UNIWERSYTET WSB MERITO W GDAŃSKU

WYDZIAŁ INFORMATYKI I NOWYCH   
TECHNOLOGII

**ANALIZA PORÓWNAWCZA KONWOLUCYJNYCH SIECI NEURONOWYCH INCEPTION RESNET V2 I XCEPTION W ZASTOSOWANIU ICH DO KATEGORYZOWANIA ZMIAN SKÓRNYCH**

Praca inżynierska

na kierunku Informatyka

Praca napisana pod kierunkiem

dr inż. Pawła Tomkiewicza

Marek Małek, 56572, Gdańsk 2024

Spis treści

[WSTĘP 3](#_Toc154777841)

[Rozdział 1 – Historia Sieci Neuronowych 5](#_Toc154777842)

[1.1 Wprowadzenie 5](#_Toc154777843)

[1.2 Konwolucyjne Sieci Neuronowe (CNN) 6](#_Toc154777844)

[1.3 Mechanizm propagacji wstecznej 7](#_Toc154777845)

[1.4 Operacja splotu 8](#_Toc154777846)

[Rozdział 2 – Architektury InceptionResNetV2 i Xception 9](#_Toc154777847)

[2.1 Architektura InceptionResNetV2 9](#_Toc154777849)

[2.2 Architektura Xception 10](#_Toc154777850)

[Rozdział 3 – Proces Trenowania Modeli 12](#_Toc154777851)

[3.1 Zbiór danych i EDA 12](#_Toc154777853)

[3.2 Trenowanie i rezultaty modeli bazowych 13](#_Toc154777854)

[3.3 Trenowanie sieci neuronowych 15](#_Toc154777855)

[3.4 Usprawnianie sieci neuronowych 15](#_Toc154777856)

[3.5 Porównanie wyników 15](#_Toc154777857)

[Rozdział 3 – Wzbogacanie Zestawu Danych 16](#_Toc154777858)

[4.1 Przyczyna i cel wzbogacania danych 16](#_Toc154777860)

[4.2 Wymagania techniczne 17](#_Toc154777861)

[Rozdział 4 18](#_Toc154777862)

[5.1 Implementacja 18](#_Toc154777864)

[5.2 Testowanie 23](#_Toc154777865)

[5.3 Wdrożenie 24](#_Toc154777866)

[BIBLIOGRAFIA 25](#_Toc154777867)

[Opracowania książkowe 25](#_Toc154777868)

[Netografia 25](#_Toc154777869)

[SPIS RYSUNKÓW 26](#_Toc154777870)

# WSTĘP

Celem pracy jest wykonanie analizy porównawczej dwóch, klasycznych już architektur konwolucyjnych sieci neuronowych, którymi są InceptionResNetV2 oraz Xception. Obie te architektury służyć będą za rdzeń kilkunastu wariacji mających na celu osiągnięcie wyników lepszych, niż byłoby to możliwe przy użyciu niezmodyfikowanych sieci i surowych danych. W jej skład wchodzić będą następujące części:

* Krótka, eksploracyjna analiza danych (en. EDA[[1]](#footnote-1)) – jej celem jest sprawdzenie rozkładu klas obrazów
* Wyszukiwanie modelu bazowego – aby wyraźnie zobaczyć przewagę sieci neuronowych w danym zagadnieniu, dobrze jest najpierw zastosować prostsze i szeroko znane techniki oraz sprawdzić jaką one mają wydajność
* Sprawdzenie jak dobrze na zestawie danych HAM10000[[2]](#footnote-2) działa architektura InceptionResNetV2[[3]](#footnote-3), na razie ignorując dysbalans klas; sprawdzenie, czy będzie w stanie pobić model bazowy
* Dostrajanie wybranej architektury tak, aby lepiej radziła sobie na zbiorze danych HAM10000
* Wytrenowanie na tym samym zestawie klas architektury Xception[[4]](#footnote-4) i porównanie jej wyników z InceptionResNetV2

Poza wymienionymi składowymi, wykonane zostaną także skrypty ułatwiające zarządzanie kodem oraz automatyzujące wielokrotne uruchamianie procesów treningowych. Jest to potrzebne, aby móc wyciągnąć poprawne wnioski, ponieważ w szkoleniu każdej sieci neuronowej jest miejsce na element losowości. Stąd dla zrozumienia, która modyfikacja miała pozytywny wpływ na trening potrzebne jest kilkukrotne uruchomienie go oraz uśrednienie wyników, a następnie analiza statystyczna.

W rozdziale pierwszym i drugim została zawarta część teoretyczna pracy. Opisano w nich krótko historię sieci neuronowych, w szczególności splotowych sieci neuronowych (CNN[[5]](#footnote-5)), przybliżono w nich informacje dotyczące technik, które sprawiły że są tak skuteczne w rozpoznawaniu obrazów, a także opisano dwie wybrane architektury porównując je w miejscach, w których najmocniej się różnią.

Na początku rozdziału trzeciego opisano proces i skutki trenowania modeli bazowych, które ostatecznie posłużyły jako punkt wyjścia do poszukiwań właściwych architektur neuronowych. Następnie opisano modyfikacje mające usprawnić sieci InceptionResNetV2 oraz Xception i analogicznie do początku rozdziału – opisano proces i skutki trenowania tych modeli.

W rozdziale czwartym przedstawiono proces analizy statystycznej stanowiącej sposób dojścia do opisanych dalej wniosków. Opisano tu też problemy, z którymi spotkano się w trakcie prac.

# Rozdział 1 – Historia Sieci Neuronowych

## Wprowadzenie

Większość wytwarzanego oprogramowania od samego początku była algorytmiczna. Program miał zawierać procesy i reguły biznesowe danego przedsiębiorstwa i w ten sposób automatyzować powtarzalne czynności, które można przelać na kod w postaci szeregu instrukcji. Oprogramowanie takie zawsze jest specyficzne, rzadziej dla całej domeny, w której porusza się firma, częściej dla konkretnego jej wycinku, specyfiki działania samego przedsiębiorstwa. Oczywiście istnieją gotowe produkty, takie jak SAP ERP[[6]](#footnote-6), które operują w obrębie całych domen, jednak wciąż ich zasada działania jest taka, jak mniejszych programów – do aplikacji trafiają dane zewnętrzne, które po przetworzeniu przez ściśle określone reguły biznesowe dają spodziewany rezultat. Od zawsze istniała jednak potrzeba istnienia oprogramowania posługującego się logiką rozmytą[[7]](#footnote-7) tak, aby dla pewnych klas problemów nie trzeba było tworzyć tak sztywnych i dokładnych reguł przetwarzania danych. W istocie, być może dla większości tych klas problemów nie byłoby to nawet możliwe. Dla przykładu, w praktyce niemożliwe jest napisanie dobrego klasyfikatora pisma odręcznego.

Różne algorytmy uczenia maszynowego znane były od połowy ubiegłego stulecia, a wciąż krótką historię samych sieci neuronowych podzielić można na kilka okresów:

* Wczesne koncepcje lat 40-tych i 50-tych: wtedy to Warren McCulloch i Walter Pitts tworzą pierwszy „obliczeniowy” model biologicznego neuronu[[8]](#footnote-8).
* Wynalezienie Perceptronu[[9]](#footnote-9) w późnych latach 50-tych: była to wczesna sieć neuronowa zdolna nauczyć się rozpoznawania prostych wzorców, miała jednak spore ograniczenia, np. nie była w stanie rozwiązywać problemów, które nie były liniowo rozdzielne (np. problem XOR[[10]](#footnote-10))
* „Pierwsza Zima Sztucznej Inteligencji” w późnych latach 60-tych i 70-tych: przez ograniczenia perceptronu oraz niewielkiej dostępnej wtedy badaczom mocy obliczeniowej zainteresowanie inwestorów mocno ucierpiały przez co rozwój AI bardzo spowolnił.
* Wynalezienie algorytmu Propagacji Wstecznej[[11]](#footnote-11): umożliwiło to sieciom neuronowym uczenie się bardziej skomplikowanych wzorców.
* „Druga Zima Sztucznej Inteligencji” w późnych latach 80-tych i 90-tych: nastała z powodu zbyt dużych oczekiwań i ograniczonej dostępności mocy obliczeniowej. Innymi słowy sprzęt komputerowy wciąż był zbyt słaby, aby sieci neuronowe mogły pokazać swoje prawdziwe możliwości. Wtedy też na popularności zyskały mechanizmy uczenia maszynowego niezwiązane z sieciami neuronowymi, takie jak Support Vector Machines[[12]](#footnote-12)
* Ponowna faza wzrostu zainteresowania AI oraz czas przełomów: od początku lat 2000 ponownie widać było rosnące zainteresowanie sieciami neuronowymi przez badaczy i wielkie firmy technologiczne. W okolicach początku drugiego dziesięciolecia lat dwutysięcznych zaczęły się pojawiać kolejne przełomowe odkrycia i czas ten trwa do teraz. Kilka największych z nich to np. AlexNet[[13]](#footnote-13), która to sieć stanowiła przełom w dziedzinie rozpoznawania obrazów, odkrycie architektur GAN[[14]](#footnote-14) oraz Transformer[[15]](#footnote-15) (która to jest używana np. w święcącym sukcesy ChatGPT).

## Konwolucyjne Sieci Neuronowe (CNN)

Chociaż historia rozpoznawania obrazów przy pomocy sieci neuronowych, przynajmniej na poziomie koncepcji, sięga lat 80-tych, to pierwszym najgłośniejszym zastosowaniem tego typu sieci była architektura LeNet-5, stworzona w późnych latach 90-tych przez obecnie szeroko znanego badacza sztucznej inteligencji – Yanna LeCuna. Celem jej stworzenia było umożliwienie Amerykańskiemu Biuru Pocztowemu automatycznego odczytu kodów pocztowych z kopert. Była to dość prosta (z dzisiejszego punktu widzenia) architektura, której składowe to warstwa wejścia o wymiarach 32x32x1 (gdzie pierwsze dwie liczby to wysokość i szerokość obrazów, a ostatnia to ilość kanałów), warstwa splotowa, warstwa uśredniająca, jeszcze jedna warstwa splotowa, po niej uśredniająca i ostatnia splotowa, po której znajdują się dwie warstwy gęsto połączone oraz warstwa wyjścia, służąca do klasyfikacji liczb. Warstwy splotowe wykrywają wzorce przestrzenne, natomiast warstwy uśredniające zmniejszają rozmiar danych pochodzących z ich poprzedników. Zostało to zilustrowane na Rysunku 1.

Rysunek 1 – Architektura Sieci LeNet-5

Obraz zawierający tekst, diagram, zrzut ekranu, linia

Opis wygenerowany automatycznie

Źródło: https://www.analyticsvidhya.com/blog/2021/03/the-architecture-of-lenet-5/

Architektura ta stała się punktem wyjścia dla twórców bardziej złożonych sieci neuronowych takich jak AlexNet, VGGNet, czy ResNet.

Być może największym wkładem Yanna LeCuna w rozwój sztucznej inteligencji było zastosowanie mechanizmu tzw. Propagacji Wstecznej. Nie była to całkowicie nowa idea, jednak była świeża w świecie wielowarstwowych modeli konwolucyjnych. Jako, że jest to technika leżąca u podstaw każdego rodzaju sieci, nie tylko CNN, zostanie ona opisana jako pierwsza.

## Mechanizm propagacji wstecznej

Mechanizm ten składa się z kilku kroków:

* Przejście wprzód: warstwa po warstwie, aż do wyjścia przekazywane są dane. W przypadku sieci konwolucyjnej będą to coraz bardziej zmienione – zmniejszone, a na końcu spłaszczone do postaci wektora, dane obrazu.
* Obliczenie straty: funkcja straty (która jest jednym z tzw. hiperparametrów) oblicza różnicę między predykcją modelu, a stanem faktycznym – tym, co chciano by, aby model zwrócił.
* Przejście w tył (propagacja wsteczna): w tym miejscu algorytm oblicza gradienty funkcji straty względem każdej z wag w sieci. Każda z owych wag przyczynia się do uzyskania ostatecznego rezultatu, dlatego też sygnał błędu musi zostać rozpropagowany wstecz.
* Gradienty używane są do aktualizacji wag sieci.
* Cały cykl powtarza się na nowo, w każdym przejściu minimalizując wynik funkcji straty.

## Operacja splotu

Tak jak mechanizm propagacji wstecznej jest kluczowym elementem dla wszystkich rodzajów sieci neuronowych, to operacja splotu jest takim elementem w sieciach konwolucyjnych. Dzięki niej, sieci splotowe są w stanie lepiej, od np. sieci gęsto połączonych wychwytywać złożone wzorce, ale też, do pewnego stopnia są w stanie znajdywać je w różnych miejscach obrazów wejściowych. Składowymi operacji splotu są:

* Filtry (lub kernele): są to małe macierze wag, które podobnie jak wagi warstw gęstych ustawione są początkowo na losowe wartości.
* „Aplikowanie” filtrów na danych wejściowych: w tym momencie każda z wag mnożona jest przez wartość danego piksela. Kiedy mnożenia na obecnej pozycji filtra zostają zakończone, ich wyniki są do siebie dodawane – wynik tych operacji to wynik działania filtra na bieżącym obszarze obrazu.
* „Przesuwanie” filtrów: w tym momencie następuje przesunięcie kerneli o określoną ilość pikseli, aż do „końca” obrazu.
* Wyjściem z tej operacji są tzw. mapy cech, które reprezentują to, co w danym przejściu udało się wykryć warstwie splotowej.
* Na końcu, na każdej mapie cech wywoływana jest nieliniowa funkcja aktywacji (np. ReLU), po to, żeby model był zdolny wychwytywać skomplikowane wzorce.

# Rozdział 2 – Architektury InceptionResNetV2 i Xception



## Architektura InceptionResNetV2

Największą zmianą wprowadzoną przez autorów tej architektury w stosunku do jej poprzedników jest użycie modułów incepcyjnych oraz połączeń skrótowych (residual connections). Obie te techniki istniały już wcześniej, jednak użycie ich kombinacji umożliwiło utworzenie dużo głębszych struktur, które oprócz cechowania się wyższą dokładnością, mniej cierpiały z powodu problemu tzw. zanikającego gradientu.

Problem ten został zauważony przez badaczy, kiedy próbowali uczyć bardzo głębokie sieci. Opisany wcześniej mechanizm propagacji wstecznej może cechować się tym, że wagi warstw położonych najdalej są w procesie treningu dostosowywane w najmniejszym stopniu, co z kolei skutkuje tym, że warstwy te przestają się uczyć i sieć przykłada nieproporcjonalnie wyższą uwagę to położonych wyżej. To z kolei sprawia, że stosowanie głębokich sieci neuronowych może nie przynieść żadnych korzyści, lub wręcz sprawić, że ich dokładność będzie niższa od ich prostszych rodzajów. W poszukiwaniu architektur zdolnych rozpoznawać coraz bardziej skomplikowane wzorce, narodziła się konieczność zapobieżenia temu zjawisku. Połączenia skrótowe (pomijające) działają tak, że warstwa poprzedzająca warstwę bezpośrednio nad nią, zwraca swój rezultat do warstwy położonej jeszcze wyżej – oba wyniki są dodawane i w ten sposób mechanizm propagacji wstecznej może w bardziej wydajny sposób rozpropagować gradient funkcji straty.

Jeśli zaś chodzi o ideę, która stała za zastosowaniem modułów incepcyjnych, to badaczom chodziło o to, aby móc wykrywać skomplikowane wzorce w różnych skalach. Stąd, moduły incepcyjne przeprowadzają operację splotu używając jednocześnie filtrów o różnych wielkościach (np. 5x5, 3x3).

Na rysunku 2 zaprezentowane zostały oba omawiane moduły. Jest to bardzo podstawowa prezentacja tej idei, a w jej konkretnych implementacjach występują różne bloki incepcyjne oraz pomijające, bardziej rozbudowane od poniższych.

Po lewej stronie widać, że blok incepcyjny rozgałęzia się na kilka „ramion” przechodzących przez dane filtrami różnych wielkości. Na końcu wyniki tych równoległych operacji łączone są do postaci jednej mapy cech.

Rysunek 2 - przykład modułu incepcyjnego

Obraz zawierający tekst, diagram, linia, zrzut ekranu

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznie

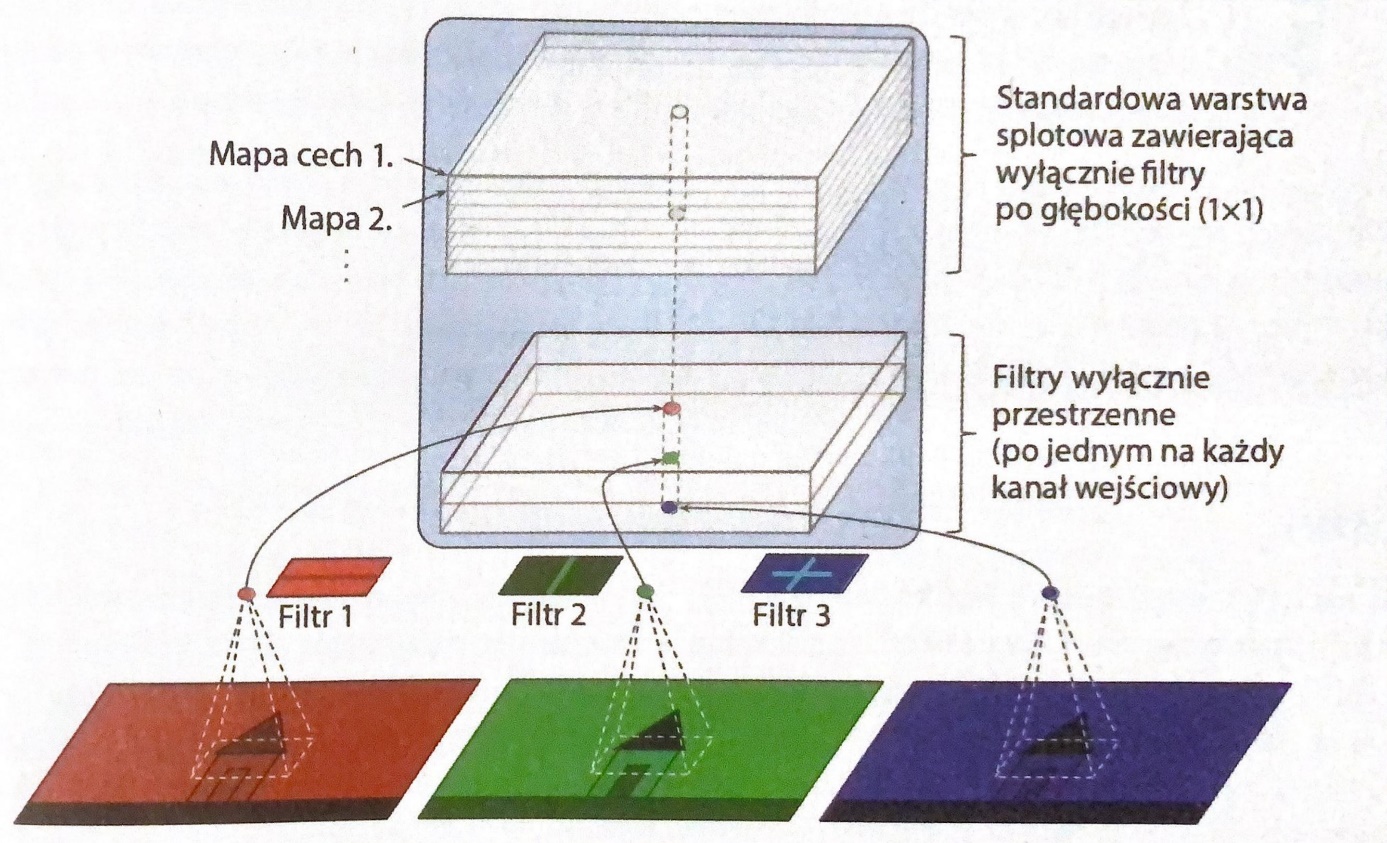
Źródło: https://www.researchgate.net/figure/Basic-architecture-of-inception-block\_fig4\_330511306  
Źródło: https://arxiv.org/pdf/1602.07261v2.pdf? https://arxiv.org/pdf/1602.07261v2.pdf?

Po prawej stronie widać schemat bloku zawierającego połączenie pomijające. W tym przykładzie, dla uproszczenia użyto dwóch warstw splotowych, jednak nic nie stoi na przeszkodzie, aby w ich miejscu pojawił się blok incepcyjny taki, jak na schemacie z lewej strony.

## Architektura Xception

Architektura Xception została zaproponowana jako wariacja podejścia prezentowanego przez Inception, z tą różnicą, że bloki incepcyjne zastąpione zostały rozdzielnymi po głębokości warstwami splotowymi. Klasyczna sieć splotowa wykorzystuje filtry starające się jednocześnie wykrywać wzorce przestrzenne i międzykanałowe, natomiast w rozdzielnej warstwie splotowej użyto założenia, że wzorce przestrzenne i międzykanałowe mogą być modelowane oddzielnie. Między tą architekturą a kolejnymi podejściami do pomysłów zawartych w linii modeli Inception istnieje też sporo innych różnic, jednak ta jest kluczowa i największa. Dzięki użyciu operacji splotu w opisany powyżej, alternatywny sposób, Xception uznawane jest za generalnie skuteczniejszą architekturę. Rysunek 3 zawiera diagram kluczowego elementu:

Rysunek 3- wykorzystanie filtrów po głębokości



Źródło: Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow

# Rozdział 3 – Proces Trenowania Modeli



## Zbiór danych i EDA

Jako zbiór danych wybrano dostępny w portalu kaggle.com HAM10000. Zawiera on 10 000 obrazów zmian skórnych dla 7 kategorii. Oprócz samych zdjęć, obecne są także pewne informacje o pacjentach, takie jak wiek, płeć, umiejscowienie zmiany i metoda jej wykrycia. Ostatecznie nie zdecydowano się na ich użycie, aby lepiej skupić się na rdzeniu rozpatrywanego problemu, czyli rozpoznawaniu obrazów. Rysunek 4 prezentuje histogramy zliczające wystąpienia określonych zmian skórnych wśród płci, gdzie na niebiesko zaznaczono mężczyzn, a na pomarańczowo kobiety:

Rysunek 4 - EDA

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

Źródło: opracowanie własne

Ciekawą obserwację być może stanowi fakt, że dla większości zmian skórnych obecnych w zbiorze HAM 10000 ich ofiarami nieco częściej padają mężczyźni. Nie ma to jednak wpływu na efektywność treningu, wybranych architektur.

## Trenowanie i rezultaty modeli bazowych

Przez modele bazowe, w kontekście tej pracy należy rozumieć modele niewykorzystujące sieci neuronowych. Wybrano to podejście, aby unaocznić, że prostsze algorytmy uczenia maszynowego nie są skuteczne w obliczu zbioru danych tak dużego i skomplikowanego jak HAM 10000. Chociaż były w stanie osiągnąć zbieżność szybciej od sieci neuronowych, to ich dokładność nigdy nie przekroczyła 41%.

Pierwszy z wybranych algorytmów to KNeighborsClassifier, obecny w pythonowej bibliotece scikit-learn (w skrócie sklearn). W świecie data science jest on dość popularny i często wybierany jest jako klasyfikator nawet dla bardziej złożonych problemów – przez swoją skuteczność i prostotę. Rysunek 5 pokazuje punkt wyjściowy, czyli trening tego klasyfikatora bez żadnych modyfikacji:

Rysunek 5 - pierwsze podejście do treningu KNN

Obraz zawierający tekst, zrzut ekranu, Czcionka, oprogramowanie

Opis wygenerowany automatycznie

Źródło: opracowanie własne

Dokładność lekko poniżej 41% była niezadawalająca, dlatego też spróbowano usprawnić ten wynik dodając do procesu treningu przeszukiwanie siatki. Polega ono na tym, że interfejs sklearn pozwala użytkownikowi na stworzenie tzw. pipeline, w którym to kolejne kroki uruchamiane są z różnymi kombinacjami parametrów, w ten sposób próbując przetestować każdą taką kombinację i na końcu wybierając najlepszą z nich. Najpopularniejszym podejściem używanym przy treningu modeli dostarczanych przez sklearn jest użycie PCA. Jest to technika redukcji wymiarowości, która powinna nieco uprościć (lub odszumić) dane, dekorelując je i zmniejszając liczbę ich wymiarów. Rysunek 6 pokazuje skutki treningu z użyciem PCA:

Rysunek 6 - trenowanie KNN z pomocą PCA

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznie

Źródło: opracowanie własne

Jak widać, nastąpiła nieznaczna poprawa, wciąż jednak było to zbyt mało. Jako kolejny klasyfikator wybrano SGD, również z biblioteki sklearn. Występują między nimi zasadnicze różnice, które da się streścić w słowach. KNeighborsClassifier może działać dobrze w sytuacji, kiedy granice decyzyjne są nietrywialne (a więc zbiór danych jest złożony), ale gorzej radzi sobie z dużymi zbiorami danych i ich wysoką wymiarowością, natomiast SGD lepiej radzi sobie wieloma danymi, ale wymaga większej dozy strojenia hiperparametrów. Efekty szkolenia drugiego z nich widoczne są na rysunku 5:

Rysunek 7 - trenowanie SGD

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznie

Źródło: opracowanie własne

Jest to najlepszy wynik, który udało się osiągnąć przy użyciu tego klasyfikatora, jednocześnie jak widać gorszy nawet od uzyskanego przez KNeighborsClassifier.

## Trenowanie sieci neuronowych

## Usprawnianie sieci neuronowych

## Porównanie wyników

# Rozdział 3 – Wzbogacanie Zestawu Danych



## Przyczyna i cel wzbogacania danych

Sieci neuronowe w najlepszy sposób odnajdują rozmaitości, jeśli dysponują dużymi i zróżnicowanymi zestawami danych. Mogą wyuczyć się rozpoznawania niektórych cech, na mniejszych zestawach, ale zasady ich działania sprawiają, że tym lepiej będą to robić, im lepsze dane dostaną w trakcie uczenia. W przypadku tego konkretnego problemu – rozpoznawania zmian skórnych na podstawie obrazów dermatoskopowych – zastosowano kilka popularnych technik augmentacji, takich jak:

* Przesunięcie w pionie lub poziomie – wiele obrazów pokazuje zmianę skórną mniej więcej w centrum, podczas gdy w rzeczywistości, osoba wykonująca zdjęcie może nie być tak dokładna. Aby więc zapewnić dobrą generalizację modelu na obrazach zmian skórnych obecnych w innym, niż centralne miejsce obrazu, można zastosować ten rodzaj augmentacji.
* Zmiana skali obrazu – czyli jego powiększenie, lub pomniejszenie. Przyczyna zastosowania tej techniki jest taka sama, jak w poprzednim podpunkcie.
* Zmiana jasności obrazu – aby symulować różne warunki oświetlenia, w których zdjęcia mogłyby być robione.

Istnieje również wiele innych technik wzbogacania obrazów, jednak uznano, że dla rozpatrywanego problemu będą one nieskuteczne, lub wręcz szkodliwe. Kilka z nich to:

* Obrót obrazu – zmiany skórne nie posiadają elementów, które powtarzają się w tych samych miejscach. Nie ma więc znaczenia pod jakim kątem zdjęcie zostało wykonane, ponieważ sieć neuronowa i tak powinna je poprawnie rozpoznać. Obroty obrazów mają więcej sensu, kiedy trenowana jest sieć rozpoznająca twarze, przedmioty, czy istoty żywe – obiekty, które wykazują sporą regularność.
* Odbijanie obrazu w pionie, czy poziomie – jw.
* Przesunięcie wartości kanałów – w przypadku rozpoznawania zmian skórnych ta technika mogłaby być szkodliwa dla dokładności modelu, ponieważ kolor takiej zmiany jest jednym z czynników używanych podczas klasyfikacji, Przesuwając wartości kanałów, istotna informacja zostałaby zniszczona.

## Wymagania techniczne

Aplikacja ta została przygotowana do używania na urządzeniach typu PC (w tym, przede wszystkim wszelkiej klasy środowiska serwerowe) i laptop. Z uwagi na wysokie wymagania dotyczące ilości rdzeni procesora i dostępnej pamięci, tylko te dwa środowiska zdolne będą ją uruchomić.

Jeśli chodzi o wymagania sprzętowe stawiane środowisku lokalnemu, takiemu jak laptop, są to:

* system operacyjny zdolny uruchamiać skrypty Powershell – Windows lub Linux z zainstalowanymi stosownymi pakietami
* zainstalowane oprogramowanie Docker Desktop
* procesor zdolny uruchomić 5 wątków
* 14gb pamięci RAM

Są to wymagania minimalne, takie które sprawią, że uruchomienie tej aplikacji nie sprawi, że komputer przestanie być responsywny.

Jeśli chodzi o wymagania sprzętowe stawiane środowisku serwerowemu, są to:

* system operacyjny Linux na serwerze uruchamiającym aplikację
* co najmniej trzy niezależne serwery hostujące węzły bazy danych Cassandra (4-rdzeniowe procesory, co najmniej 128gb pamięci RAM)
* co najmniej trzy niezależne serwery hostujące węzły bazy danych ElasticSearch (technicznie, jw.)

# Rozdział 4

**IMPLEMENTACJA APLIKACJI**



## Implementacja

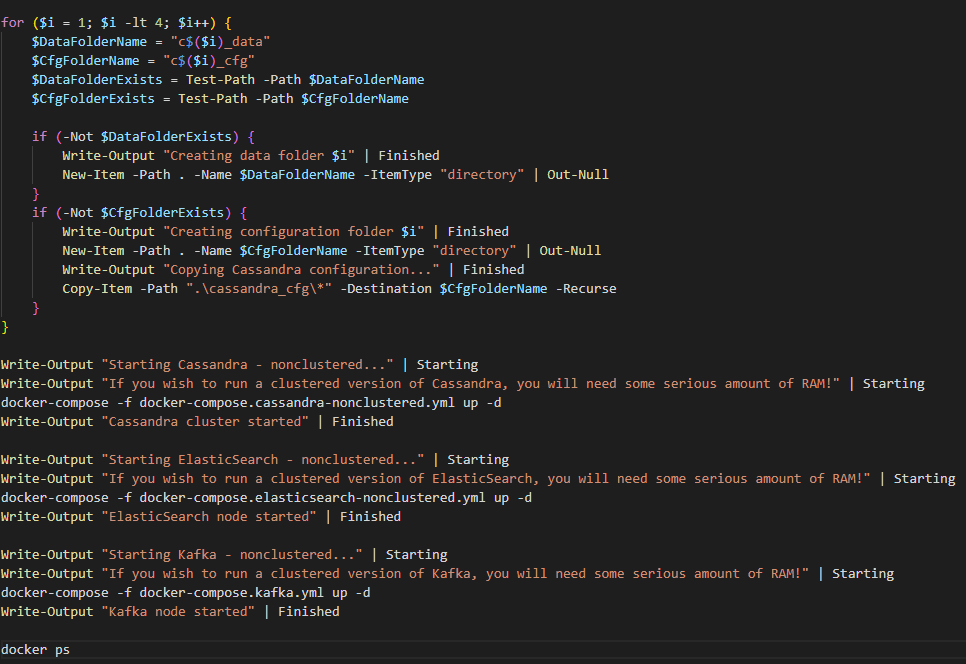
Tworzenie aplikacji tego typu należy rozpocząć od instalacji niezbędnych komponentów infrastrukturalnych. Pierwszym tego typu elementem jest oprogramowanie Docker Desktop. Dla systemu Windows jego autorzy przewidzieli dedykowany instalator, który należy ściągnąć z internetu i uruchomić. Podczas instalacji należy zaakceptować domyślnie wybrane opcje.

Po zainstalowaniu Dockera należy uruchomić w nim bazy danych wymagane przez aplikację do działania. W rzeczywistości, kiedy aplikacja uruchomiona byłaby na serwerze, Docker wymagany byłby tylko po to, aby móc ją uruchamiać, natomiast same bazy danych uruchamiane byłyby dla zapewnienia najwyższej wydajności w trybie tzw. „bare metal” – bezpośrednio na serwerze.

Projekt zawiera skrypt automatyzujący instalację wymaganego oprogramowania. Znajduje się on w ścieżce (zaczynając od głównego folderu projektu): /devops/runme.ps1.

Rysunek 8

Skrypt automatyzujący infrastrukturę

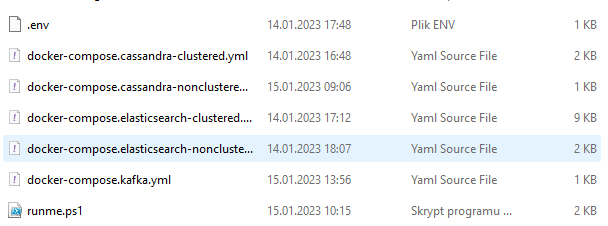


Źródło: opracowanie własne

Jego jedyną rolą jest uruchomienie oprogramowania docker, które automatycznie zainstaluje elementy zdefiniowane w plikach docker compose widocznych poniżej:

Rysunek 9

Lista plików docker-compose

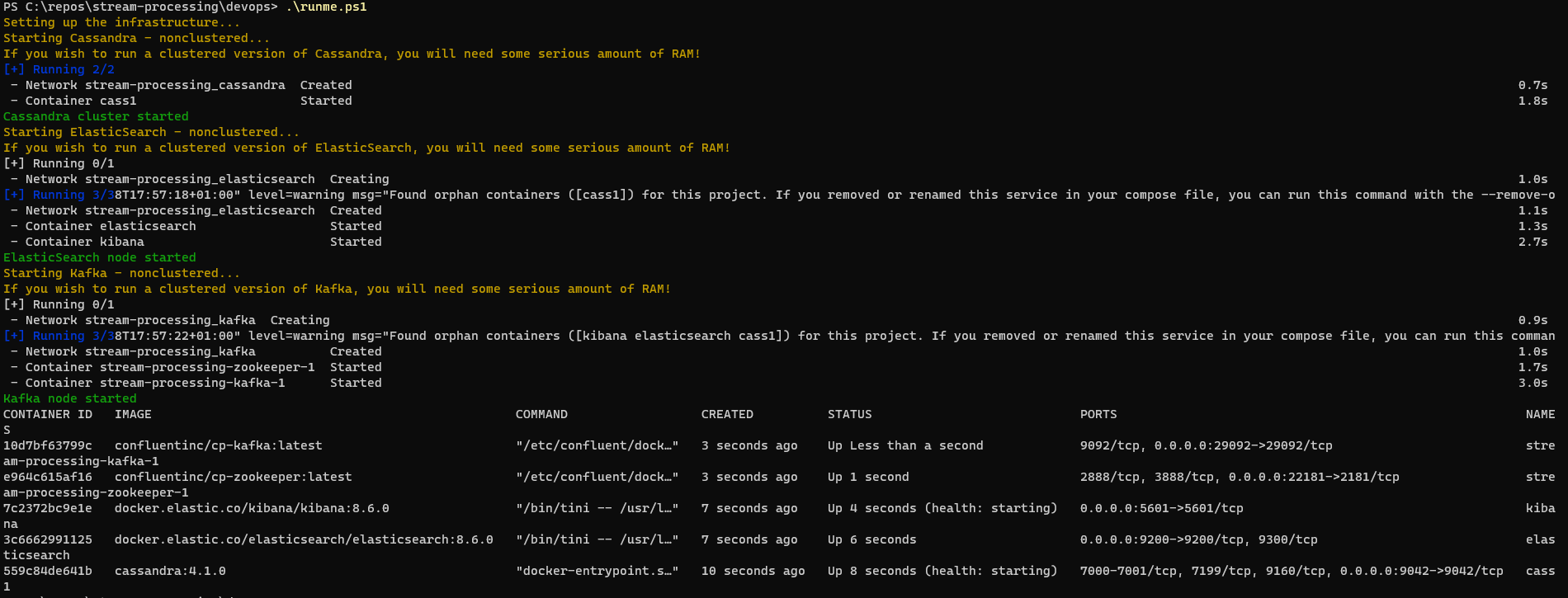


Źródło: opracowanie własne

Powszechną praktyką, w wypadku konfiguracji skomplikowanej infrastruktury przy użycia Dockera jest użycie oddzielnych plików definiujących obrazy[[16]](#footnote-16) dla ułatwienia zarządzania nimi. Po zakończeniu pracy skryptu, w konsoli powinniśmy zobaczyć rezultat podobny do poniższego:

Rysunek 10

Rezultat uruchomienia skryptu runme.ps1

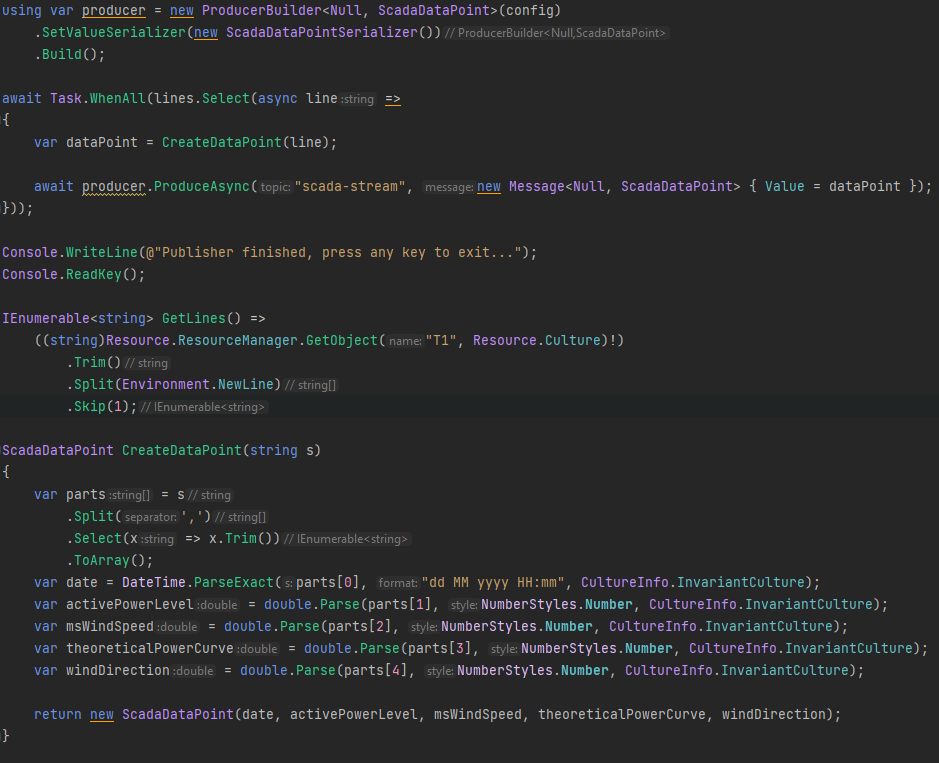


Źródło: opracowanie własne

Kolejnym krokiem jest stworzenie aplikacji umieszającej dane w kolejce Kafka. Potrzebujemy jej z uwagi na testowanie – tworząc aplikację nie mamy możliwości interakcji z prawdziwym systemem SCADA. Nie mamy jej aż do momentu wejścia na środowisko preprodukcyjne, chociaż równie często spotykana sytuacja wygląda tak, że nie ma się styczności z faktycznymi turbinami aż do wejścia na produkcję. Stąd, aby móc przetestować cały proces, musimy jego początek emulować. Nie jest to trudne, a mając dobrze zdefiniowane wymagania biznesowe dotyczące tego, co mamy składować w bazach danych kod takiego emulatora (po zignorowaniu importu bibliotek i początkowych linii konfiguracyjnych) wygląda następująco:

Rysunek 11

Kod programu - producenta danych



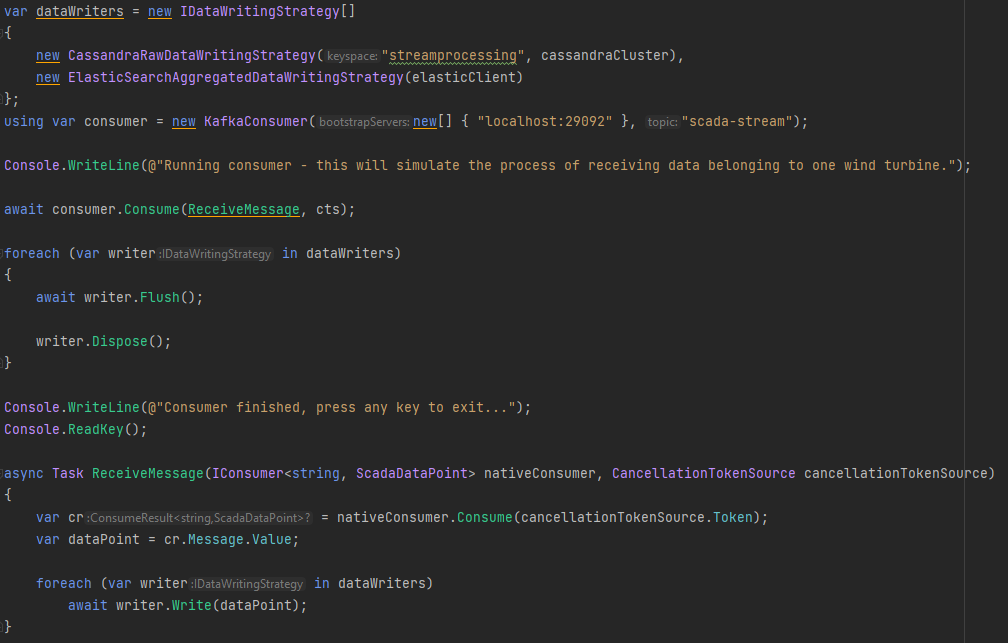
Źródło: opracowanie własne

Pierwsze linie widoczne na powyższym rysunku starają się równolegle wysłać wiele wierszy danych. Odczyt owych linii (z pliku) definiuje funkcja GetLines, następnie obiekt typu ScadaDataPoint tworzony jest z kolejnych linii przy użyciu metody CreateDataPoint.

Jak zostało wspomniane wcześniej, program ten jest tylko emulatorem – pomocą developerską istniejącą po to, aby lokalny rozwój aplikacji był możliwy i sam nie będzie umieszczany na produkcji. Dużo ciekawszym zatem programem jest aplikacja konsumenta. Kod jej punktu wejścia zaprezentowany został na kolejnym rysunku:

Rysunek 12

Punkt wejściowy aplikacji - konsumenta



Źródło: opracowanie własne

Pierwsze linie definiują dwie strategie[[17]](#footnote-17) zapisu danych. Aby wykorzystać polimorfizm[[18]](#footnote-18), a więc uprościć kod programu i uczynić go gotowym na ewentualne, przyszłe zmiany, a jednocześnie bardziej ogólnym, wzmiankowane strategie implementują interfejs IDataWritingStrategy. W tej prototypowej aplikacji powstały jedynie dwie strategie, ale z tak napisanym kodem, łatwo można rozszerzyć ich listę o kolejne.

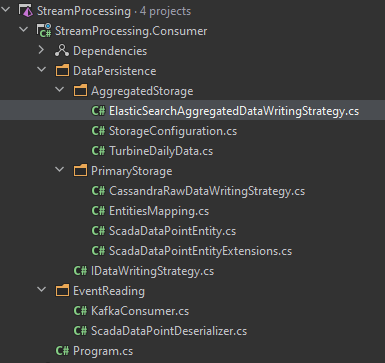
Pierwszą z nich jest CassandraRawDataWritingStrategy. Ma ona na celu zapis surowych danych – w niezmienionej formie. Powodem jej istnienia jest to, że nasza firma chce, aby jej analitycy mogli z takich surowych danych budować własne modele dynamicznie. Innymi słowy dzisiaj dla firmy może najbardziej istotny będzie poziom produkowanej mocy, a więc wszystkie analizy powinny być prowadzone na tej tylko kolumnie. W przyszłości jednak wiemy, że istotne będzie też badanie awaryjności turbin wiatrowych. Oprócz poziomu mocy istnieją też inne wskaźniki, które mogą analitykom i ich modelom uczenia maszynowego sugerować, że dla danej turbiny istnieje pewne wysokie prawdopodobieństwo nadchodzącej awarii.

Drugą zaimplementowaną strategią jest ElasticSearchAggregateDataWritingStrategy. Ta ma na celu budowanie dziennych agregatów danych. Oczywiście analitycy mogliby je sami budować, jednak tutaj cel jest nieco inny. Wymaganie biznesowe jest takie, aby aplikacja frontendowa potrafiła generować pewne typy prostych raportów na żądanie, bez udziału analityków. Aby temu wymaganiu sprostać, powstał pomysł dziennych agregacji danych.

Prócz tych dwóch, kluczowych elementów aplikacji konsumenckiej, zawiera ona kolejny – KafkaConsumer. Nie zapisuje on danych, lecz odczytuje je z kolejki Kafka i przekazuje dwóm opisywanym wcześniej strategiom. Aby podsumować kształt aplikacji konsumenta, załączony zostaje kolejny rysunek:

Rysunek 13

Wysokopoziomowy rzut na elementy aplikacji



Źródło: opracowanie własne

## Testowanie

Podczas testowania zostały przeprowadzone:

* Testy funkcjonalne
  + Sprawdzenie poprawności wysyłki danych
  + Sprawdzenie poprawności składowania danych w obu bazach
* Testy wydajnościowe
  + Sprawdzenie wydajności aplikacji pod kątem wolumenu danych, który zdolna jest obsłużyć

Dzięki poprawnemu użyciu technologii przeznaczonych dokładnie dla opisywanego scenariusza udowodniono, że cały wybrany stos technologiczny będzie skalował się liniowo wraz z dodawaniem do systemu kolejnych danych.

## Wdrożenie

Sugerowanych jest tutaj kilka rozwiązań. Aplikacja została przygotowana w taki sposób, aby mogła operować zarówno w środowisku „on-premise[[19]](#footnote-19)”, jak i chmurowym. Poza infrastrukturą w postaci wymienionych już bazy danych Cassandra, ElasticSearch oraz systemu wysyłki zdarzeń Kafka, aplikacja nie ma zewnętrznych zależności. Zatem, jeśli poda się jej odpowiednie dane konfiguracyjne takie jak hasła dostępowe i adresy serwerów, bez zmian samego kodu możliwe będzie jej przeniesienie na dowolny serwer, komunikujący się z dowolną ilością serwerów hostujących wymienione części infrastruktury.

# BIBLIOGRAFIA

## Opracowania książkowe

1. Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides: **Wzorce projektowe. Elementy oprogramowania obiektowego wielokrotnego użytku.** – Wydawnictwo Helion
2. Jeff Carpenter, Eben Hewitt: **Cassandra: The Definitive Guide, 3rd Edition.** – Wydawnictwo O’Reilly
3. Clinton Gormley: **Elasticsearch: The Definitive Guide**. – Wydawnictwo O’Reilly
4. Neha Narkhede, Gwen Shapira, Todd Palino: **Kafka: The Definitive Guide.** – Wydawnictwo O’Reilly

## Netografia

1. https://en.wikipedia.org/wiki/SCADA (data odczytu 17.01.2023)
2. https://learn.microsoft.com/pl-pl/aspnet/core/introduction-to-aspnet-core?view=aspnetcore-7.0 (data odczytu 17.01.2023)
3. https://learn.microsoft.com/pl-pl/aspnet/core/blazor/?view=aspnetcore-7.0 (data odczytu 17.01.2023)
4. https://learn.microsoft.com/pl-pl/xamarin/ (data odczytu 17.01.2023)
5. https://unity.com/how-to/learning-c-sharp-unity-beginners (data odczytu 17.01.2023)
6. https://learn.microsoft.com/en-us/dotnet/standard/garbage-collection/fundamentals (data odczytu 17.01.2023)
7. https://www.jetbrains.com/rider/ (data odczytu 17.01.2023)
8. https://cassandra.apache.org/\_/index.html (data odczytu 17.01.2023)
9. https://pl.wikipedia.org/wiki/Elasticsearch (data odczytu 17.01.2023)
10. https://lucene.apache.org/ (data odczytu 17.01.2023)
11. https://blog.packagecloud.io/what-is-a-docker-image/ (data odczytu 18.01.2023)
12. https://refactoring.guru/design-patterns/strategy (data odczytu: 18.01.2023)
13. https://www.modestprogrammer.pl/co-to-jest-polimorfizm-w-programowaniu-obiektowym (data odczytu: 18.01.2023)
14. https://www.forbes.com/sites/forbestechcouncil/2019/03/22/cloud-based-vs-on-premise-servers/?sh=a86be3f79e20 (data odczytu: 18.01.2023)

# SPIS RYSUNKÓW

Rysunek 1 9

Rysunek 2 10

Rysunek 3 10

Rysunek 4 11

Rysunek 5 12

Rysunek 6 13

1. https://en.wikipedia.org/wiki/Exploratory\_data\_analysis [↑](#footnote-ref-1)
2. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000 [↑](#footnote-ref-2)
3. https://arxiv.org/pdf/1602.07261v2.pdf [↑](#footnote-ref-3)
4. https://arxiv.org/pdf/1610.02357.pdf [↑](#footnote-ref-4)
5. https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html [↑](#footnote-ref-5)
6. https://pl.wikipedia.org/wiki/SAP\_ERP [↑](#footnote-ref-6)
7. https://towardsdatascience.com/a-very-brief-introduction-to-fuzzy-logic-and-fuzzy-systems-d68d14b3a3b8 [↑](#footnote-ref-7)
8. https://towardsdatascience.com/mcculloch-pitts-model-5fdf65ac5dd1 [↑](#footnote-ref-8)
9. https://maelfabien.github.io/deeplearning/Perceptron/#the-classic-model [↑](#footnote-ref-9)
10. https://www.niser.ac.in/~smishra/teach/cs460/2020/lectures/lec19/ [↑](#footnote-ref-10)
11. https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd [↑](#footnote-ref-11)
12. https://pl.wikipedia.org/wiki/Maszyna\_wektor%C3%B3w\_no%C5%9Bnych [↑](#footnote-ref-12)
13. https://proceedings.neurips.cc/paper\_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf [↑](#footnote-ref-13)
14. https://arxiv.org/pdf/1406.2661.pdf [↑](#footnote-ref-14)
15. https://arxiv.org/pdf/1706.03762.pdf [↑](#footnote-ref-15)
16. https://blog.packagecloud.io/what-is-a-docker-image/ (data odczytu 18.01.2023) [↑](#footnote-ref-16)
17. https://refactoring.guru/design-patterns/strategy (data odczytu: 18.01.2023) [↑](#footnote-ref-17)
18. https://www.modestprogrammer.pl/co-to-jest-polimorfizm-w-programowaniu-obiektowym (data odczytu: 18.01.2023) [↑](#footnote-ref-18)
19. https://www.forbes.com/sites/forbestechcouncil/2019/03/22/cloud-based-vs-on-premise-servers/?sh=a86be3f79e20 (data odczytu: 18.01.2023) [↑](#footnote-ref-19)