Uniwersytet Warszawski
Wydział Fizyki

Joanna Nowakowska
Nr albumu: 370486

Testowanie skuteczności algorytmu klasyfikującego obrazy medyczne wykorzystującego parametryzację za pomocą konwolucyjnej sieci neuronowej

Praca licencjacka
na kierunku Zastosowania Fizyki w Biologii i Medycynie
specjalność Fizyka Medyczna

Praca wykonana pod kierunkiem
dra Józefa Gintera
Zakład Fizyki Biomedycznej
Instytut Fizyki Doświadczalnej

Warszawa, <miesiąc-i-rok-złożenia-pracy>

Oświadczenie kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

<Krótkie (maks. 800 znaków) streszczenie pracy, na przykład:

Lorem ipsum – tekst składający się z łacińskich i quasi-łacińskich wyrazów, mający korzenie w klasycznej łacinie, wzorowany na fragmencie traktatu Cycerona „O granicach dobra i zła" (De finibus bonorum et malorum) napisanego w 45 r. p.n.e. Tekst jest stosowany do demonstracji krojów pisma (czcionek, fontów), kompozycji kolumny itp. Po raz pierwszy został użyty przez nieznanego drukarza w XVI w.

Tekst w obcym języku pozwala skoncentrować uwagę na wizualnych aspektach tekstu, a nie jego znaczeniu.

Cytat z `https://pl.wikipedia.org/wiki/Lorem_ipsum` >

Słowa kluczowe

czerniak, SVM, klasyfikator, sieć neuronowa

Dziedzina pracy (kody wg programu Socrates-Erasmus)

13.2 Fizyka

Tytuł pracy w języku angielskim

<Tytuł pracy w tłumaczeniu na język angielski>

Spis treści

Cel pracy

. . . blablablabbjj[**?**]blblb

Rozdział 1

Wstęp

Nowoczesne technologie oraz coraz szybsze komputery są obecnie kluczowymi czynnikami napędzającymi rozwój niemal wszystkich gałęzi przemysłu. Jedną z najważniejszych branży, która dzięki temu może się rozwijać jest medycyna. Wprowadzane na bieżąco medyczne innowacje pozwalają ratować życie coraz większej liczbie ludzi na całym świecie.

Jednym z najbardziej dynamicznie rozwijających się obecnie obszarów informatyki jest sztuczna inteligencja, w tym jedna z jej gałęzi - tzw. sieci neuronowe (ang. *neural networks*). Służą one głównie do klasyfikacji i ropoznawania obiektów oraz analizy danych. Posiadają wiele zastosowań, m.in w technice – do analizy obrazów i przetwarzania sygnałów, czy w ekonomii do tworzenia prognoz i optymalizacji decyzji gospodarczych.

Dzięki ich zdolnościom do ciagłego uczenia się, sieci wykorzystuje się do przewidywania wyników na podstawie zgromadzonych danych wejściowych. Z tego powodu konwolucyjne sieci neuronowe (ang. *convolutional neural networks*), jeden z typów sieci neuronowych (o których będzie mowa w kolejnych rozdziałach), wykazują duży potencjał w obszarze diagnostyki medycznej. By wspomóc lekarzy, systemy te mogą przeprowadzać dogłebną analizę szerokiego zakresu danych na temat danego pacjenta i na tej podstawie dokonywać decyzji na temat jego stanu zdrowia. Takie zastosowanie sieci neuronowych odciążyłoby lekarzy i, przy odpowiednio dobranych sposobach klasyfikacji, zwiększyło liczbę prawidłowo postawionych diagnoz.

Jednym z obszarów diagnostyki medycznej, w którym wciąż w dużej mierze diagnozę opiera się na subiektywnej ocenie lekarza jest dermatologia. Przy coraz większej liczbie ludności mającej problemy ze zmianami skórnymi, szybka i odpowiednia diagnoza staje się kluczowym elementem procesu leczenia. Szczególnie niebezpieczne są czerniaki – złośliwe nowotwory skóry, których częstość występowania bardzo wzrosła w przeciągu ostatnich 30 lat[PRZYPIS]. Są obecnie najczęstszą przyczyną zgonów spowodowanych nowotworami złośliwymi skóry. Z tego powodu w ostatnich latach wzrosło zainteresowanie badaniami mającymi na celu opracowanie algorytmów wspomagających rozpoznawanie czerniaków. Między innymi w 2017 roku na Międzynarodowym Sympozjum Obrazowania Biomedycznego (ISBI – ang. *International Symposium on Biomedical Imaging*) zorganizowano konkurs przy współpracy z *International Skin Imaging Collaboration* (ISIC), którego celem była prezentacja algorytmów klasyfikujących wyżej opisane zmiany skórne. Jako danych wejściowych użyto zestawu zdjęć zmian barwnikowych łagodnych oraz złośliwych udostępnionych przez stronę ISIC.

W pracy użyto tego samego zbioru zdjęć w celu przetestowania różnych algorytmów klasyfikujących obrazy medyczne.

1.1. Sztuczne sieci neuronowe

Sztuczne sieci neuronowe (ang. *Artifical Neural Networks*, ANNs) są to pewne struktury matematyczne wzorujące się na działaniu neuronów biologicznych. Naśladując zdolności ludzkiego umysłu, mają one na celu naukę podejmowania odpowiednich decyzji, klasyfikacji wzorów czy formułowania prognoz na podstawie wcześniej zgromadzonych danych i dokonania analizy retrospektywnej. Przy konstruowaniu sieci należy na początku zadać pewne ograniczenia(?) związane m.in z architekturą sieci (typ, funkcja aktywacji czy liczba neutronów w warstwach), sposobem uczenia sieci czy metody analizy i weryfikacji otrzymanych wyników.

Sieć neuronowa składa się z pojedynczych neuronów pogrupowanych w warstwy. Sieć można podzielić na trzy główne części: warstwę wejściową, warstwy ukryte oraz warstwę wyjściową. Warstwa wejściowa wprowadza dane do sieci. Ostatnia warstwa, warstwa wyjściowa, służy do wyznaczania wartości wyjściowych sieci. Pomiędzy pierwszą i ostatnią warstwą znajdują się warstwy ukryte. Neurony znajdujące się w warstwach ukrytych mają za zadanie przetwarzać informacje wejściowe poprzez kolejne etapy uczenia na informacje wyjściowe. Neurony , które znajdują się w sąsiadujących warstwach są ze sobą połączone. Poprzez te połączenia następuje przesył informacji przez sieć. Ogólną budowę sztucznej sieci neuronowej przedstawiono na Rysunku []. [ambroch]

Pojedynczy neuron jest przetwornikiem sygnału - na wejściu dostaje wektor x danych wejściowych (sygnałów) pochodzących z warstwy wejściowej sieci lub z neuronów warstwy poprzedniej. Każdemu połączeniu przypisywana jest waga w_i dobierana w procesie uczenia. W neuronie sumowane są dane wejściowe i odpowiadające im wagi. Suma iloczynów daje sumaryczne wejście s neuronu:

$$s = \sum_{i=0}^{n} x_i * w_i = w^T x, \tag{1.1}$$

gdzie x_i - kolejne sygnały wejściowe, w_i - odpowiadające im wagi, n - liczba cech (sygnałów) na wejściu. Wartość s jest argumentem twz. funkcji aktywacji a według której obliczana jest wartość wyjściowa neuronu, określonej wzorem

$$a = f(s). \tag{1.2}$$

Stosuje się różne funkcje aktywacji, m.in.:
• funkcję progową:

$$f(s,p) = \{\, 0 \quad \text{dla} \quad s<p1 \quad \text{dla} \quad s>p, \tag{1.3}$$

gdzie p - wartość progowa.
• funkcję logistyczną:

$$f(s) = \frac{1}{1 + exp^{-s}}, \tag{1.4}$$

• tangens hiperboliczny:

$$f(s) = \frac{exp^{s} - exp^{-s}}{exp^{s} + exp^{-s}}. \tag{1.5}$$

1.2. Algorytm SVM (ang. *Support Vector Machine*)

Maszyny wektorów nośnych (ang. *Support Vector Machine* - SVM) to jeden z najczęściej stosowanych algorytmów w nadzorowanym uczeniu maszynowym jako klasyfikator. W modelu SVM każdy element badanego zbioru danych traktowany jest jako punkt w n-wymiarowej

przestrzeni, gdzie n to liczba parametrów opisująca dany element. Zadaniem klasyfikatora jest znalezienie hiperpłaszczyny, która 'najlepiej' rozseparuje dane należące do różnych klas. Proces uczenia się algorytmu to proces dobierania wag, w którym maksymalizowany jest margines separacji oddzielający skrajne punkty obu klas (wektory nośne) leżące najbliżej wyznaczanej hiperpłaszczyzny. Ideę działania algorytmu przedstawiono na Rysunku [].

Klasyfikator SVM przyjmuje na wejściu parę zmiennych (x,y), gdzie x to n-wymiarowy wektor cech (parametrów), a y to liczba wskazująca prawdziwą przynależność zmiennej do wybranej klasy (dla wygody poźniejszych obliczeń, -1 lub +1). Przy założeniu o liniowej separalności klas, równanie hiperpłaszczyzny możemy zapisać w postaci

$$h(x) = w^T x + b = 0, \tag{1.6}$$

gdzie w - wektor wag $(w_1, w_2, ..., w_n)$, x - wektor danych wejściowych, b - polaryzacja. W takim wypadku równania przynależnośći dla obydwu klas wyglądają następująco:

$$\text{jeśli} \quad w^T x + b \geqslant 0, \quad \text{to} \quad y_i = +1, \tag{1.7}$$

$$\text{jeśli} \quad w^T x + b \leqslant 0, \quad \text{to} \quad y_i = -1, \tag{1.8}$$

co można uogólnić do postaci:

$$y_i(w^T x + b) \geqslant 1. \tag{1.9}$$

Spełnienie równania 1.9 przez pary punktów (x_i, y_i) definiuje wektory nośne decydujące o położeniu hiperpłaszczyzny i wielkości marginesu separacji. Odległość r(xsv) wektorów nośnych x_{SV} od hiperpłaszczyzny określona jest wzorem:

$$r(x_{SV}) = \frac{y_i(x_{SV})}{||w||} = \frac{\pm 1}{||w||}, \tag{1.10}$$

gdzie $||w||$ to normalizacja wektora w. Szerokość d marginesu separacji możemy w takim wypadku określić jako podwojoną bezwzględną odległość wektorów nośnych od hiperpłaszczyzny:

$$d = 2 * ||r(x_{SV}|| = \frac{2}{||w||}. \tag{1.11}$$

Pragnąc uzyskać zmaksymalizwany margines separacji d trzeba zminimalizować $||w||$, co jest równoważne z minimalizacją wyrażenia $\frac{1}{2}||w||^2$. Podana funkcja jest funkcją wypukłą, a więc mającą jedno minimum globalne. Istnieje wiele gotowych bibliotek numerycznych przeprowadzających bardzo szybko wspomnianą minimalizację. Jednak do rozwiązania podanego wyrażenia często stosuje się tzw. mnożniki Lagrange'a, ułatwiające rozwiązanie problemu przy bardzo wielowymiarowych wektorach cech. Więcej o tej metodzie w [Przypiss].
Trzeba jednak pamiętać, że bardzo często pracuje się na początkowo nieseparowalnym liniowo zbiorze danych. W takim wypadku można zastosować pewne przekształcenia (tzw. mapowanie) pozwalające na przejście do większowymiarowej przestrzeni, w której dane stają się liniowo separowalne. Wprowadza się tzw. funkcję mapującą $\phi(x)$, która przenosi punkty z oryginalnej przestrzeni wejściowej do 'powiekszonej' przestrzeni cech: (w algorytmie uczącym zamienia się x na $\phi(x)$).

$$\begin{gathered} \phi : R^n \to R^{n'} \\ x \to \phi(x) \\ (x_1, x_2, \ldots, x_n) \to (z_1, z_2, \ldots, z_{n'}) \end{gathered}$$

Dla każdego mapowania możemy zdefiniować tzw. jądro (ang. *kernel*) K:

$$K(x,z) = \phi(x)^T \phi(z) \tag{1.12}$$

Wtedy za odwzorowanie danych w nowej przestrzeni odpowiada funkcja kernelowa, mogąca przyjmować różne postaci. Do najważniejszych funkcji kernelowych należą:
• Kernel RBF (ang. *Radial Basis Function*):

$$K(x,z) = e^{-\frac{||x-z||^2}{2\sigma^2}} \tag{1.13}$$

• Kernel wielomianowy:

$$K(x,z) = (\langle x,z\rangle + 1)^p \tag{1.14}$$

• Kernel sigmoidalny:

$$K(x,z) = \text{tgh}(\langle x,z\rangle + 1) \tag{1.15}$$

gdzie p to parametr jądrowy. W pracy zastosowano kernel RBF, którego używa się najczęściej.
Algorytm SVM domyślnie zwraca wartości binarne. Jednak do niektórych analiz statystycznych, również w niniejszej pracy, potrzebne jest otrzymanie prawdopodobieństw przynależności danego elementu do każdej z klas (zamiast wartości zero-jedynkowych). W tym celu stosuje się pewne algorytmy, w tym tzw. skalowanie Platta (ang. *Platt scaling*), zwracające prawdopodobieństwo przynależności zmiennej do klasy 1, określone wzorem:

$$P(y=1|f(x)) = \frac{1}{1+exp(Af(x)+B)}, \tag{1.16}$$

gdzie $f(x)$ - ....., A i B - parametry wyznaczone na drodze uczenia przy użyciu ("maximum likelihood estimation") na zbiorze uczącym ($f(x_i, y_i)$. [prypiss - art.pap]

1.3. Klasyfikacja danych i ocena jakości klasyfikatora

Proces klasyfikacji danych podczas uczenia maszynowego składa się zazwyczaj z 3 etapów. Pierwszym jest konstrukcja modelu najlepiej odpowiadającego zbiorowi danych wejściowych (danym treningowym), a następnie zastosowanie uzyskanego dopasowania do klasyfikacji nowych danych (danych testowych). Na koniec dokonuje się weryfikacji "jakości" wybranego modelu, czyli bada się jego parametry takie jak trafność, szybkość czy stopień generalizacji. Aby ocenić poprawność otrzymanego klasyfikatora niezbędna jest znajomość prawdziwej przynależności danych do badanych klas i porównanie jej z przyporządkowaniem do klas zaproponowanych w procesie klasyfikacji. Oczywiste wydaje się być to, że im większą ilością danych dysponujemy oraz gdy zbiory treningowe i testowe pochodzą z różnych "źródeł" (różnych serii badań), tym otrzymamy lepszy klasyfikator. Niestety w praktyce często pracuje się na pojedynczym zbiorze danych, na dodatek o małej liczebności. W takim wypadku stosuje się różne techniki podziału zbioru na elementy treningowe i testowe. Jedną z nich jest zastosowana w poniższej pracy tzw. metoda K-krotnej walidacji (ang. *K-fold cross–validation*).

Metoda K-krotnej walidacji polega na podziale dysponowanego zbioru danych na K podzbiorów. Następnie kolejno jeden pozdbiór przyjmuje się jako zbiór testowy, a wszystkie pozostałe jako zbiór uczący, na którym trenuje się klasyfikator. Proces klasyfikacji zachodzi więc niezależnie K razy. Na koniec uśrednia się otrzymane wyniki z K przejść w celu otrzymania końcowego rezultatu. Zaletą powyższej metody jest wykorzystanie w trakcie uczenia całego kompletu informacji. Jednak z drugiej strony użycie tych samych danych na etapie trenowania i testowania może skutkować nieobiektywną oceną skuteczności klasyfikatora.

Przy binarnej ocenie jakości klasyfikacji (podział na dwie klasy) często stosuje się tzw. macierz błędów (ang. *confusion matrix*). Każdemu elementowi zbioru przypisuje się dwie etykiety: jedną zgodną ze stanem faktycznym oraz jedną uzyskaną w procesie klasyfikacji. Wszystkie możliwe kombinacje połączeń obu etykiet przedstawia ww. macierz błędów (Tabela 1.1). Zazwyczaj zamiast etykiet zero-jedynkowych stosuje się podział na klasę pozytywną oraz negatywną.

Gdy przydział do danej klasy w procesie klasyfikacji zgodził się ze stanem faktycznym,

Tabela 1.1: macierz błedów

	klasa prawdziwa pozytywna	klasa prawdziwa negatywna
klasa predykowana pozytywna	prawdziwie pozytywna TP	fałszywie pozytywna FP
klasa predykowana negatywna	fałszywie negatywna FN	prawdziwie negatywna TN

przypisujemy dany element albo do zbioru „prawdziwie pozytywnych" rozpoznań (ang. *true positive*, TP) albo do „prawdziwie negatywnych" (ang. *true negative*, TN). W przeciwnym razie, gdy przypisane etykiety nie są zgodne, przypisujemy danym wartość „fałszywie negatywną" (ang. *false negative*, FN) lub „fałszywie pozytywną" (ang. *false positive*, FP). Zliczenia wszystkich przypadków są podstawą do wyznaczenia różnych wartości diagnostycznych testu pozwalających na ocenę klasyfikatora. Podstawowym parametrem wykorzystywanym przede wszystkim w diagnostyce medycznej (ale nie tylko) jest tzw. czułość (ang. *sensitivity*). Można się również spotkać z innymi angielskimi nazwami takimi jak *true positive ratio* (TPR) czy *recall*. Mówiąc o przypadkach medycznych, określa ona jaką część chorych prawidłowo zakwalifikowano do chorych, czyli inaczej, prawdobodobieństwo prawidłowej klasyfikacji osoby chorej. Z tego powodu klasyfikatory mające pomóc rozpoznaniom medycznym powinny dążyć do jak największej wartości tego parametru. Na podstawie macierzy błędów, czułość określona jest wzorem:

$$TPR = \frac{TP}{TP + FN} \tag{1.17}$$

„Dopełnieniem" czułości jest specyficzność (ang. *specificity* (SPC), *true negative ratio*). Określa ona jaką część osób zdrowych zakwalifikowano poprawnie, czyli prawdopodobieństwo, że test pokaże wynik negatywny dla osoby zdrowej. Obliczana jest według wzoru:

$$SPC = \frac{TN}{FP + TN} \tag{1.18}$$

Ważną miarą jest również określenie ilości „fałszywych alarmów" czyli tzw. FPR (ang. *false positive ratio*). Wskazuje ono prawdopodobieństwo zakwalifikowania osoby zdrowej jako

chorej. Określana jest wzorem:

$$FPR = \frac{FP}{FP + TN} \tag{1.19}$$

Innymi parametrami badanymi w pracy były również precyzja pozytywna PPV (ang. *positive predictive value*) mówiąca o procencie liczby osób chorych wsród wszystkich kwalifikacji pozytywnych, dokładność ACC (ang. *accuracy*) określająca całkowitą dokładność klasyfikacji oraz parametr F1, czyli średnia harmoniczna z precyzji i czułości. wymienione wielkości obliczane są kolejno według wzorów:

$$PPV = \frac{TP}{TP + FP} \tag{1.20}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{1.21}$$

$$F_1 = 2\frac{PPV * TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN} \tag{1.22}$$

Wyżej wymienione parametry stosuje się w przypadku klasyfikacji binarnych. Większe możliwości mogą dawać klasyfikatory zwracające wartości z ciagłego rozkładu prawdopodobieństwa. Istnieje wtedy możliwość „ręcznego" ustalenia progu z przedziału [0,1], powyżej którego odpowiedź klasyfikatora traktuje się jako pozytywną, a poniżej jako negatywną. W takim wypadku ocenę jakości modelu przeprowadza się za pomocą tzw. krzywej ROC. Kolejne punkty na tej krzywej otrzymywane są dla ustalonej wartości progu. Na osi X znajdują się wartości 1-SPC, czyli parametr false positive ratio FPR, natomiast na osi Y wartości czułości TPR. Idealna krzywa zawierałaby punkt (0,1) oznaczający, że wszystkie dane zakwalifikowano prawidłowo. Parametr AUC (ang. *Area Under Curve*), definiowany jako pole powierzchni pod krzywą, wskazuje moc diagnostyczną testu.

W celu zbadania jakości klasyfikatora można również porównać go z innymi, prostymi klasyfikatorami, np. klasyfikatorem losowym – gdzie ostateczne wyniki są przypisywane losowo na podstawie przyjętego rozkładu prawdopodobieństwa lub klasyfikatorem zerowym/jedynkowym, kiedy predykcja zawsze wynosi odpowiednio 0 lub 1.

1.4. Czerniak - różnicowanie zmian barwnikowych w praktyce klinicznej

Czerniak jest obecnie jednym z najgroźniejszych zmian nowotworowych skóry. Wywodzi się z komórek barwnikowych skóry – melanocytów. Zdecydowana większość przypadków dotyczy tkanki skórnej, jednak w związku z występowaniem komórek barwnikowych w innych tkankach może się on pojawiać np. wewnątrz gałki ocznej, jamie ustnej czy jelitach. Charaketryzuje się wysokim stopniem złośliwości i częstymi przerzutami nawet we wczesnych stadiach choroby. Przyczyny rozwoju wciąż nie są do końca wyjaśnione. Wiadomo, że znaczenie mają zarówno czynniki genetyczne (około 10% zachorowań), jak i środowiskowe. Za najważniejszy czynnik uważa się nadmierną ekspozycję na promieniowanie UV, która ma mutagenny wpływ na DNA. Czerniak jest nowotworem wyleczalnym, jednak zasadnicze znaczenie ma jak najwcześniejsza diagnoza, ponieważ dzięki niej możliwe jest wycięcie zmiany przed rozwinięciem i przerzutami do inncyh organów. Przeżywalność powyżej 5 lat od rozpoznania spada z powyżej 90% na początkowych etapach nowotworzenia do 14% w ostatnich stadiach[przypis-artykul papierowy]. Wstępne rozpoznanie i różnicowanie zmian barwnikowych skóry przeprowadza dermatolog. Zazwyczaj używa się tzw. dermatoskopu, który pozwala na obejrzenie zmian z

kilkunastokrotnym powiększeniem. Lekarz stawia wstepną diagnozę na podstawie oceny różnych parametrów zmiany, takich jak:

- asymetria zmiany,
- nierówne brzegi (postrzępione, nieodgraniczone od skóry),
- niejednolity kolor- obecność kilku kolorów (głównie czerwonego i niebieskiego),
- duży rozmiar (średnica > 6 mm),
- ewolucja zmiany w krótkim czasie,
- niebiesko-biały welon,
- nieregularne kropki/globule,
- nieregularne plamy i pasma.

Po przeprowadzeniu badania dermatoskopowego i oceny według przyjętej skali (zazwyczaj reguła ABCD lub jedna z innych stosowanych metod [przypisss]), lekarz dokonuje subiektywnej oceny i w przypadku pozytywnej diagnozy zleca wycięcie zmiany. Ostateczne rozpoznanie następuje dopiero po badaniu histopatologicznym wyciętej zmiany.

Podawana trafność stawianych przez lekarzy prawidłowych diagnoz różni się w zależności od źródła. Najczęściej jednak nie przekraczają one 65%[przypiss], co pozostawia duże pole do poprawy w zakresie diagnostyki czerniaka oraz pokazuje potrzeby szukania różnych algorytmów wspomagających.

Rozdział 2

Dane eksperymentalne

Danymi eksperymentalnymi w niniejszej pracy był zbiór 1279 zdjęć zmian barwnikowych skóry, podzielony na zdjęcia zmian łagodnych (1031 zdjęć) oraz złośliwych (248 zdjęć). Obrazy zostały pobrane ze strony *International Skin Imaging Collaboration*[przypis] - organizacji posiadającej jedną z największych na świecie bazy zdjęć dermatologicznych, które sprawiają duże trudności diagnostyczne. Przykładowe zdjęcia pobrane z bazy pokazano na Rysunku[rysunekk].

Stosunek liczby zdjęć złośliwych i niezłośliwych dobrze odzwierciedla rzeczywistość, gdzie w większości przypadków mamy do czynienia ze zmianami łagodnymi, a tylko nieznaczna część badanych zmian skórnych to pozytywne przypadki choroby. Zgodnie z pracą [przypiss], większość eksperymentów jest przeprowadzana właśnie na niezrównoważonych danych, które dokładniej odzwierciedlają stan faktyczny, kiedy zawyczaj więcej badanych osób jest zdrowych niż chorych. Jednakże, w odniesieniu do pracy z konwolucyjnymi sieciami neuronowymi, postanowiono sprawdzić czy zrównanie liczebności zbiorów, zarówno w przypadku danych treningowych i testowych, zwiększy efektywność nauki sieci oraz poprawi wyniki końcowe.

W pracy użyto parametrów uzyskanych w wyniku przepuszczenia obrazów przez wytrenowaną wcześniej sieć konwolucyjną udostępnioną przez *Berkeley Vision and Learning Center* (BVLC). Model sieci wykorzystany do uzyskania potrzebnych danych nazywa się VGG-CNN-S i został stworzony w ramach *Image Large Scale Visual Recognition Challenge* (ILSVRC) 2012 - konkursu w ramach którego bada się najlepsze algorytmy do klasyfikacji obrazów.Sieć została wcześniej wytrenowana na bazie danych z ImageNet[przypis]. Pełną budowę sięci można zobaczyć w [przypiss].

Rozdział 3

Metodologia

3.1. Oprogramowanie

W celu analizy danych oraz przedstawienia wyników końcowych napisano skrypty w języku Python 3.4. Skorzystano w wielu wbudowanych bibiotek: NumPy, Matplotlib, Pandas i przede wszystkim Scikit-Learn. Scikit-Learn to biblioteka używana do uczenia maszynowego zawierająca różne algorytmy, m.in. do klasyfikacji, regresji czy analizy skupień.

3.2. Metodyka pracy

Do analizy w ramach niniejszej pracy licencjackiej posłużono się wektorem 4096. parametrów uzyskanych w procesie parametryzacji obrazów. Przycięte do kwadratu zdjęcia przepuszczono przez wyżej opisaną sieć VGG-CNN-S. Z sieci wyodrębniono warstwę szóstą w pełni połączoną (FC6) o długości 4096. W ten sposób uzyskano wektor 4096 parametrów opisujących dany obraz. Powyższa operacja została przeprowadzona w ramach pracy licencjackiej Urszuli Romaniuk[prypis].
Danymi wejściowymi przeprowadzonej analizy była macierz X o wymiarach mxn, gdzie m to liczba wierszy odpowiadająca ilości zdjęć użytych w pracy (1279), a n to liczba kolumn równa długości wektora parametrów opisującego pojedynczy obraz (4096). Przed procesem uczenia przeskalowano wektor cech za pomocą funkcji sigmoidalnej

$$x_j^{'} = \frac{1}{1 + exp^{-x_j}} \quad (\text{dla} \quad j = 1, \ldots, n)\,, \tag{3.1}$$

gdzie $x_j^{'}$ - przeskalowana wartość j-tej cechy, x_j - oryginalna wartość j-tej cechy, a n to długość wektora cech (4096). Dodatkowo jednowymiarowa macierz Y o długości równej liczbie zdjęć (1279) zawierała kolejno zera i jedynki, wskazując na przynależność każdego zdjęcia do jednej z dwóch klas ("0" - klasa negatywna, zmiany niezłośliwe; "1" - klasa pozytywna, zmiany złośliwe). Zbiór danych podzielono na obrazy testowe i treningowe kolejno na 3 sposoby: najpierw trenowano i testowano klasyfikatory na zbiorach zrównoważonych, następnie na zbiorach niezrównoważonych (w stosunku 4:1 zmian łagodnych do złośliwych), a na koniec trenowano na zbiorze zrównoważonym a testowano na niezrównoważonym. Tworzono instację danego klasyfikatora, dopasowywano model do danych treningowych, a następnie sprawdzano go na danych testowych. Za pomocą funkcji *predict_proba* otrzymywano prawdopodobieństwo przynależności zmiennej do danej klasy. Zmieniano próg prawdopodobieństwa z zakresu (0,1) poniżej którego obraz klasyfikowano jako "0", a powyżej jako "1". Za każdym razem obliczano parametry wskazujące jakości danego klasyfikatora (rozdziałprzypiss). Następnie wybierano

jeden z parametrów, sprawdzano dla jakiego progu dał on najlepsze wyniki i jakie wartości wskazują w tej sytuacji pozostałe parametry. Za każdym razem wyniki porównywano z klasyfikatorem losowym.

Rozdział 4

Wyniki

Zbiór danych zawierał 1279 zdjęć zmian barwnikowych skóry, w tym 1031 zdjęć zmian łagodnych o przypisanej wartości binarnej 0 i 248 zdjęć zmian złośliwych o przypisanej wartości 1. Za pomocą walidacji krzyżowej tworzono dwa podzbiory — treningowy *TR* i testowy *TE*. W zależności od analizowanego przypadku zbiory były albo zrównoważone (zawierały taką samą ilość zmian łagodnych co złośliwych) albo niezrównoważone (stosunek ilości zmian złośliwych do niezłośliwych wynosił 1:4). Sprawdzano wyniki dla czterech różnych przypadków:
• gdy obydwa zbiory *TR* i *TE* były zrównoważone,
• gdy obydwa zbiory *TR* i *TE* były niezrównoważone,
• gdy zbiór *TR* był zrównoważony, a *TE* niezrównoważony,
• gdy zbiór *TR* był niezrównoważony, a *TE* zrównoważony.
Za każdym razem wyniki były porównywane z klasyfikatorem losowym.
Rozdział został podzielony na[] podrozdziały, każdy mówiący o innym typie klasyfikatora. Dla każdego klasyfikatora analizowano wyniki dla czterech wymienionych wyżej przypadków.

4.1. Klasyfikator SVM

4.1.1. Zbiory *TR* i *TE* zrównoważone

Ze względu na dużą dysproporcję liczby zdjęć zmian łagodnych i złośliwych, w celu uzyskania zbiorów zrównoważonych potrzebne było obcięcie ilości zdjęć o wartości binarnej 0. Stosując walidację krzyżową liczbę jedynek (248) podzielono na 5 jak najbardziej równolicznych podzbiorów i do każdego dodano dokładnie taką samą liczbę zer. Za każdym razem jeden z tak stworzonych pozdbiorów pełnił funkcję zbioru testowego *TE*, a pozostałe cztery rolę zbioru *TR*. Na początek zmieniając wysokość j progu w zakresie [0,1] wybierano taką wartość j_{best} dla którego parametr *TPR* był jak najbliższy wartości 0,95. Następnie obliczano pozostałe parametry dla opisanego przypadku. Sprawdzano również wszystkie parametry ze względu na wartość C (odpowiadającej generalizacji modelu) w zakresie od 1 do 464 (wartość zwiększano logarytmicznie) i wybierano wartość C, dla której model dawał najlepsze wyniki. Uzyskany model porównywano z klasyfikatorem losowym. Wyniki przedstawiono w Tabeli 4.1.

Tabela 4.1: Wyniki dla zbiorów *TR* i *TE* zrównoważonych dla parametru $C = 46{,}4$ i ustalonym TPR = 0,95

	Klasyfikator SVM	Klasyfikator losowy
Próg j_{best}	0,21	0,11
TPR	0,952 ± 0,020	0,952 ± 0,045
F1	0,704 ± 0,009	0,664 ± 0,029
ACC	0,601 ± 0,013	0,508 ± 0,028
SPC	0,559 ± 0,008	0,509 ± 0,015

Następnie sprawdzono dla którego progu i dla której wartości C wartość parametru F1 była największa i jakie wartości przyjmowały w takim wypadku pozostałe parametry. Wyniki przedstawiono w Tabeli 4.2.

Tabela 4.2: Wyniki dla zbiorów *TR* i *TE* zrównoważonych dla parametru $C = 100$ i maksymalizacji F1

	Klasyfikator SVM	Klasyfikator losowy
Próg j_best	0,37	0,03
TPR	0,855 ± 0,050	1,000 ± 0
F1	0,738 ± 0,037	0,728 ± 0,002
ACC	0,696 ± 0,045	0,573 ± 0,004
SPC	0,649 ± 0,035	0,572 ± 0,002

Analogicznie postąpiono maksymalizując pozostałe dwa parametry - trafność ACC oraz specyficzność SPC. Otrzymane wyniki przedstawiono kolejno w Tabelach 4.3 i 4.4.

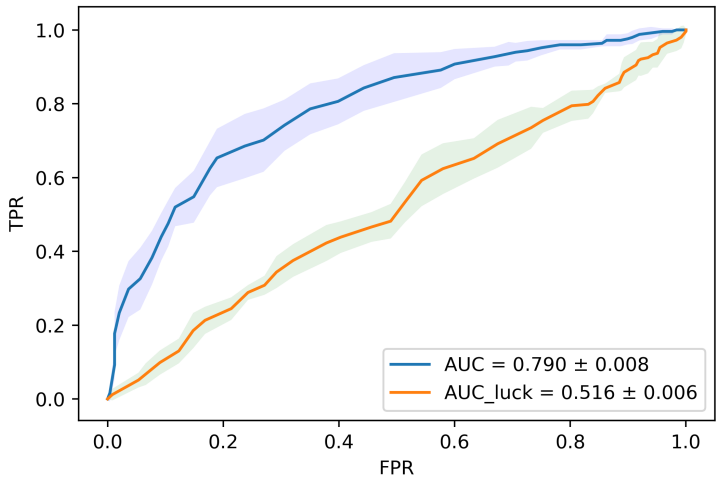
Tabela 4.3: Wyniki dla zbiorów *TR* i *TE* zrównoważonych dla parametru $C = 21{,}5$ i maksymalizacji trafności ACC

	Klasyfikator SVM	Klasyfikator losowy
Próg j_best	0,55	0,07
TPR	0,653 ± 0,078	0,996± 0,007
F1	0,704 ± 0,080	0,727 ± 0,005
ACC	0,726 ± 0,076	0,573 ± 0,009
SPC	0,767 ± 0,090	0,572 ± 0,005

Tabela 4.4: Wyniki dla zbiorów *TR* i *TE* zrównoważonych dla parametru $C = 46{,}4$ i maksymalizacji precyzji SPC

	Klasyfikator SVM	Klasyfikator losowy
Próg j_best	0,82	0,76
TPR	0,217± 0,075	0,183± 0,047
F1	0,347± 0,106	0,278± 0,056
ACC	0,603± 0,036	0,464 ± 0,015
SPC	0,957± 0,057	0,597 ± 0,036

Na Rysunku 4.1.1 przedstawiono krzywe ROC dla zastosowanego klasyfikatora i klasyfikatora losowego dla najlepszej wartości parametru C otrzymanej dla czułości TPR = 0,95.



Pole AUC pod krzywą jest ponad 50% większe od pola AUC_{luck} dla klasyfikatora losowego, co świadczy o dużej mocy diagnostycznej otrzymanego klasyfikatora.
Analizując powyższe Tabele można zauważyć, że maksymalizacja parametru F1 daje najlepsze wyniki - zachowana jest duża czułość TPR na poziomie 85%. Na dodatek próg 0,37 zapewnia wysoką trafność wynosząca ok. 70%. W przypadku ustawienia bardzo wysokiej czułości 95% próg j spada do wartości 0,21 co powoduje obniżenie trafności. Natomiast maksymalizacja trafności nie podnosi znacząco wyników w porównaniu do maksymalizacji F1, a jedynie pogarsza czułość o prawie 25%. Zdecydowanie najgorsze wyniki powoduje maksymalizacja precyzji.

4.1.2. Zbiory *TR* i *TE* niezrównoważone

W tej części analizy zarówno zbiór *TE* jak i *TR* były zbiorami niezrównoważonymi o stosunku zer i jedynek 4:1. Wszystkie kroki obliczeń powtórzono tak jak w podrozdziale []. Wyniki dla TPR = 0,95, maksymalizacji F1, ACC i SPC przedstawiono kolejno w Tabelach 4.5, 4.6, (()).

Tabela 4.5: Wyniki dla zbiorów *TR* i *TE* niezrównoważonych dla parametru $C = 10$ i ustalonym TPR = 0,95

	Klasyfikator SVM	Klasyfikator losowy
Próg j_best	0,07	0,03
TPR	0,944±0,030	0,956±0,032
F1	0,412± 0,012	0,333±0,008
ACC	0,460± 0,022	0,235±0,009
SPC	0,263± 0,009	0,202±0,004

Tabela 4.6: Wyniki dla zbiorów *TR* i *TE* niezrównoważonych dla parametru $C = 100$ i maksymalizacji F1

	Klasyfikator SVM	Klasyfikator losowy
Próg j_best	0,15	0,01
TPR	0,814±0,074	0,992±0,016
F1	0,709±0,042	0,333±0,004
ACC	0,867±0,018	0,206±0,005
SPC	0,630±0,033	0,200±0,002

4.1.3. Zbiór *TR* zrównoważony i *TE* niezrównoważony

W tej części przetestowano algorytmy gdy klasyfikator uczony jest na zbiorze *TR* zrównoważonym i testowany na zbiorze *TE* niezrównoważonym (stosunek zer i jedynek 4:1). Wyniki przedstawiono w Tabelach (), .

4.1.4. Zbiór *TR* niezrównoważony i *TE* zrównoważony

W ostatniej części uczono klasyfikator na zbiorze niezrównowazonym a testowano na zrównoważonym. Wyniki w Tabelach () ukazują otrzymane wyniki.

Rozdział 5

Dyskusja

Rozdział 6

Podsumowanie