ZASTOSOWANIE NARZĘDZI KLASYFIKACYJNYCH OPARTYCH NA TECHNIKACH STATYSTYCZNYCH I METODACH SZTUCZNEJ INTELIGENCJI W OCENIE PRAWDOPODOBIEŃSTWA ISTNIENIA RAKA JAJNIKA

Agata Smoleń, Uniwersytet Medyczny w Lublinie

Wprowadzenie

Ciągłym problemem jest wspomaganie przedoperacyjnej diagnostyki różnicowej guzów przydatkowych. Wymierne korzyści może przynieść zaawansowana analiza danych z użyciem technik statystycznych i metod sztucznej inteligencji oraz możliwość połączenia uzyskanych wyników w jednym projekcie. Wydobycie z danych użytecznej wiedzy stanowi podstawę do jej wykorzystania w poprawie przedoperacyjnej diagnostyki różnicowej i podejmowania decyzji w oparciu o szacowane prawdopodobieństwa istnienia procesu złośliwego. Odkrycie skomplikowanych zależności, wyjaśnienie zaobserwowanych tendencji i możliwość prospektywnego wykorzystania wydają się być pomocne w praktyce klinicznej.

Do wydobycia z danych użytecznej wiedzy najczęściej nie wystarczają tradycyjne techniki statystyczne. Decyduje o tym konieczność zbadania dużej liczby danych, przy uwzględnieniu wielu parametrów. Ponieważ zależności lub ich interakcje są zazwyczaj nieznane, nie można zakładać istnienia prostych związków. Powoduje to konieczność zastosowania „inteligentnych" metod uczących się, przy wykorzystaniu odpowiedniego zasobu wiedzy w zakresie prowadzonych badań.

Predykcja prawdopodobieństwa przynależności do grupy guzów złośliwych lub niezłośliwych na podstawie wartości zmiennej zależnej (wynik badania histopatologicznie określającego, czy dany guz jest złośliwy czy niezłośliwy) i niezależnych (wyniki badań cech klinicznych, ultrasonograficznych i stężenia antygenu CA-125) jest jednym z ważniejszych zagadnień wspomagania wstępnej diagnostyki różnicowej guzów przydatkowych u kobiet.

Badania zostały przeprowadzone w oparciu o dane chorych diagnozowanych i leczonych z powodu guzów przydatkowych w Klinice Ginekologii Onkologicznej i Ginekologii Akademii Medycznej w Lublinie w latach 2002-2007. Analiza mocy testu wykazała, że podstawą do realizacji badań powinny być dane zebrane od minimum 300 pacjentek leczonych z powodu wykrytego guza przydatków.

Spośród 1003 kobiet z wykrytymi w badaniu ultrasonograficznym guzami przydatkowymi do analizy zakwalifikowano 379 chorych, u których również oceniono badane zmiany z użyciem trójwymiarowej ultrasonografii i oznaczono stężenie antygenu CA-125 przed zabiegiem operacyjnym. Dodatkową grupę prospektywną (zbiór testowy), włączoną w celach oceny dokładności diagnostycznej konstruowanych modeli, stanowiło 33 chorych. Weryfikację rodzaju guza stanowiły wyniki pooperacyjnego badania histopatologicznego. Analizowano wyniki badania ultrasonograficznego z zastosowaniem skali szarości (2D), oceny przepływu z zastosowaniem kolorowego i spektralnego Dopplera oraz badania z obrazowaniem 3D sonoangiografii z pomiarem objętości brył na podstawie oprogramowania VOCAL™ i funkcją angiohistogramu. Dla każdej pacjentki oceniono cechy kliniczne i ultrasonograficzne wymienione w tabeli 1.

Tabela 1. Charakterystyka analizowanych cech klinicznych i ultrasonograficznych badanej grupy kobiet.

Oceniane cechy
wiek
status menopauzalny
wskaźnik masy ciała BMI (kg/m2)
umiejscowienie guza (jedno/obustronne)
typ histologiczny guza wg WHO
stopień zaawansowania guza złośliwego wg FIGO
przegrody w guzie (>3mm lub brak)
elementy lite w guzie (obecne/brak)
wyrośla endofityczne w guzie (>3mm lub brak)
największy wymiar guza (mm)
objętość guza (ml)
lokalizacja naczyń (centralna/obwodowa/brak naczyń)
kolor wg IOTA
indeks pulsacji PI
indeks oporu RI
najwyższa prędkość skurczowa PSV (cm/s)
średnia maksymalna szybkość przepływu TAMXV (cm/s)
indeks waskularyzacji VI
indeks przepływu FI
indeks naczyniowo-przepływowy VFI
stężenie antygenu CA-125 (U/ml)

Wśród 379 zakwalifikowanych do analizy chorych z guzami jajnika było 160 (42,2%) kobiet z nowotworami złośliwymi oraz 219 (57,8%) kobiet z niezłośliwymi guzami przydatkowymi. W prospektywnej grupie 33 chorych weryfikacja histopatologiczna potwierdziła 10 (30,3%) guzów złośliwych oraz 23 (69,7%) guzów niezłośliwych.

Na podstawie wartości badanych zmiennych budowany jest model, który może wspomagać przewidywanie prawdopodobieństwa istnienia procesu złośliwego dla nowych przypadków. Jest to problem klasyfikacyjny, w którym w wyniku działania predykcyjnego algorytmów eksploracji danych odpowiedzią będzie guz złośliwy lub niezłośliwy, na podstawie przyjętych progów prawdopodobieństwa. Do rozwiązywania tego typu problemów klasyfikacyjnych można zastosować następujące metody: analizę dyskryminacyjną, uogólnione modele liniowe (regresja logistyczna i probitowa), uogólnione modele addytywne (MARS – algorytm rekurencyjnego podziału przestrzeni cech do budowy modelu w postaci krzywych sklejanych), drzewa klasyfikacyjne standardowe lub ze wzmacnianiem, sztuczne sieci neuronowe oraz inne metody eksploracji danych, tj. metoda wektorów wspierających, naiwny klasyfikator Bayesa, metoda K najbliższych sąsiadów [Smoleń 2002, 2008].

Jednoczynnikowa analiza statystyczna

Do wykrycia istotności różnic między porównywanymi grupami guzów złośliwych i niezłośliwych (zmiennymi niepowiązanymi) użyto testu U Manna-Whitneya dla cech ilościowych (ze względu na brak zgodności z rozkładem normalnym i niejednorodność wariancji) oraz testu jednorodności χ^2 dla cech jakościowych. Przyjęto 5% ryzyko błędu wnioskowania i związany z nim poziom istotności $p<0,05$ wskazujący na istnienie istotnych statystycznie różnic.

Ponadto w celu podsumowania i potwierdzenia istotności badanych parametrów (predyktorów) w różnicowaniu guzów złośliwych i niezłośliwych zastosowano algorytm doboru i eliminacji zmiennych, który można zastosować dla zmiennych zależnych (guz złośliwy lub niezłośliwy) oraz niezależnych ilościowych i jakościowych. Odpowiadające prawdopodobieństwo testowe (poziom p) oznacza, że im mniejsza jest jego wartość, tym silniejsza jest sugestia, że istnieje związek między zmienną zależną a predyktorem. Graficzna prezentacja uzyskanego wyniku pozwala na porównanie, który z parametrów wykazuje najsilniejszy związek z różnicowaniem guzów oraz umożliwia porównanie wpływu interakcji między badanymi parametrami.

Dokładność testu została określona przez porównanie pomiędzy wynikami testu a pooperacyjnym wynikiem histopatologicznym, który stwierdza rzeczywistą sytuację chorobową.

Sposoby grupowania danych i oceny wartości diagnostycznej podstawowych parametrów przedstawia tabela 2.

Tabela 2. Tablica kontyngencji przedstawiająca związek badanego czynnika ryzyka z pooperacyjną weryfikacją histopatologiczną.

Wynik testu	Rozpoznanie kliniczne		Razem
	Guz złośliwy	Guz niezłośliwy	
Pozytywny	TP	FP	TP+FP
Negatywny	FN	TN	FN+TN
Razem	TP+FN	FP+TN	TP+FP+FN+TN

Skróty w tabeli oznaczają: TP - wyniki prawdziwie pozytywne; FP - wyniki fałszywie pozytywne; FN - wyniki fałszywie negatywne; TP - wyniki prawdziwie negatywne.

Odsetki wyników prawdziwie dodatnich, prawdziwie ujemnych, fałszywie dodatnich i fałszywie ujemnych to wyznaczniki testu, na podstawie których decyduje się, czy dany test zastosować. Po wykonaniu testu bardzo istotna jest odpowiedź, jakie jest prawdopodobieństwo istnienia procesu złośliwego.

Zastosowane podstawowe miary dokładności testu to:

Czułość (*sensitivity* – SENS) – prawdopodobieństwo zarejestrowania testu pozytywnego (nieprawidłowe poszczególne cechy kliniczne) u kobiet rzeczywiście chorych na nowotwór złośliwy: $SENS = \frac{TP}{TP + FN} \cdot 100\%$.

Specyficzność (*specifity* – SPEC) – inaczej swoistość, czyli prawdopodobieństwo uzyskania wyniku negatywnego (prawidłowe cechy kliniczne) u kobiet z guzem niezłośliwym: $SPEC = \frac{TN}{FP + TN} \cdot 100\%$.

Wskaźnik wiarygodności (*likelihood ratio* – LR) odzwierciedla wartość predykcji wyniku testu. Określa stopień, w jakim wynik testu zmienia prawdopodobieństwo obecności guza złośliwego.

Wskaźnik wiarygodności dodatniego wyniku testu (LR+) wskazuje, na ile wynik dodatni zwiększa szansę wystąpienia guza złośliwego w porównaniu z szansą określoną przed zastosowaniem testu:

$$LR+ = \frac{SENS}{1 - SPEC}.$$

Wskaźnik wiarygodności ujemnego wyniku testu (LR-) wskazuje, na ile wynik ujemny zmniejsza szansę wystąpienia guza złośliwego w porównaniu z szansą określoną przed zastosowaniem testu:

$$LR- = \frac{1 - SENS}{SPEC}.$$

Przyjmuje się, że test ma rzeczywistą wartość diagnostyczną, jeżeli LR wynosi około 10 i więcej lub około 0,1 i mniej. Wyniki między 5 a 10 oraz między 0,1 a 0,2 wskazują, że test jest przydatny. LR między 0,5 a 2 oznacza, że test nie wpływa w sposób istotny na ocenę prawdopodobieństwa złośliwości guza. Porównanie LR dla różnych testów umożliwia również szybką ocenę użyteczności badanych parametrów (testów).

Zastosowano także dwie inne metody porównania stosowanych testów:

Dodatnią wartość predykcyjną (*positive predictive value* - PPV), która określa prawdopodobieństwo stwierdzenia nowotworu złośliwego na podstawie pozytywnego wyniku testu

$$PPV = \frac{TP}{TP + FP} \cdot 100\%.$$

Ujemną wartość predykcyjną (*negative predictive value* - NPV), która określa prawdopodobieństwo wykluczenia obecności nowotworu złośliwego przy negatywnym wyniku testu

$$NPV = \frac{TN}{FN + TN} \cdot 100\%.$$

Dokładność (*accuracy* - ACC) - odsetek pozytywnego wyniku testu wśród kobiet, u których rzeczywiście występuje guz złośliwym, oraz negatywnego wyniku testu u kobiet rzeczywiście bez guza złośliwego. Dokładność obliczono wg wzoru:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \cdot 100\%$$ [Smoleń 2002, 2008].

Wartości decyzyjne przyjęte dla mierzonych parametrów to wartości graniczne, których przekroczenie pozwala na dokonanie kwalifikacji osoby badanej, u której istnieje podejrzenie guza złośliwego. W sytuacjach, kiedy wartości testu diagnostycznego były ciągłe lub miały wiele kategorii, obliczono wartości predykcyjne dla najlepiej różnicujących charakter guza jajnika wartości granicznych (*cutoff point*). W przeprowadzonych analizach przyjęto, że wartość graniczna obliczana dla zmiennych mierzalnych ma oddzielać pacjentki z guzami niezłośliwymi od chorych z nowotworami złośliwymi.

W przeprowadzonych obliczeniach zmiany czułości i specyficzność testów przy przesuwaniu wartości granicznej przedstawiono w formie krzywych ROC (*Receiver Operating Characteristic Curves*) oraz wykorzystano dodatkowe informacje, porównując pola pod krzywą ROC. Konstrukcja krzywych ROC jest efektywną metodą, pomocną w ustaleniu wartości progowej dla testu diagnostycznego [Hanley i McNeil 1982, 1983]. Najbardziej użyteczną klinicznie metodą przedstawiania właściwości testu jest opisanie jego wyników za pomocą LR i/lub pola powierzchni pod krzywą ROC [Jaeschke i wsp. 1998].

Wieloczynnikowa analiza statystyczna

W dalszej części pracy oceniono, czy i które z wybranych parametrów (zmienne niezależne) mają istotny statystycznie wpływ na prawdopodobieństwo istnienia nowotworu złośliwego lub niezłośliwego (zmienna zależna). Oprócz cech klinicznych, ultrasonograficznych

i dopplerowskich do konstrukcji modelu użyto stężenia antygenu CA-125, jako najczęściej oznaczanego markera raka jajnika. Użyto klasyfikacji wzorcowej, której zadaniem było przypisanie każdego badanego przypadku do jednej z dwóch klas (guz złośliwy bądź niezłośliwy). W celu znalezienia najlepszej kombinacji cech istotnie wpływających na istnienie procesu złośliwego i obliczenie indywidualnego prawdopodobieństwa istnienia nowotworu jajnika zastosowano analizę dyskryminacyjną, regresję logistyczną i probitową, model krzywych sklejanych - MARS, drzewa klasyfikacyjne standardowe lub ze wzmacnianiem, sztuczne sieci neuronowe oraz metodę wektorów wspierających, naiwny klasyfikator Bayesa i metodę K najbliższych sąsiadów.

Dostępny zbiór danych podzielono w sposób losowy na dwa zbiory: uczący i walidacyjny. Trzeci zbiór testowy stanowiła prospektywna grupa 33 pacjentek.

Analiza dyskryminacyjna

W celu oceny, które z badanych parametrów najbardziej przyczyniają się do dyskryminacji guzów złośliwych i niezłośliwych zastosowano analizę ogólnych modeli dyskryminacyjnych metodą krokową oraz metodą najlepszego podzbioru z minimalną liczbą efektów wynoszącą 2. Obliczona została wartość wielowymiarowej statystyki lambda Wilksa, przybliżona wartość statystyki F oraz odpowiadający jej poziom istotności w celu oceny istotności mocy dyskryminacyjnej z uwzględnieniem kombinacji analizowanych parametrów. Im wartość lambda Wilksa jest bliższa zeru tym lepsza jest moc dyskryminacyjna. Ze względu na dyskryminację tylko dwóch grup: guzy złośliwe i niezłośliwe skonstruowano tylko jedną funkcję dyskryminacyjną. Istnienie funkcji i istotność mocy dyskryminacyjnej potwierdzono testem χ^2, a przydatność funkcji dyskryminacyjnych potwierdzono wartościami współczynników korelacji kanonicznej. Do obliczenia wartości funkcji dyskryminacyjnych dla każdego przypadku użyto wartości surowych współczynników. Natomiast do interpretacji wkładu każdej zmiennej do funkcji dyskryminacyjnej posłużono się wartościami współczynników standaryzowanych. Standaryzowane współczynniki funkcji dyskryminacyjnej wskazują, jak silny jest wpływ danego parametru na różnicowanie grup guzów złośliwych i niezłośliwych, czyli jaki jest indywidualny wkład każdego parametru do funkcji dyskryminacyjnej. Przy wyliczaniu standaryzowanych współczynników brane są pod uwagę wszystkie pozostałe zmienne. Obliczona została również wartość własna, na postawie której wylicza się procent wariancji międzygrupowej wyjaśniającej udział danej funkcji w dyskryminowaniu. Dodatkowe informacje uzupełniające uzyskano, wyznaczając współczynniki struktury czynnikowej, czyli współczynniki korelacji między pojedynczymi zmiennymi a funkcją dyskryminacyjną, które nie są pod wpływem pozostałych zmiennych, przeciwnie do współczynników standaryzowanych. Po określeniu i sprawdzeniu istotności funkcji dyskryminacyjnej klasyfikację przypadków umożliwiły funkcje klasyfikacyjne. Dany przypadek klasyfikowano do grupy, dla której wartość funkcji klasyfikacyjnej była większa. Do poprawy klasyfikacji uwzględniono prawdopodobieństwa a priori proporcjonalnie do wielkości grup. Ponadto zastosowano inną miarę w wielowymiarowej przestrzeni zmiennych dyskryminacyjnych z użyciem klasyfikacji na zasadzie minimalizacji odległości Mahalanobisa i prawdopodobieństw a posteriori. Dany przypadek był klasyfikowany do

grupy, dla której uzyskano największe prawdopodobieństwo a posteriori [Dobosz 2004, Stanisz 2007].

Regresja logistyczna i probitowa

W celu oceny prawdopodobieństwa istnienia procesu złośliwego skonstruowano modele regresji logistycznej i probitowej. Zastosowano wstępującą analizę regresji logistycznej z estymacją quasi-Newtona. Metoda regresji logistycznej pozwoliła na obliczenie tzw. logarytmu wiarygodności, służącego do oceny dopasowania modelu wraz z wartością statystyki χ^2 i poziomem prawdopodobieństwa p. Gdy p<0,05, analizowany model różnił się istotnie od modelu uwzględniającego tylko wyraz wolny. Do oceny mocy predykcji uzyskanego modelu obliczono współczynniki $R^2_{McFaddena}$ i $R^2_{Nagelkerke'a}$, które są uogólnieniem współczynnika determinacji stosowanego w regresji liniowej, określającego, jak duża część zmienności analizowanej cechy zależnej powiązana jest ze zmiennością badanych parametrów niezależnych. Ponadto obliczone zostały: oceny parametrów modelu z 95% przedziałami ufności i asymptotycznymi błędami standardowymi. Istotność parametrów modelu sprawdzono przy pomocy wartości testu t-Studenta lub wartości statystyki χ^2 Walda i poziomu prawdopodobieństwa p. Obliczone zostały również: iloraz szans dla zmiany jednostkowej poszczególnych parametrów oraz iloraz szans dla zmiany równej zakresowi analizowanych zmiennych z 95% przedziałami ufności. W tzw. powiązaniu hierarchicznym zastosowano test ilorazu wiarygodności LR funkcji straty (największej wiarygodności) między poprzednim modelem a modelem po dodaniu kolejnej zmiennej dołączonej do modelu. Do oceny, czy dołączenie zmiennej w sposób istotny poprawia dopasowanie modelu do danych użyto testu χ^2 i poziomu p dla precyzji dopasowania.

Ponadto zastosowano wstępującą analizę regresji probitowej z estymacją quasi-Newtona. Obliczono wartości χ^2 przy porównywaniu skonstruowanego modelu i modelu tylko z wyrazem wolnym. Obliczone zostały: oceny parametrów modelu z asymptotycznymi błędami standardowymi. Istotność estymatorów parametrów modelu sprawdzono przy pomocy wartości testu t-Studenta i poziomu prawdopodobieństwa p [Gore i Altman 1997, Stanisz 2000, Dobosz 2004, Stanisz 2007].

Model regresji logistycznej przedstawiał się następująco: $p = \frac{1}{1+e^{-z}}$, gdzie *e* jest stałą matematyczną, granicą ciągu $\lim_{n\to\infty}(1+\frac{1}{n})^n = e = 2{,}71828...$, a *z* jest kombinacją liniową zmiennych x_i i ocen b_i zawartych w modelu: $z = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$. Wartości przewidywane należące do przedziału <0,1> można traktować jako prawdopodobieństwo istnienia procesu złośliwego bądź niezłośliwego. W zastosowanym modelu wartości prawdopodobieństwa $\leq 0{,}5$ klasyfikowały nowotwór jako niezłośliwy, a przypadki z wartościami predykcyjnymi >0,5 jako nowotwór złośliwy. Przy pomocy klasyfikacji przypadków i ilorazu szans >1 oceniano, czy klasyfikacja jest lepsza niż można by oczekiwać przez przypadek oraz obliczano wartości predykcyjne uzyskanego modelu [Hosmer i Lemeshow 1989].

Model krzywych sklejanych - MARS

Uogólnione modele addytywne, a w szczególności metoda MARS, ze względu na możliwości wyjaśniania badanego problemu w sytuacji, gdy analizuje się wiele parametrów oraz obserwuje nieliniowe korelacje i interakcje między czynnikami wykazuje bardzo często wysokie wartości predykcyjne. Model konstruowany przy użyciu metody MARS wykorzystuje metodę rekurencyjnego podziału przestrzeni cech do budowy modelu klasyfikacyjnego w postaci krzywych sklejanych. Algorytm ten to dwuetapowa procedura stosowana sukcesywnie, aż do otrzymania żądanego modelu. W pierwszym etapie budowano model, rozpoczynając od najprostszego modelu, z funkcją bazową o stałej wartości, następnie zwiększając jego złożoność poprzez dodanie kolejnych funkcji bazowych Przeszukiwanie następuje dla każdej zmiennej i możliwych węzłów, przestrzeni funkcji bazowych i dodawanie do modelu tych funkcji, które maksymalizują pewną miarę dobroci dopasowania modelu, czyli minimalizują błąd predykcji, aż do osiągnięcia maksymalnego żądanego stopnia złożoności modelu. Definiowano więc parametry modelu, takie jak: maksymalna liczba funkcji bazowych (50 funkcji bazowych), rząd interakcji (testowano rząd 2, 3 i 4) między zmiennymi, wielkość „kary" (C) za dodanie kolejnej funkcji do modelu (przyjęto 1 lub 2), oraz wartość próg (przyjęta 0,0005), która zabezpiecza przed nadmiernym dopasowaniem modelu. Następnie uruchamiana była procedura wsteczna, usuwania z modelu najmniej istotnych funkcji bazowych, czyli takich, które dawały mały wkład do poprawy dobroci modelu i których usunięcie najmniej pogarszało dopasowanie modelu. Usuwanie nieistotnych funkcji bazowych stosowane jest dla lepszej generalizacji „wiedzy" wydobytej z danych uczących, tzn. mniejszy jest wtedy błąd predykcji dla nowych przypadków. Miarą dopasowania jest błąd (GCV) uogólnionego sprawdzianu krzyżowego [Pasztyła 2005, StatSoft Polska, IPS].

Modele drzew klasyfikacyjnych standardowych i ze wzmacnianiem

W celu oceny możliwości zróżnicowania guzów złośliwych i niezłośliwych z zastosowaniem badanych parametrów przeprowadzono analizy z użyciem standardowych drzew klasyfikacyjnych CART i CHAID oraz drzew wzmacnianych i algorytmu losowych lasów.

Przed przystąpieniem do analizy z użyciem drzew CART wybrano następujące parametry:

- reguła podziału – miara Giniego,
- prawdopodobieństwa a priori – szacowane z próby uczącej,
- koszty błędnych klasyfikacji – równe,
- kryterium stopu – bezpośrednie zatrzymanie typu FACT (frakcja obiektów – przyjęto 5%),
- minimalna liczność węzła końcowego (10% próby uczącej),
- maksymalna liczba poziomów (głębokość) drzewa n=10,
- maksymalna liczba węzłów n=1000,
- szacowanie błędu za pomocą zbioru testowego.

Graficzną prezentację uzyskanego modelu stanowi drzewo klasyfikacyjne. Pod każdym węzłem, który nie jest liściem, zapisywane jest kryterium podziału podgrupy docierającej do tego węzła na mniejsze podgrupy, które trafiają do węzłów-potomków. Kryterium podziału przeprowadzanego w danym węźle jest takie samo dla wszystkich elementów próby uczącej, które znalazły się w tym węźle. Elementy próby są przesuwane aż do węzła końcowego, czyli liścia drzewa, któremu zwykle przypisuje się etykietę tej klasy analizowanego problemu dyskryminacji, z której pochodzi najwięcej elementów próby uczącej, które dotarły do tego liścia. Obok gałęzi podano warunki podziału, jakie muszą być spełnione, aby element próby losowej, który znalazł się w danym węźle macierzystym, trafił do danego węzła potomka. W każdym węźle podawana jest klasa większościowa, czyli ta, do której należy większość elementów podgrup próby uczącej, które znalazły się w tym węźle.

Kolejnym krokiem było zestawienie rankingu ważności predyktorów (rangi od 0 do 100), których najwyższa wartość oznacza największy wpływ danej zmiennej na zmienną zależną (guz złośliwy lub niezłośliwy). Następnym etapem analizy była analiza trafności przewidywania przez zastosowanie skonstruowanego drzewa klasyfikacyjnego do przewidywania przynależności do klas w próbie testowej. Jakość modelu oceniono również z użyciem współczynnika LIFT, którego wartość większa od 1 oznacza wysoką jakość predykcyjną danego liścia w skonstruowanym drzewie klasyfikacyjnym.

Drugą metodą użytą był algorytm CHAID budowy drzew klasyfikacyjnych, w której kolejne węzły mogą być dzielone na wiele grup: z każdego węzła mogą wychodzić więcej niż dwie gałęzie.

CART budował drzewa nadmiernie rozbudowane i „przycinał" je z uwzględnieniem zmian współczynnika błędu. CHAID buduje drzewo do momentu, aż spełnione zostaną założenia kryterium stopu.

Przed przystąpieniem do analizy z użyciem drzew CHAID wybrano następujące parametry:

- koszty błędnych klasyfikacji – równe,
- minimalna liczność węzła końcowego (przyjęto n=50),
- maksymalna liczba węzłów n=1000,
- szacowanie błędu za pomocą v-krotnego sprawdzianu krzyżowego i zbioru testowego.

Ponadto zastosowano do prawdopodobieństw poprawkę Bonferroniego, która jest stosowana w celu „utrudnienia" uznania za statystycznie istotny wyniku pojedynczego testu, przy wielokrotnym przeprowadzaniu testów w oparciu o te same dane.

Ostatecznie zbudowano drzewa typu CART i CHAID w sposób interakcyjny [Łapczyński 2005, StatSoft Polska, IPS].

Do oceny możliwości poprawy mocy predykcji z zastosowaniem drzew klasyfikacyjnych użyto algorytmu drzew wzmacnianych i losowych lasów.

Przed przystąpieniem do analizy z użyciem wzmacnianych drzew klasyfikacyjnych wybrano następujące parametry:

- prawdopodobieństwa a priori – szacowane z próby uczącej,
- koszty błędnych klasyfikacji – równe,
- kryterium stopu – najmniejsza liczność węzła drzewa podlegającego podziałom (przyjęto n=15),
- minimalna liczność węzła powstającego w wyniku podziału (przyjęto n=1),
- maksymalna liczba poziomów drzewa n=10,
- maksymalna liczba wszystkich węzłów tworzących drzewo n=3,
- szacowanie błędu za pomocą zbioru testowego.

Zaprezentowano wykres średniego błędu dla różnej liczby drzew tworzących model i wybrano optymalną liczbę na podstawie oceny błędu w próbie testowej. Przedstawiono również wykres ważności predyktorów. Ponadto wykorzystano wykres zysku, który pokazuje procent obserwacji poprawnie zaklasyfikowanych do grup guzów złośliwych i niezłośliwych, jeśli uwzględnia się x procent przypadków (oś x) o największych prawdopodobieństwach klasyfikacyjnych dla wybranej klasy, to będą one stanowić y procent wszystkich poprawnie zaklasyfikowanych przypadków (oś y) z tej klasy.

Ponadto użyto algorytmu losowy las, który polega na zbudowaniu zespołu drzew, a następnie stosowaniu ich do przewidywania wartości zmiennej zależnej albo przynależności obserwacji do klasy.

Dla algorytmu losowych lasów przyjmowano następujące parametry:

- prawdopodobieństwa a priori – szacowane z próby uczącej,
- koszty błędnych klasyfikacji – równe,
- kryterium stopu – najmniejsza liczność węzła drzewa podlegającego podziałom (przyjęto n=15),
- minimalna liczność węzła powstającego w wyniku podziału (przyjęto n=5),
- maksymalna liczba poziomów drzewa n=10,
- maksymalna liczba wszystkich węzłów tworzących drzewo n=100,
- zatrzymanie zaawansowane: liczba cykli dla wyznaczenia błędu – 10 i procentowy spadek błędu – 5,
- szacowanie błędu za pomocą v-krotnego sprawdzianu krzyżowego.

Do interpretacji wyników zastosowano ranking ważności predyktorów oraz wykresy zysku [Demski 2005, StatSoft Polska, IPS].

Sztuczne sieci neuronowe

Do obliczenia prawdopodobieństwa istnienia procesu złośliwego na podstawie danych klinicznych, ultrasonograficznych i markera CA-125 zastosowano sztuczne sieci neuronowe.

Zadanie klasyfikacji guzów złośliwych i niezłośliwych na podstawie obliczonych prawdopodobieństw było realizowane przy pomocy sieci liniowych, perceptronów wielowarstwowych, sieci o radialnych funkcjach bazowych oraz probabilistycznych sieci neuronowych.

Podstawową metodą uczenia sieci wielowarstwowych był algorytm wstecznej propagacji błędów z uwagi na kierunek przepływu informacji o błędzie. Użyto funkcji logistycznej do transformacji każdego z węzłów w warstwie ukrytej i wyjściowej. Ta nieliniowa funkcja przedstawia się następująco: $Tf(\sum_{i=1}^{n} w_i x_i) = \frac{1}{1+e^{-\sum_{i=1}^{n} w_i x_i}}$, gdzie $Tf(\sum_{i=1}^{n} w_i x_i)$ jest wartością wyjściową aktywującą neuron i, a $w_i x_i$ jest wartością wejściową neuronu do sieci (w_i – wagi; x_i – pojedyncze wartości zmiennej).

Uczenie jednokierunkowych sieci wielowarstwowych przebiegało w trybie z nauczycielem, na podstawie zbioru uczącego. Celem uczenia była minimalizacja wartości błędu modelu neuronowego przeprowadzona poprzez modyfikację wartości parametrów, tzw. wag sieci. Działanie modelu neuronowego wymagało zdefiniowania metody przekształcenia numerycznych wartości wyjściowych sieci w informację dotyczącą przypisanej klasy. Dla wartości nominalnej zmiennej wyjściowej reprezentowanej przez pojedynczy neuron stosowana była metoda dwustanowa, która klasom przypisywała odpowiedniki numeryczne 0 lub 1. O interpretacji zaliczonego obiektu do klasy reprezentowanej przez daną wartość numeryczną decydowały określane parametry: poziom akceptacji (przyjęto 0,75) i poziom odrzucenia (przyjęto 0,25). Jeżeli wartość wyjściowa neuronu jest większa od poziomu akceptacji, to nowotwór został zaklasyfikowany jako złośliwy, jeżeli mniejsza od poziomu odrzucenia - jako niezłośliwy. Natomiast jeżeli wartość zawierała się pomiędzy poziomem akceptacji i odrzucenia, sieć nie była w stanie podjąć decyzji o przynależności przypadku. Przed podjęciem decyzji o wyborze architektury sieci do rozwiązania danego problemu przetestowano wiele różnorodnych modeli sieci. Kontrolowanie zdolności do generalizacji było możliwe dzięki walidacji i zdefiniowaniu zbioru walidacyjnego, przy pomocy którego możliwa jest ocena wartości błędu (zwiększanie się wartości błędu w zbiorze walidacyjnym przy jednoczesnym zwiększaniu się wartości błędu dla zbioru uczącego świadczy o przeuczeniu sieci). Powodem może być błędnie dobrana struktura sieci (zbyt rozbudowana). Sprawdzenie poprawności działania sieci umożliwiły statystyki klasyfikacyjne, na podstawie których obliczono wartości predykcyjne. Strukturę otrzymanych najlepszych modeli przedstawiono graficznie z uwzględnieniem neuronów wejściowych, ukrytych i wyjściowych. Model neuronowy pozwolił również uszeregować według ważności przyjęte zmienne objaśniające, co zostało przeprowadzone przy pomocy analizy wrażliwości [Tadeusiewicz i Lula 2001, Tadeusiewicz 1993, 1999, 2006]. Ze skonstruowanych modeli stworzono zespół, wyznaczając prognozę za pomocą każdego z nich, a następnie poddając ją uśrednieniu [Lula 2005].

SVM

W celu klasyfikacji badanych guzów złośliwych i niezłośliwych na podstawie badanych parametrów użyto metody SVM, która polegała na budowaniu nieliniowych granic decyzyjnych. Ze względu na fakt, iż zagadnienie przynależności do grup guzów złośliwych i niezłośliwych jest zagadnieniem klasyfikacji, w metodzie SVM użyto dwóch różnych typów modeli: pierwszy ze stałą C (C-SVM) i drugi ze stałą ni (ni-SVM). Do wyboru optymalnych wartości algorytmu (wartości stałych uczenia: C lub ni), które nie są znane a priori, użyto 10-krotnego sprawdzianu krzyżowego, który pozwolił na ustalenie najlepszych parametrów, dla których uzyskano najmniejszy średni błąd. Następnie wybierano typ funkcji jądrowej, który był używany przy konstruowaniu przestrzeni cech modelu SVM. Użyto jądra typu liniowego, a następnie RBF. Ponadto ze względu na duże zakresy zmienności zmiennych wejściowych, co mogłoby generować dużo wektorów wspierających i skutkować przeuczeniem modelu, przeskalowano zmienne wejściowe, tak aby ich wartości należały do przedziału [0, 1]. Ze względu na niezrównoważoną liczbę guzów złośliwych i niezłośliwych w badanej grupie zastosowano różne wartości kary (koszty błędnych klasyfikacji) wynoszące 2 lub 3 dla grupy guzów złośliwych, by przeciwdziałać tendencji modelu SVM do błędnego klasyfikowania przypadków jako guzy niezłośliwe (grupa liczniejsza). Im wyższa kara w danej klasie, tym trudniej model będzie przypisywał do niej nowe przypadki [Demski 2005, StatSoft Polska, IPS].

Naiwny klasyfikator Bayesa

Do różnicowania guzów złośliwych i niezłośliwych zastosowano również naiwny klasyfikator Bayesa, ponieważ jest to ogólnie przyjęta metoda zaprojektowana dla zadań klasyfikacyjnych, na podstawie założenia, że rozkład badanych parametrów w grupach guzów złośliwych i niezłośliwych jest niezależny. Wykorzystano różne rozkłady warunkowe zmiennych niezależnych: normalny, lognormalny i Poissona. Analiza ta pozwoliła na wybór zmiennych istotnie różnicujących grupy guzów złośliwych i niezłośliwych. Zobrazowano je na wykresach prawdopodobieństawa a posteriori w funkcji wartości predyktora i punkty krytyczne przy ustalonych wartościach innych predyktorów [Demski 2005, StatSoft Polska, IPS].

Metoda K-najbliższych sąsiadów

Zastosowano metodę K-NN, w której zamiast dopasowywać model poszukiwano podobnych obiektów, przyjmując, że podobne obiekty trafią do tej samej klasy. Kluczowymi parametrami metody są: przyjęta miara odległości i liczba najbliższych sąsiadów. Użyto standaryzowanych wartości parametrów mierzalnych i jednorodnego uśredniania odległości sąsiadów od badanej obserwacji bez uwzględniania wag, przy zastosowaniu miary odległości Manhattan. Optymalną liczbę sąsiadów dobierano z użyciem 10-krotnego sprawdzianu krzyżowego. Przewidywania metody k- najbliższych sąsiadów wyznaczane były na podstawie głosowania odpowiedzi dla k obiektów [Demski 2005, StatSoft Polska, IPS].

Zespoły modeli

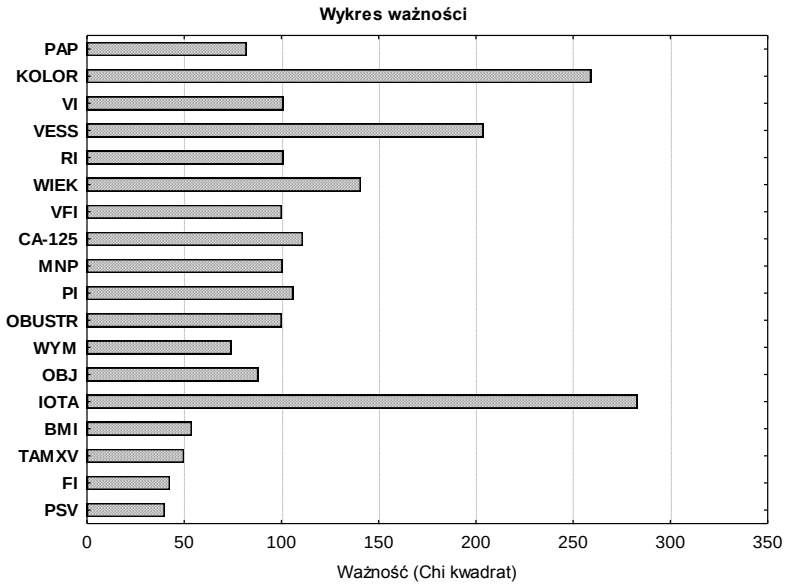
Tworzenie dla jednego zbioru danych wielu różnych modeli jest często stosowaną praktyką. Używając niezależnie techniki różnych metod (np. analiza dyskryminacyjna, uogólnione modele liniowe, uogólnione modele addytywne, drzewa klasyfikacyjne, sztuczne sieci neuronowe oraz inne metody sztucznej inteligencji), wybiera się model sprawdzający się najlepiej na danych testowych. Zadaniem modelu predykcyjnego jest obliczenie przewidywanej wartości zmiennej zależnej (np. guz złośliwy lub niezłośliwy) na podstawie zadanych wartości zmiennych zależnych. Uzyskiwane predykcje dla zbioru uczącego mogą znacznie różnić się dla wielu modeli, mimo że wszystkie modele uczone były na tych samych danych. Wynika to z faktu, iż dane testowe z konieczności dają wynik w pewnym stopniu losowy, tzn. model działający dobrze z konkretnymi danymi testowymi może inaczej zachować się w przypadku innego zestawu danych. Problem ten można rozwiązać, łącząc modele w zespół. Zespoły modeli predykcyjnych mają zwykle większą zdolność do generalizacji wiedzy niż ich pojedyncze składowe, ponieważ lepiej przewidują wartości zmiennej zależnej dla nowych zestawów wartości zmiennych niezależnych. Często również zespół modeli jest lepszy od najlepszego, wchodzącego w jego skład modelu. Dlatego najlepiej użyć całego zestawu modeli [Bishop 1995]. Zespoły modeli predykcyjnych tworzą wspólną predykcję zmiennej zależnej dla danego zestawu wartości zmiennych niezależnych. W takim przypadku jednym ze sposobów przewidywania, do jakiej klasy trafi dany obiekt, jest proste głosowanie, tzn. jako wynik wybierana jest klasa, którą wskazało najwięcej modeli (tzw. zasada „głosowania k-sędziów"). Można także zastosować ważone głosowanie lub uśrednianie [StatSoft Polska, IPS].

Podsumowaniem uzyskanych wyników analiz dla wszystkich skonstruowanych modeli była ocena poprawnie i niepoprawnie zaklasyfikowanych przypadków i obliczone wartości predykcyjne dla zbioru uczącego, walidacyjnego i testowego oraz ogółem dla całej grupy badanej i obu zbiorów testujących skuteczność. Ponadto do porównania skonstruowanych modeli z pojedynczymi parametrami i uzyskanych modeli między sobą skonstruowano krzywe ROC.

Analizy statystyczne przeprowadzono z użyciem programu *STATISTICA Data Miner*.

Podsumowanie wyników jednoczynnikowej analizy badanych parametrów w przedoperacyjnym różnicowaniu guzów przydatkowych

Podsumowując jednoczynnikową analizę badanych parametrów i ich przydatność w różnicowaniu guzów złośliwych i niezłośliwych, zastosowano algorytm doboru i eliminacji istotnych predyktorów jakościowych i ciągłych na podstawie statystyki χ^2 i uzyskiwanego poziomu p. Graficzną prezentację uzyskanych wyników posortowanych wg poziomu istotności $p<0,01$ dla najlepszych predyktorów przedstawia rys. 1.



Rys. 1. Histogram ważności predyktorów posortowanych wg poziomu istotności p.

Najistotniejszymi parametrami posortowanymi wg poziomu istotności p okazały się wyrośla endofityczne (PAP), cecha „Kolor", indeks unaczynienia (VI), lokalizacja naczyń (VESS), indeks oporu (RI), wiek chorej, indeks naczyniowo-przepływowy (VFI), a następnie stężenie antygenu CA-125, status menopauzalny chorej (MNP), indeks pulsacji (PI), obustronność guza (OBUSTR), największy wymiar guza (WYM), objętość guza (OBJ), morfologia ultrasonograficzna (IOTA), wskaźnik masy ciała (BMI), średnia maksymalna prędkość przepływu (TAMXV), indeks przepływu (FI) i maksymalna prędkość skurczowa (PSV). Ze względu na poziom istotności $p<0,01$ wszystkie przedstawione na rysunku zmienne były następnie przedmiotem dokładniejszych poszukiwań najlepszych kombinacji cech z wykorzystaniem metod sztucznej inteligencji.

Przydatność diagnostyczną analizowanych pojedynczych testów dla wartości progowej najlepiej dyskryminującej guzy złośliwe i niezłośliwe przedstawia tabela 3. Ocenę przeprowadzono z uwzględnieniem liczby wyników prawdziwie pozytywnych (TP), fałszywie pozytywnych (FP), fałszywie negatywnych (FN) i prawdziwie negatywnych (TN) oraz wartości prognostycznych, tj. czułość (SENS), specyficzność (SPEC), iloraz wiarygodności (LR+ i LR-), dodatnia i ujemna wartość predykcyjna (PPV i NPV) oraz dokładność diagnostyczna (ACC).

Tabela 3. Wartości prognostyczne poszczególnych parametrów dla optymalnej wartości progowej w różnicowaniu złośliwych i niezłośliwych guzów przydatkowych.

Parametr (optymalna wartość graniczna)	TP	FP	FN	TN	SENS	SPEC	LR+	LR-	PPV	NPV	ACC
WIEK (55 lat)	103	58	99	377	0,51	0,87	3,82	0,57	0,64	0,79	0,75
BMI (25 kg/m^2)	112	115	90	320	0,55	0,74	2,10	0,61	0,49	0,78	0,68
MNP (po mnp)	120	85	82	350	0,59	0,80	3,04	0,50	0,59	0,81	0,74
OBUSTR	100	56	102	379	0,50	0,87	3,85	0,58	0,64	0,79	0,75
SEPT >3mm	158	188	44	247	0,78	0,57	1,81	0,38	0,46	0,85	0,64
PAP > 3mm	91	55	111	380	0,45	0,87	3,56	0,63	0,62	0,77	0,74
WYM (80mm)	143	149	59	286	0,71	0,66	2,07	0,44	0,49	0,83	0,67
OBJ (500ml)	62	36	140	399	0,31	0,92	3,71	0,76	0,63	0,74	0,72
VESS (obwod.)	165	96	37	339	0,82	0,78	3,70	0,24	0,63	0,90	0,79
KOLOR (3)	100	18	102	417	0,50	0,96	11,96	0,53	0,85	0,80	0,81
PI (0,69)	108	184	94	251	0,53	0,58	1,26	0,81	0,37	0,73	0,56
RI (0,48)	104	184	98	251	0,51	0,58	1,22	0,84	0,36	0,72	0,56
PSV (58)	3	1	199	434	0,01	1,00	6,46	0,99	0,75	0,69	0,69
TAMXV (18,45)	23	14	179	421	0,11	0,97	3,54	0,92	0,62	0,70	0,70
VI (1,985)	105	72	97	363	0,52	0,83	3,14	0,58	0,59	0,79	0,73
FI (41,7)	21	21	181	414	0,10	0,95	2,15	0,94	0,50	0,70	0,68
VFI (2,25)	50	16	152	419	0,25	0,96	6,73	0,78	0,76	0,73	0,74
CA-125 (115)	98	13	62	206	0,61	0,94	10,32	0,41	0,88	0,77	0,80

Najwyższą wartość dokładności testu dla najlepiej dyskryminującej wartości progowej stwierdzono dla półilościowej oceny: cechy „Kolor" - 81%, markera CA-125 - 80% oraz dla centralnej lokalizacji unaczynienia - 79%. Kolejnymi parametrami były wiek i obustronność guza o dokładności 75%, status menopauzalny i wyrośla endofityczne - 74%. Dla indeksów VFI i VI wartości te wynosiły odpowiednio 74% i 73%. Uzyskiwane w badaniu z wykorzystaniem spektralnego Dopplera indeksy przepływu krwi PI i RI nie wykazały znaczącej przydatności diagnostycznej.

Obliczone wartości pola powierzchni pod krzywą ROC (AUROC) z błędem standardowym (SE) oraz 95% przedziałem ufności (95% CI) dla analizowanych parametrów przedstawia tabela 4.

Tabela 4. Porównanie wartości diagnostycznej ocenianych parametrów w oparciu o pole powierzchni pod krzywą ROC.

Parametr (optymalna wartość graniczna)	AUROC	SE	95% CI
WIEK	0,788	0,018	0,752 - 0,824
BMI	0,652	0,023	0,606 - 0,697
MNP	0,699	0,023	0,654 - 0,745
OBUSTR	0,683	0,024	0,636 - 0,731
IOTA	0,635	0,022	0,592 - 0,677
PAP	0,662	0,025	0,614 - 0,710
WYM	0,683	0,023	0,638 - 0,727
OBJ	0,714	0,022	0,671 - 0,758
VESS	0,788	0,020	0,748 - 0,828
KOLOR	0,831	0,020	0,793 - 0,870
PI	0,533	0,023	0,489 - 0,577
RI	0,522	0,023	0,478 - 0,567
PSV	0,589	0,023	0,544 - 0,634
TAMXV	0,661	0,022	0,617 - 0,705
VI	0,741	0,021	0,700 - 0,783
FI	0,651	0,022	0,607 - 0,695
VFI	0,741	0,021	0,700 - 0,783
CA-125	0,870	0,019	0,834 - 0,907

Najwyższe wartości pola powierzchni pod krzywą ROC w jednoczynnikowej analizie dla parametrów ultrasonograficznych uzyskano w przypadku: cechy „Kolor” (0,83) oraz indeksów przepływu w badaniu 3D, tj. VI i VFI (0,83). Dla stężenia antygenu CA-125 uzyskano wartość najwyższą wynoszącą 0,87.

Porównanie najwyższych wartości pól pod krzywymi ROC: CA-125 vs „Kolor” nie wykazało różnicy (p=0,15) [Smoleń 2008].

Podsumowanie wyników wieloczynnikowej analizy badanych parametrów w przedoperacyjnej diagnostyce różnicowej guzów przydatkowych

Najistotniejsze parametry w skonstruowanych modelach oraz uzyskaną dokładność dla zbiorów testowych w badanej grupie przedstawia tabela 5.

Tabela 5. Podsumowanie najistotniejszych parametrów i dokładności diagnostycznej dla zbiorów testowych w skonstruowanych modelach.

model	5 najistotniejszych parametrów	dokładność dla zbiorów testowych
AD	cecha „Kolor", lokalizacja naczyń, wiek, status menopauzalny, obustronność guza	91%
MLRA	stężenie CA-125, wyrośla endofityczne, cecha „Kolor", indeks PI, wiek	89%
MARS	wiek, indeks FI, stężenie CA-125, wyrośla endofityczne, cecha „Kolor"	75%
CART	cecha „Kolor", stężenie CA-125, morfologia ultrasonograficzna wg IOTA, lokalizacja naczyń, status menopauzalny	91%
CHAID	stężenie CA-125, lokalizacja naczyń, morfologia ultrasonograficzna wg IOTA, status menopauzalny	82%
BT	stężenie CA-125, wiek, morfologia ultrasonograficzna wg IOTA, cecha „Kolor", indeks VI	**93%**
RF	stężenie CA-125, morfologia ultrasonograficzna wg IOTA, cecha „Kolor", wiek, indeks VI	91%
MLP	morfologia ultrasonograficzna wg IOTA, objętość guza, wskaźnik masy ciała BMI, cecha „Kolor", lokalizacja naczyń	**93%**
RBF	morfologia ultrasonograficzna wg IOTA, cecha „Kolor", wyrośla endofityczne, największy wymiar guza, lokalizacja naczyń	**95%**
PNN	morfologia ultrasonograficzna wg IOTA, objętość guza, wskaźnik masy ciała BMI, lokalizacja naczyń, obustronność guza	**93%**
SVM	wszystkie badane parametry	88%
NB	stężenie CA-125, VFI, TAMXV, cecha „Kolor", lokalizacja naczyń	79%
KNN	wszystkie badane parametry	81%

Najistotniejszymi parametrami w skonstruowanych modelach okazały się: cecha „Kolor" (uwzględniona w 11 modelach), stężenie CA-125 (9), morfologia ultrasonograficzna wg IOTA (9), lokalizacja naczyń (9), wiek (7)/status menopauzalny (5), wyrośla endofityczne (5), objętość (4), obustronność (4), VI (4).

Najwyższymi wartościami predykcyjnymi dla klasyfikacji zbiorów testowych i zdolnością do uogólniania cechowały się modele sieci neuronowych i wzmacnianych drzew klasyfikacyjnych.

Najwyższe wartości pola powierzchni pod krzywą ROC w jednoczynnikowej analizie uzyskano dla stężenia antygenu CA-125 (0,87) i cechy „Kolor" (0,83). Podsumowanie wartości pól pod krzywymi ROC dla skonstruowanych modeli przedstawia tabela 6.

Tabela 6. Porównanie przydatności diagnostycznej skonstruowanych modeli w oparciu o wartość pola powierzchni pod krzywą ROC.

Model	AUROC	SE	95% CI
AD	0,929	0,013	0,903-0,955
MLRA	0,940	0,012	0,917-0,964
MARS	0,947	0,011	0,926-0,969
CART	0,934	0,013	0,908-0,959
CHAID	0,916	0,015	0,887-0,945
BT	0,974	0,007	0,961-0,986
RF	0,974	0,007	0,961-0,986
MLP	0,999	0,001	0,997-1
RBF	0,930	0,013	0,905-0,956
PNN	0,996	0,002	0,993-0,999
SVM z jądrem liniowym	0,861	0,021	0,821-0,902
SVM z jądrem RBF	0,891	0,018	0,855-0,927
NB	0,707	0,028	0,652-0,763
KNN	0,931	0,012	0,906-0,955

Wysokie wartości pól pod krzywymi ROC uzyskano dla modeli skonstruowanych przy użyciu regresji logistycznej, metody krzywych sklejanych MARS, sztucznych sieci neuronowych, wzmacnianych drzew klasyfikacyjnych i metody K-najbliższych sąsiadów. Wartości pól pod krzywą ROC różniły się istotnie dla cechy „Kolor" i dla stężenia antygenu CA-125 w porównaniu do tych modeli ($p<0,0001$). Najwyższymi wartościami pól pod krzywą ROC cechowały się sieci neuronowe typu MLP. Najniższymi natomiast modele skonstruowane z użyciem metody naiwnego klasyfikatora Bayesa. Ze względu na fakt, iż przy użyciu sieci neuronowych uzyskiwano najwyższe wartości dokładności diagnostycznej i pól pod krzywymi ROC, stworzono zespoły sieci, łącząc sieć MLP, RBF i PNN. Ponadto obliczono prognozy klasyfikacyjne, wykorzystując metodę kontaminacji najlepszych modeli (zasada głosowania k-sędziów). W celu porównania jako prognozę wybierano wynik „głosowania" trzech i pięciu najlepszych modeli oraz przedstawiono najlepszy uzyskany wynik sieci neuronowej MLP i zespołu sieci. Wartości prognostyczne obliczone dla sieci neuronowej (MLP), zespołu sieci oraz wyników „głosowania" modeli w zbiorze uczącym (U), walidacyjnym (W) i testowym (T) oraz ogółem dla grupy badanej (U+W) i ogółem w zbiorach testujących skuteczność (W+T) przedstawia tabela 7.

Tabela 7. Wartości prognostyczne dla najlepszego modelu sieci MLP, zespołu sieci i wyników „głosowania" najlepszych modeli.

Model		TP	FP	FN	TN	SENS	SPEC	LR+	LR-	PPV	NPV	ACC
sieć MLP	U	148	0	0	207	1,00	1,00	-	0,00	1,00	1,00	1,00
	W	12	0	0	12	1,00	1,00	-	0,00	1,00	1,00	1,00
	T	9	3	1	20	0,90	0,87	6,90	0,12	0,75	0,95	0,88
	U+W	160	0	0	219	1,00	1,00	-	0,00	1,00	1,00	**1,00**
	W+T	21	3	1	32	0,95	0,91	11,14	0,05	0,88	0,97	**0,93**
zespół sieci (MLP, RBF, PNN)	U	146	9	2	198	0,99	0,96	22,69	0,01	0,94	0,99	0,97
	W	12	0	0	12	1,00	1,00	-	0,00	1,00	1,00	1,00
	T	10	3	0	20	1,00	0,87	7,67	0,00	0,77	1,00	0,91
	U+W	158	9	2	210	0,99	0,96	24,03	0,01	0,95	0,99	**0,97**
	W+T	22	3	0	32	1,00	0,91	11,67	0,00	0,88	1,00	**0,95**
głosowanie (MLRA, BT, MLP)	U	138	9	10	198	0,93	0,96	21,45	0,07	0,94	0,95	0,95
	W	12	0	0	12	1,00	1,00	-	0,00	1,00	1,00	1,00
	T	10	2	0	21	1,00	0,91	11,50	0,00	0,83	1,00	0,94
	U+W	150	9	10	210	0,94	0,96	22,81	0,07	0,94	0,95	**0,95**
	W+T	22	2	0	33	1,00	0,94	17,50	0,00	0,92	1,00	**0,96**
głosowanie (MLRA, MARS, BT, MLP, KNN)	U	133	13	15	194	0,90	0,94	14,31	0,11	0,91	0,93	0,92
	W	12	0	0	12	1,00	1,00	-	0,00	1,00	1,00	1,00
	T	10	2	0	21	1,00	0,91	11,50	0,00	0,83	1,00	0,94
	U+W	145	13	15	206	0,91	0,94	15,27	0,10	0,92	0,93	**0,93**
	W+T	22	2	0	32	1,00	0,94	17,00	0,00	0,92	1,00	**0,96**

Najwyższą dokładność diagnostyczną uzyskano dla zespołu sieci oraz „głosowania" trzech modeli, tj. regresji logistycznej, wzmacnianych drzew klasyfikacyjnych i sieci MLP [Smoleń 2008].

Podsumowanie

Idea wykorzystania modeli statystycznych w celu szacowania indywidualnego prawdopodobieństwa istnienia raka jajnika w guzach przydatkowych wydaje się być bardzo atrakcyjna. Modele statystyczne mogłyby być alternatywą do metod rozpoznawania układu badanych cech guza wykrytych w badaniu klinicznym i ultrasonograficznym. Faktem jest, że uwzględnienie kombinacji cech klinicznych, ultrasonograficznych i stężenia antygenu CA-125 w skonstruowanych modelach, z użyciem metod sztucznej inteligencji, pozwala na uzyskanie lepszej dyskryminacji od otrzymywanej na podstawie pojedynczych parametrów

diagnostycznych w przedoperacyjnym różnicowania guzów złośliwych i niezłośliwych jajnika.

Należy podkreślić, że opisane modele mogą być gotowe do zastosowania w rutynowej praktyce klinicznej po spełnieniu pewnych warunków. Ze względu na subtelne różnice w definicjach, jakie mogą występować pomiędzy różnymi ośrodkami lub nawet poszczególnymi badającymi (którzy mogą np. używać różnego sprzętu usg) oraz różnicami populacji badanej i testowanej, przedstawione modele powinny zostać przetestowane w wieloośrodkowych badaniach na danych prospektywnych dostatecznie dużej grupy chorych, tak aby wykazać ich przydatność w innej populacji kobiet, przy zastosowaniu bardzo rygorystycznych i szczegółowych reguł postępowania. W przypadkach chorych z guzami jajnika należy upewnić się, że obejmują one całą gamę nowotworów przydatkowych o charakterze zarówno złośliwym, jak i niezłośliwym. Ponieważ modele statystyczne generalnie są bardzo przydatne, muszą być tworzone i testowane na podstawie bardzo dużej liczby nowotworów, zmienne w modelach muszą być jasno zdefiniowane, a techniki badawcze muszą zostać ujednolicone. Połączenie badanych parametrów w wielu modelach może prowadzić do przewidywania porównywalnego z subiektywną oceną doświadczonego lekarza ultrasonografisty.

Zaproponowane modele mogą stanowić prostą i niedrogą metodę, która mogłaby przyczynić się do wspomagania przedoperacyjnej diagnostyki różnicowej guzów jajnika, umożliwiając lepsze przedoperacyjne różnicowanie [Smoleń 2002, 2008].

Bibliografia

1. Bishop C. Neural Networks for Pattern Recognition. Oxford: University Press. 1995.
2. Demski T. Data mining II b – modele i metody. Cz. III. Wyd. StatSoft Kraków 2005.
3. Dobosz M. Wspomagana komputerowo statystyczna analiza wyników badań. Akademicka Oficyna Sikorski RJ [red.]. Wydawnicza Exit. Warszawa 2004.
4. Gore SM, Altman DG. Metody oceny - wiele zmiennych W: Statystyka w praktyce lekarskiej. Wyd. Naukowe PWN. Warszawa 1997: 95-99.
5. Hanley JA, McNeil B. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology. 1982; 143: 29-36.
6. Hanley JA, McNeil B. A method of comparing of the areas under the receiver operating characteristic curve derived from the same cases. Radiology 1983; 148: 839-843.
7. Hosmer DW, Lemeshow S. Applied logistic regression. New York. Wiley-Interscience 1989.
8. Jaeschke R, Cook D, Guyatt G. Ocena artykułów na temat testów diagnostycznych. Cz. II - metody określania przydatności testu Medycyna Praktyczna 1998; 11: 184-191.
9. Łapczyński M. Data mining II b – modele i metody. Cz. I. Wyd. StatSoft Kraków 2005.
10. Lula P. Data mining II b – modele i metody. Cz. II. Wyd. StatSoft Kraków 2005.
11. Pasztyła A. Data mining II b – modele i metody. Cz. IV. Wyd. StatSoft Kraków 2005.

12. Smoleń AB. Zastosowanie zaawansowanych metod modelowania statystycznego w ocenie prawdopodobieństwa istnienia raka jajnika u kobiet z guzami przydatkowymi: praca doktorska. Akademia Medyczna Lublin 2002.
13. Smoleń AB. Zastosowanie zaawansowanych metod eksploracji danych i wspomagania procesów decyzyjnych we wstępnej diagnostyce różnicowej guzów przydatkowych u kobiet: rozprawa habilitacyjna. Akademia Medyczna Lublin 2008.
14. Stanisz A. Przystępny kurs statystyki z wykorzystaniem programu *STATISTICA PL* na przykładach z medycyny t. II. Wyd. StatSoft Kraków 2000.
15. Stanisz A. Przystępny kurs statystyki z zastosowaniem *STATISTICA PL* na przykładach z medycyny t. II Modele liniowe i nieliniowe. Wyd. StatSoft Kraków 2007.
16. StatSoft Polska. Internetowy podręcznik statystyki (IPS): http://www.statsoft.pl/textbook/stathome.html.
17. Tadeusiewicz R. Sieci neuronowe. Akademicka Oficyna Wydawnicza RM. Warszawa 1993.
18. Tadeusiewicz R. Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami. Akademicka Oficyna Wydawnicza PLJ Warszawa 1999: 1-51.
19. Tadeusiewicz R. Data mining jako szansa na relatywnie tanie dokonywanie odkryć naukowych poprzez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych. W: Statystyka data mining w badaniach naukowych. StstSoft Polska. Kraków 2006.
20. Tadeusiewicz R, Lula P. Sieci neuronowe. Wyd. StatSoft Kraków 2001.