Acta Paediatr., 1994, 83(1), 19-24.

[13] Fawcett T., *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*, Technical Report HPL-2003-4, HP Labs, 2003.

[14] Geisser S., *The predictive sample reuse method with applications*, Journal of The American Statistical Association, Vol. 70, No. 350, 1975, 320-328.

[15] Gilbert R., Keighley J., *The arterial/alveolar oxygen tension ratio. An index of gas exchange applicable to varying inspired oxygen concentrations.*, Am Rev Respir Dis., 1974, 109, 142-145.

[16] Goldberg D.E., *Algorytmy genetyczne i ich zastosowania*, Wydawnictwa Naukowo-Techniczne, 2003.

[17] Goutte C., Larsen J., *On Optimal Data Split For Generalization Estimation And Model Selection*, Proceedings of the 1999 IEEE Signal Processing Society Workshop, 1999, 225-234.

[18] Greenough A., Milner A.D., *Chronic lung disease*, Roberton's Textbook of Neonatology, Rennie J.M, London, 2005, 4th ed., 554-572.

[19] Gupta M.M., Jin L., Homma N., *Static and dynamic neural networks. From fundamentals to Advanced Theory.*, Wiley, New Jersey, 2003

[20] Hanley J.A., McNeil B.J., *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology, 1982, 143, 29-36.

[21] Henery R.J., *Methods of comparison.* in *Machine Learning, Neural and Statistical Classification*, Michie D., Spiegelhalter D.J., Taylor C.C. (ed.), Prentice Hall, 1994, 107-124.

[22] IEEE 802.3 (Ethernet) standard specifications, http://www.ieee802.org/3/

[23] Kleinbaum D.G., *Logistic Regression: A Self-Learning Text*, Springer-Verlag Telos, 1994.

[24] Kohavi R., John G.H., *Wrappers for Feature Subset Selection*, Artificial Intelligence, Vol. 97, 1997, 273-324.

[25] Koller D., Sahami M., *Toward optimal feature selection*, Machine Learning: Proceedings of the 13th International Conference, 1996, 284-292.

[26] Kruczek P., *Ocena przydatności prognostycznej sztucznej sieci neuronowej u urodzonych przedwcześnie noworodków z zespołem zaburzeń oddychania*, rozprawa doktorska, Collegium Medicum Uniwersytetu Jagiellońskiego, Wydział Lekarski, 2002.

[27] Kruczek P., Pietrzyk J.J., Sukiennik A., Wajs W., *Blood gases values forecasting by artificial neural network in prematurely born infants with respiratory distress*, Przegl Lek. 2002, 59, 34-37.

[28] Kwinta P, Kruczek, P., Stoch P., Wajs W., Pietrzyk J., *Results of continuous monitoring of hemoglobin saturation during the first month of life as predictors of retinopathy of prematurity*, Pediatric Research, 2005,58, 390.

[29] Lachenbruch P.A., Mickey M.R., *Estimation of Error Rates in Discriminant Analysis*, Technometrics, Vol. 10, No. 1, 1968, 1-11.

[30] Lampinen J., Laaksonen J., Oja E., *Neural Network Systems, Techniques and Applications in Pattern Recognition. Research rept.*, Technical Report B1, Helsinki University of Technology, Laboratory of Computational Engineering, 1997.

[31] Larose D.T.: *Data mining methods and models*, Wiley, New York, 2006

[32] Larose D.T., *Discovering knowledge in data. An introduction to data mining*, Wiley, New York, 2005.

[33] Lehmann E.L., *Testing statistical hypothesis*, Wiley, 2nd ed., New York, 1986.

[34] MacQueen J.B., *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1966, 281-297.

[35] MEDLINE - U.S. National Library of Medicine's bibliographic database http://www.nlm.nih.gov

[36] Menard S., *Applied Logistic Regression Analysis*, Sage Publications Series: Quantitative Applications in the Social Sciences, No. 106, 1995.

[37] Mitchell T., *Machine Learning*, McGraw-Hill Science/Engineering/Math, 1997.

[38] Mosteller F., Tukey J.W., *Data analysis, including statistics*, Handbook of Social Psychology, 2nd ed., 1968, 80-203.

[39] Nickalls R.W.D., Ramasubramanian R., *Interfacing the IBM-PC to medical equipment. The art of serial communication*, Cambridge University Press, 1st ed., 1995.

[40] NPB-295 Pulse Oximeter Operator's Manual, NELLCOR Puritan Bennett Inc.

[41] Rozycki H.J., Narla L., *Early versus late identification of infants at high risk of developing moderate to severe bronchopulmonary dysplasia*, Pediatr. Pulmonol., 1996, 21(6), 345-352.

[42] Rutkowski L., *Metody i techniki sztucznej inteligencji*, WN PWN, Wyd. 1, 2006.

[43] Schwarz G., *Estimating the dimension of a model*, The Annals of Statistics, 6(2), 1978, 461-464.

[44] Shapiro S.S., Wilk M.B., *An analysis of variance test for normality*, Biometrika, 52, 1965, 591-611.

[45] Somech R., Zangen S., Davidson S., Merlob P., *Prediction of bronchopulmonary dysplasia during first day of life in preterm infants*, The International Journal of Risk and Safety in Medicine, 1999, 101-107.

[46] Srisuparp P., Marks J.D., Khoshnood B., Schreiber M.D., *Predictive power of initial severity of pulmonary disease for subsequent development of bronchopulmonary dysplasia*, Biol Neonate, 2003 ,84(1), 31-36.

[47] Stone M., *Cross–validatory choice and assessment of statistical predictions*, Roy.Stat. Soc. B36, 1974, 111–147.

[48] Sukiennik A., *Sieci neuronowe jako narzędzie wykorzystane do prognozy wartości parametrów gazometrii krwi*, rozprawa doktorska, AGH, Wydział EAIiE, 2002.

[49] Tabachnick B.G., Fidell L.S., *Using Multivariate Statistics*, Harper Collins College Publishers, New York, 1996.

[50] Tadeusiewicz R., *Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami*, Akademicka Oficyna Wydawnicza, Warszawa, 1998.

[51] Tadeusiewicz R., *Sieci neuronowe*, Akademicka Oficyna Wydawnicza RM, Warszawa, 1993.

[52] Tadeusiewicz R., *Telemedycyna jako źródło nowych wyzwań naukowych*, rozdział w książce: Cierniak R. (red.): *Ewolucja czy rewolucja - nowoczesne techniki informatyczne*, Katedra Inżynierii Komputerowej Politechniki Częstochowskiej, Częstochowa, 2003, 237-277.

[53] Tadeusiewicz R., Izworski A., Majewski J., *Biometria*, Wydaw. AGH, Kraków, 1993.

[54] Tapia J.L., Agost D., Alegria A., Standen J., Escobar M., Grandi C., Musante G., Zegarra J., Estay A., Ramírez R., *Bronchopulmonary dysplasia: incidence, risk factors and resource utilization in a population of South American very low birth weight infants*, J Pediatr (Rio J). 2006, 82(1), 15-20.

[55] Tsai C.L., Hurvich C.M., *Regression and time series model selection in small samples*, Biometrika 76, 1989, 297-307

[56] Vinterbo S., Ohno-Machado L., *A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction*, Proc AMIA Symp., 1999, 984-988.

[57] Wahba G., *Spline Models for Observational Data*, SIAM Press, Philadelphia, 1990.

[58] Wahba G., Wold S., *A completely Automatic French Curve: Fitting spline functions by cross–validation*, Communications in Statistics, 4(1), 1975, 1-17.

[59] Wajs W., Wais P., Święcicki M., Stoch P., Maj G., Sukiennik A., Kruczek P., Pietrzyk J.J., *Classification of vectorized medical data sets using artificial immune algorithms*, Proceedings of IFAC workshop on Programmable Devices and Systems, 2004, 395-400.

[60] Witten I.H., Frak E., *Data mining. Practical Machine Learning Tools and Techniques.*, 2nd ed., Morgan Kaufmann series, 2005

[61] Yang J., Honavar V., *Feature subset selection using a genetic algorithm*, IEEE Intelligent Systems, Vol.13, 1998, 44-49.

Dodatek A

Dane wykorzystane w obliczeniach pobrane z bazy Neonatal Information System

Lp	MASAUR [g]	WIEKPL [tyg]	RESPIMV	AA	PDA	SURFACT	BPD
1	1360	29	0	0.2790	0	0	0
2	890	25	1	0.2475	1	0	1
3	1400	31	0	0.5130	0	0	0
4	700	24	1	0.3147	0	0	1
5	880	28	0	0.5429	0	0	0
6	985	32	0	0.2031	0	0	0
7	1100	28	0	0.3293	0	0	0
8	1110	28	1	0.0725	1	0	1
9	1100	28	0	0.8553	0	0	0
10	1300	29	0	0.1378	0	0	0
11	760	28	1	0.0809	0	1	1
12	1200	29	1	0.1221	0	1	1
13	700	25	1	0.2254	1	0	1
14	960	26	1	0.4877	1	0	1
15	900	28	0	0.1734	0	0	0
16	760	25	0	0.2568	1	0	1
17	1100	30	0	0.5227	0	0	0
18	1400	30	1	0.1043	0	1	0
19	1100	28	0	0.5591	0	0	0
20	1000	30	1	0.5295	0	0	0
21	1270	30	1	0.2165	1	1	0
22	1120	32	0	0.4280	0	0	0
23	1420	32	1	0.1146	1	1	0
24	860	27	1	0.2100	0	0	1
25	600	24	1	0.0591	1	1	1
26	750	27	1	0.1765	1	1	0
27	800	27	1	0.1828	1	1	0
28	1440	29	1	0.0502	0	0	0
29	1170	32	0	1.0000	1	0	0
30	860	28	1	0.1370	1	1	1
31	1300	29	1	0.1346	1	1	1
32	1100	26	1	0.1037	0	1	0
33	1200	28	1	0.0581	0	1	0
34	1100	28	1	0.0936	1	0	0
35	900	28	1	0.8175	0	0	0
36	1400	31	1	0.0960	1	1	1
37	1250	31	0	0.6849	0	0	0
38	940	28	1	0.0722	1	1	1
39	800	28	1	0.1033	1	1	1
40	1250	30	1	0.3167	0	0	0
41	1000	27	1	0.1433	1	1	1
42	600	25	1	0.1137	1	0	1
43	930	32	0	0.2939	0	0	0
44	950	28	1	0.1410	1	1	1
45	1200	31	1	0.0737	0	0	0
46	1400	31	0	0.5409	0	0	0
47	1400	28	1	0.2100	1	0	0
48	1095	28	1	0.0600	0	1	1

49	980	28	0	0.9500	0	0	0
50	1140	28	1	0.1700	0	0	0
51	820	26	1	0.1500	0	0	0
52	1500	31	1	0.1000	0	0	0
53	800	25	1	0.3100	0	0	1
54	760	25	1	0.2500	0	0	1
55	990	26	1	0.3400	0	0	0
56	980	27	0	0.3800	0	0	0
57	1400	32	1	0.2100	0	0	0
58	770	24	1	0.1000	0	1	1
59	1500	30	1	0.0800	0	1	1
60	1500	30	1	0.2600	0	0	0
61	1130	27	1	0.2700	0	0	0
62	1200	30	1	0.1100	0	0	1
63	1400	32	0	0.9500	0	0	0
64	1240	32	1	0.2300	1	0	0
65	650	28	1	0.2100	0	1	1
66	700	28	1	0.2000	0	1	0
67	800	28	1	0.2800	0	0	0
68	720	25	1	0.1000	1	1	1
69	650	25	1	0.0600	0	0	0
70	1280	32	1	0.1200	1	1	0

Dane wykorzystane w obliczeniach uzyskane z systemu rejestracji danych medycznych.

Lp	SPO2MEAN [%]	SPO2DEV [%]	LOW85 [%]	HIGH94 [%]	BPMMEAN [bpm]
1	97.53	4.03	1.00	90.35	151.02
2	96.11	3.78	1.60	78.96	148.63
3	97.97	1.42	0.03	98.47	124.69
4	96.32	2.81	0.60	80.58	152.48
5	98.05	2.17	0.38	96.46	135.09
6	97.67	2.01	0.09	94.82	131.76
7	96.52	3.77	0.91	76.65	155.23
8	97.66	2.70	0.45	89.21	153.06
9	97.97	3.29	0.91	91.26	138.80
10	94.83	4.49	2.22	58.99	147.46
11	94.78	5.48	4.40	69.12	142.26
12	93.59	4.11	1.98	46.36	148.18
13	92.78	3.60	2.79	32.65	132.48
14	97.23	3.46	0.74	84.27	148.38
15	93.98	3.62	1.20	47.22	148.20
16	93.14	3.52	2.18	39.05	139.93
17	96.74	2.45	0.24	85.32	144.40
18	95.33	3.29	1.11	69.56	144.39
19	95.42	2.94	0.87	70.32	134.20
20	96.42	3.55	0.93	77.69	146.07
21	97.39	3.51	1.51	87.00	138.83
22	98.99	2.82	0.97	96.00	148.18
23	95.61	3.29	1.05	73.22	146.00
24	95.95	4.04	1.93	77.05	147.06
25	94.10	3.72	0.94	53.85	141.62
26	95.72	2.88	0.35	71.93	141.56

27	95.55	3.39	1.40	73.04	128.87
28	96.14	4.20	2.17	78.44	144.98
29	97.14	3.12	1.17	92.27	138.56
30	94.04	5.01	4.78	55.05	146.18
31	95.35	4.01	2.01	67.49	150.46
32	96.26	3.15	0.63	79.25	131.09
33	95.11	4.17	3.08	70.37	126.25
34	98.29	2.40	0.27	94.94	146.97
35	96.15	2.85	0.82	78.41	145.23
36	93.93	4.88	3.88	57.11	154.37
37	97.91	2.81	0.87	93.11	138.20
38	96.91	3.64	1.04	82.38	132.42
39	95.68	4.71	3.24	73.74	125.08
40	95.94	2.63	0.54	78.63	137.01
41	94.74	3.55	1.53	62.74	141.18
42	94.95	3.74	0.82	63.36	150.48
43	96.24	2.12	0.10	81.05	140.09
44	94.50	3.75	1.89	60.88	155.51
45	95.63	2.68	0.31	74.81	132.25
46	96.95	2.37	0.19	86.63	143.25
47	95.61	3.25	1.07	73.74	132.80
48	93.86	5.93	4.22	55.19	156.15
49	97.27	2.24	0.13	88.47	150.87
50	95.48	5.35	4.33	72.63	136.18
51	95.70	3.29	0.97	73.86	136.81
52	93.77	3.06	0.66	44.61	146.41
53	95.60	3.88	1.13	75.35	148.91
54	94.28	2.42	0.34	47.30	144.66
55	94.90	2.97	0.70	64.43	138.11
56	95.96	3.66	1.73	81.60	145.70
57	92.81	3.60	1.92	33.87	139.08
58	93.42	4.07	2.17	41.64	152.90
59	94.17	3.42	1.55	55.74	151.36
60	98.63	3.60	0.83	96.28	141.88
61	95.91	2.59	0.39	79.29	133.97
62	96.27	3.94	1.51	82.55	150.79
63	98.32	2.16	0.29	96.41	142.92
64	94.62	4.64	4.42	65.80	140.50
65	95.08	2.87	0.97	69.03	136.85
66	96.56	2.79	0.60	86.82	143.53
67	96.86	3.14	0.94	87.83	132.85
68	95.13	2.87	0.77	66.47	155.07
69	94.36	3.05	0.61	51.41	149.57
70	96.55	4.87	3.23	77.29	144.67

Dodatek B

Listing programu MatLab realizującego poszukiwanie optymalnego podzbioru zmiennych niezależnych przy użyciu algorytmu genetycznego (opis w rozdziale 5.4)

file genetic_search.m

```
clear;

% --- initial parameters ---
population_size=10;
pcross=0.75;
pmut=0.05;
max_iterations=300;


% --- data loading ---
load('medical_data.mat');
xall=[spo2dev_tr spo2mean_tr bpmmean_tr spo2dev spo2mean bpmmean high94 low85 masa aa pda respimv surfact
wiekpl];
yall=[bpd ones(length(bpd),1)];

n0=length(find(yall(:,1)==0));
n1=length(find(yall(:,1)==1));
n=length(yall(:,1));
dev0=-2*(n0*log(n0/n)+n1*log(n1/n));


% --- initial population generation ---
indep_var=size(xall,2);
models_evaluated=zeros(1,2^indep_var);
parents=floor(rand(1,population_size)*(2^indep_var-1))+1;


% --- genetic algorithm main loop ---

iteration=1;
mean_fitness(1)=dev0;
min_fitness(1)=dev0;
mov_avg(1)=dev0;

while iteration<max_iterations
   iteration=iteration+1;

   % --- evaluate population ---
   for i=1:population_size
      if models_evaluated(parents(i))==0
         eval_result(i)=check_fit_parent(parents(i),xall,yall);
         models_evaluated(parents(i))=eval_result(i);
      else
         eval_result(i)=models_evaluated(parents(i));
      end
   end

   % --- roulette wheel selection ---
   eval_ranking=dev0-eval_result-min(dev0-eval_result)+10;
   roulette_prob=eval_ranking/sum(eval_ranking);

   roulette_prob_cum(1)=roulette_prob(1);
   for i=2:population_size
      roulette_prob_cum(i)=roulette_prob_cum(i-1)+roulette_prob(i);
   end

   clear offspring selected_parents;
```

```
   for i=1:population_size
      rand_nr=rand(1,1);
      selected_parent(i)=min(find(roulette_prob_cum>rand_nr));
   end


   % --- crossover ---
   i=1;
   while i<=population_size
      rand_nr=rand(1,1);
      if rand_nr<=pcross
         offspring(i)=crossover_parents(parents(selected_parent(i)),parents(selected_parent(i+1)),indep_var);
         offspring(i+1)=crossover_parents(parents(selected_parent(i+1)),parents(selected_parent(i)),indep_var);
      else
         offspring(i)=parents(selected_parent(i));
         offspring(i+1)=parents(selected_parent(i+1));
      end
      i=i+2;
   end


   % --- mutation ---
   for i=1:population_size
      rand_nr=rand(1,1);
      if rand_nr<=pmut
         offspring(i)=mutate(offspring(i),indep_var);
      end
   end

   % --- new population ---
   clear parents;
   parents=offspring;

   % --- evaluation of some fitness indicators ---
   mean_fitness(iteration)=mean(eval_result);
   min_fitness(iteration)=min(eval_result);
   if iteration<=30
      mov_avg(iteration)=mean(mean_fitness(1:iteration));
   else
      mov_avg(iteration)=mean(mean_fitness(iteration-30:iteration));
   end
end

optimal_pop_result=min(eval_result);
optimal_pat=parents(find(eval_result==min(eval_result)));

fprintf(1, 'optimal AIC from last population : %f\n',optimal_pop_result);
fprintf(1, 'optimal independent variables binary representation : %d\n',optimal_pat);
```

file crossover_parents.m

```
function crossresult = crossover_parents(parent1,parent2,indep_var)

binparent1=dec2bin(parent1,indep_var);
binparent2=dec2bin(parent2,indep_var);

crosspt=floor(rand(1,1)*(indep_var-1))+1;
binresult=cat(2,binparent1(1:crosspt),binparent2(crosspt+1:indep_var));

crossresult=bin2dec(binresult);
```

file mutate.m

```
function mutation = mutate(obj,indep_var)

binobj=dec2bin(obj,indep_var);
mutpoint=floor(rand(1,1)*indep_var)+1;
if (binobj(mutpoint)=='0') binobj(mutpoint)='1'; else binobj(mutpoint)='0'; end

mutation=bin2dec(binobj);
```

file check_fit_parent.m

```
function fitparent = check_fit_parent(komb,xall,yall)

clear xvect;
binmap=dec2bin(komb,size(xall,2));
i=1;
for j=1:size(binmap,2)
   if strcmp(binmap(j),'1')
      xvect(:,i)=xall(:,j);
      i=i+1;
   end
end


% --- logistic regression ---
[logitCoef,dev,stats] = glmfit (xvect, yall, 'binomial', 'logit');
p=size(xvect,2)+1;
n=size(xvect,1);

fitparent=dev+2*p*(n/(n-p-1));
```