Objaśnianie modeli uczenia maszynowego

inż. Michał Brzozowski
Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej, Politechnika Łódzka
Łódź, Polska
254151@edu.p.lodz.pl

inż. Norbert Śmiechowicz
Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej, Politechnika Łódzka
Łódź, Polska
254248@edu.p.lodz.pl

STRESZCZENIE

Artykuł wprowadza do zagadnień interpretowalnej sztucznej inteligencji oraz porównuje wyniki różnych metod zastosowanych do oceny wytrenowanych wielowarstwowych perceptronów i konwolucyjnych sieci neuronowych klasyfikujących zestawy danych Iris, Wine, Breast Cancer, MNIST oraz CIFAR 10.

Interpretowalna Sztuczna Inteligencja, Konwolucyjne Sieci Neuronowe, Perceptrony Wielowarstwowe,

I. WPROWADZENIE

Na przestrzeni ostatnich lat modele uczenia maszynowego zyskały na znaczeniu oraz stały się fundamentalnym narzędziem w wielu dziedzinach nauki i przemysłu, jak rozpoznawanie obrazów i dźwięków czy klasyfikacja danych i szukanie trendów. Algorytmy takie jak sieci neuronowe (ang. *Neural Networks*), lasy losowe czy maszyny wektorów nośnych (ang. SVM – *Support Vector Machines*) potrafią przewidywać wyniki z imponującą dokładnością, często przewyższając tradycyjne metody analizy danych. Pomimo ich skuteczności oraz łatwości w implementacji, postrzegane są jednak jako czarne skrzynki, co stanowi jedno z głównym wyzwań związanych z ich szerszym wdrażaniem.

A. Problem czarnej skrzynki

Problem czarnej skrzynki (ang. *black box*) opisuje sytuacje w której operator mechanizmu nie posiada wiedzy o jego działaniu, a jedynie wyniki otrzymane dla kolejnych wartości wejściowych. Czarne skrzynki są kluczowym elementem szybkiego rozwoju wiedzy i technologii, ponieważ umożliwiały niezależną specjalizację w projektowaniu osobnych funkcjonalności koniecznych do działania większych systemów. Historycznie, czarne skrzynki miały gwarancję przewidzianego zachowania wystawianą przez projektantów, cecha, której brakuje modelom otrzymanym w wyniku uczenia maszynowego.

B. Interpretowalna Sztuczna Inteligencja

By zaradzić temu problemowi rozwinęła się dziedzina badań nad Interpretowalną Sztuczną Inteligencją (ang. Explainable Artificial Intelligence), w zakresie tego artykułu określaną również skrótem xAI. Dostarczone przez nią metody mają na celu kwantyfikować wpływy wyników modeli sztucznej inteligencji względem ich wejść tak, by dać operatorom intuicję w ich działaniu. Mają również umożliwić weryfikację czy model wykonuje zadania do których był zaprojekowany, a nie zadania o bliskich wynikach w dziedzinie wejść na których był trenowany. Jest to szczególnie istotne w sektorach, gdzie pomyłka wiąże się z poważnymi konsekwencjami, np. samo jeżdżące samochody, asysta w diagnozach chorób czy analizy finansowe.

Istotnym problemem jest wykrywanie błędnej klasyfikacji i stronniczości modelu, który bez świadomości operatorów rozpoznaje nie pożądane cechy, a cechy im silnie skorelowane w zestawie danych treningowych. Przykład opisany w [1] dotyczy modelu mającego za zadanie rozróżniać profesję lekarza i pielęgniarza. Ze względu jednak na stronniczość danych model w esencji wytrenowany został do rozróżniania kobiet od mężczyzn.

II. POWIĄZANE PRACE

Spośród metod xAI wyróżnić można dwa dominujące podejścia, metody oparte na gradientach i metody oparte na teorii gier.

Motywacja metod gradientowych wynika z inherentnej matematycznej konstrukcji modeli, które w gotowej postaci są niczym innym jak zestawem wielu powiązanych ze sobą funkcji. Badania nad zmiennością funkcji są obszerną, a zarazem wciąż rozwijającą się dziedziną matematyki oferującą wiele metod gotowych do rekontekstualizacji. Najprostszą z nich jest gradient, czyli bezpośrednia miara wpływu zmiany każdej cechy na zmianę wyjścia modelu, opisany między innymi w [2].

Metody teorii gier analizują wyniki modeli jako wartości osiągane poprzez kooperację cech wejściowych, a udział poszczególnej cechy rozumiany jest jako sprawiedliwy przydział wartości wynikowej. Popularną miarą są wartości Shapley'a [3], które modelują przyrost wartości wynikowej przypisanej danej cesze rozważając różnice między sytuacjami pokrewnymi z jej obecnością i brakiem.

III. WYKORZYSTANE METODY

Podczas przeprowadzonych eksperymentów skorzystano z metod objaśniających dostarczonych przez moduł Captum, odpowiednio dobranych do analizowanych zbiorów danych. W dalszej części artykułu stosowane są następujące pojęcia:

- Zadaniem klasyfikatora jest przypisanie wejściu modelu odpowiedniej klasy, gdzie wejście modelu rozumiane jest jako wektor cech analizowanego obiektu postaci $x = x_i \epsilon_i$, dla $i \in \{0, \dots, n-1\}$,
- Funkcja predykcji definiowana jest dla poszczególnych modeli i zbiorów danych, a rozumiana jako przypisanie wektora prawdopodobieństw wszystkich klas dla danego wejścia modelu. Wybrany element funkcji predykcji określany jest mianem funkcji predykcji klasy,
- Funkcja klasyfikacji lub klasyfikacja rozumiana jest jako przypisanie wejściu modelu klasy, której stowarzyszona funkcja predykcji przypisała najwyższą wartość dla rozważanego wejścia. Funkcja ta będzie określana mianem dominującej i oznaczona jako f_d.

A. *Moduł Captum*

Do przeprowadzenia eksperymentów wyjaśniających działanie modeli uczenia maszynowego wykorzystano metody dostarczone przez moduł Captum dla PyTorch [4]. Metody te przypisują elementom wejść modelu wskaźniki interpretowane jako udział poszczególnych elementów w otrzymanej klasyfikacji.

1) Metoda Map Ważności (ang. Saliency Map)

przypisuje elementom wejścia gradient względem dominującej funkcji predykcji klasy. Popularnym rozwiązaniem jest przypisanie wartości bezwzględnej gradientu by zwrócić uwagę na siłę wpływu cechy, jednak na potrzeby badań przeprowadzonych w tym artykule zdecydowano się na wartości względne. Mapa ważności S(x) dla modelu f i wejścia x dana jest przez gradient [5]:

$$S(x) = \frac{\partial f_d}{\partial x} \tag{1}$$

Mapa ważności zatem obrazuje, która cecha wejściowa miała największy wpływ na decyzję modelu.

2) Metoda Próbkowania Wartości Shapley'a (ang. Shapley Values Sampling)

polega na szacowaniu wartości Shapley'a przez sprawdzanie tylko losowo wybranych podzbiorów. Sprawdzana ilość pozostawiona została jako domyślna, równa 25. Dokładne wartości Shapley'a definiowane są jako:

$$SH_{x,b}(i) = \frac{1}{n}\sum_{S\subseteq N\setminus\{i\}} \binom{n-1}{|S|}^{-1} \left(f_d\left(v_{x,b}(S\cup\{i\})\right) - f_d\left(v_{x,b}(S)\right)\right) \tag{2}$$

gdzie:

$N = \{ m \in \mathbb{N}_0 \mid m < n \}$

$v_{x,b}(S) = \sum_{j=0}^{n-1} v_{x,b,j}(S)\,\epsilon_j, \qquad v_{x,b,j}(S) = \begin{cases} x_j, & dla\ j \in S \\ b_j, & dla\ j \notin S \end{cases}$

i – indeks rozważanej cechy,
b – wektor odniesienia.

3) Metoda Scałkowanych Gradientów (ang. Integrated Gradients)

przypisuje cechom ważność na podstawie różnicy między odpowiadającej jej wartości odniesienia i gradientu dominującej funkcji predykcji klasy według następującego wzoru [6]:

$$IG_i(x) = (x_i - b_i)\int_0^1 \frac{\partial f_d(b+\alpha(x-b))}{\partial x_i}\,d\alpha \tag{3}$$

gdzie:
i – indeks rozważanej cechy,
b – wektor odniesienia,
α – współczynnik skali, transformujący wartość odniesienia w wartość rozważaną.

4) Metoda Ablacji Cech (ang. Feature Ablation)

polega na pojedynczym usuwaniu cech i ocenie wpływu ich braku na predykcję modelu. Zmiana w wyniku modelu po usunięciu i-tej cechy dana jest przez [7]:

$$FA_i(x) = f_d(x) - f_d(x - (x_i - b_i)\epsilon_i) \tag{4}$$

gdzie:
b_i – i-ta cecha wektora odniesienia

5) Metoda Naprowadzanych, Ważonych Gradientem Mapowań Aktywacji Klas (ang. Guided Grad-CAM)

w zakresie niniejszego artykułu określana również w skrócie jako GG-CAM, polega na wizualizacji ważnych regionów w danych wejściowych modelu. Metoda ta łączy ze sobą gradienty z aktywacjami w wybranych warstwach sieci neuronowej (najczęściej na ostatniej warstwie konwolucyjnej w konwolucyjnych sieciach neuronowych) w celu uzyskania bardziej precyzyjnej mapy cech, które wpływają na decyzję modelu [1]. GG-CAM dana jest przez:

$$GC_i(x) = ReLU\left(\sum_k \alpha_k^c A^k\right), \alpha_k^c = \frac{1}{Z}\sum_i\sum_j -\frac{\partial f_d}{\partial A_{ij}^k} \tag{5}$$

gdzie:

α_k^c – waga przypisana aktywacji A^k dla klasy c,

Z – liczba elementów w mapie aktywacji A^k

IV. WYNIKI EKSPERYMENTÓW

Przeprowadzono eksperymenty objaśniające zasadę działania w podejmowaniu decyzji przez modele uczenia maszynowego, wykorzystując modele wielowarstwowych perceptronów (ang. *Multi Layer Perceptron*) oraz konwolucyjnych sieci neuronowych (ang. *Convolutional Neural Networks*).

A. *WYKORZYSTANE ZBIORY DANYCH*

Do przeprowadzenia eksperymentów wykorzystano zbiory danych pochodzące z następujących modułów:

a) Moduł scikit-learn [8]

- Iris - opisuje wymiary płatków i działki kielicha dla trzech różnych gatunków irysów (*łac. setosa, łac. virginica, łac. versicolor*), analizowanymi cechami są długość działki kielicha, szerokość działki kielicha, długość płatka, szerokość płatka [9],
- Wine – reprezentuje rezultaty analizy chemicznej dla trzech różnych rodzajów win. Analizowanymi cechami są alkohol, kwas jabłkowy, popiół, zasadowość popiołu, magnez, całkowite fenole, flawonoidy, nie fenolowe fenole, pro antocyjaniny, intensywność koloru, barwa, stosunek od280/od315 rozcieńczonych win, prolina [9],
- Breast Cancer Wisconsin – opisuje charakterystyki jąder komórkowe widoczne na zdigitalizowanych obrazach aspiratu cienkoigłowego (FNA) guza piersi dla dwóch różnych rodzajów M (*ang. malignant*) oraz B (*ang. benign*). Analizowanymi cechami są wartości średnie dla promienia, tekstury, obwodu, powierzchni, gładkości, zwartości, wklęsłości, liczby wklęsłych punktów, symetrii, wymiaru fraktalnego, odchylenia standardowe dla promienia, tekstury, obwodu, powierzchni, gładkości,

zwartości, wklęsłości, liczby wklęsłych punktów, symetrii, wymiaru fraktalnego, oraz największe wartości dla promienia, tekstury, obwodu, powierzchni, gładkości, zwartości, wklęsłości, liczby wklęsłych punktów, symetrii oraz wymiaru fraktalnego [9],

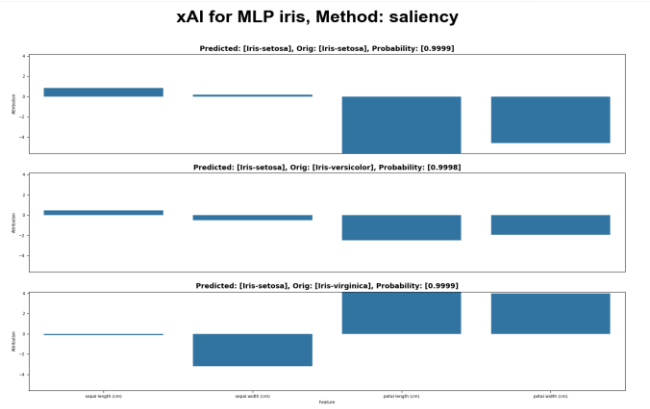
- MNIST – zbiór 60 000 czarno-białych obrazów cyfr o rozdzielczości 28x28 pikseli, przedstawiających ręcznie pisane cyfry [10].

b) Moduł Torchvision [11]

- CIFAR10 – zbiór 60 000 kolorowych obrazów o rozdzielczości 32x32 pikseli, przedstawiających elementy należące do zbioru {Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, Truck} [12].

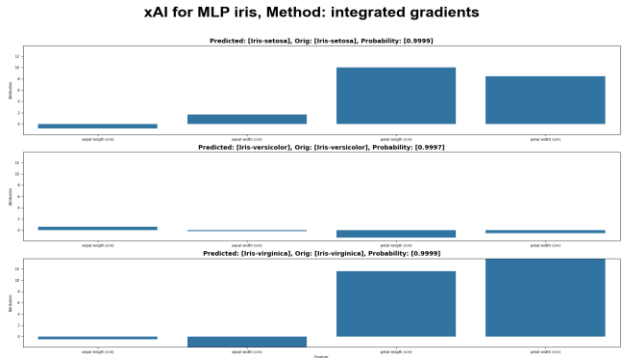
B. *Model MLP dla zbioru danych Iris*

Mapa ważności dla zbioru danych Iris wykazała, że dominującymi cechami wpływającymi na decyzję modelu dla irysów setosa oraz versicolor są długość płatka (*ang. petal length*) oraz szerokość płatka (ang. *petal width*), a dla virginica dominującymi cechami są długość płatka, szerokość płatka oraz szerokość kielicha, która przyczynia się do obniżenia wartości predykcji, jak przedstawiono na Rysunek 1.



Rysunek 1 Mapa ważności dla zbioru danych Iris, model MLP.

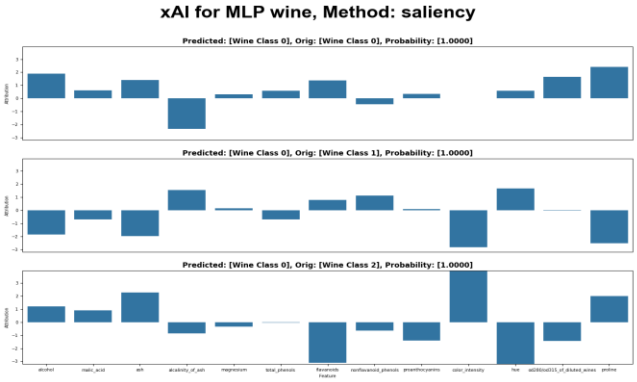
W przypadku zastosowania metody scałkowanych gradientów ważnymi cechami dla wszystkich trzech rodzajów irysów wpływającymi na decyzję modelu okazały się być długość płatka oraz szerokość płatka, chociaż w przypadku irysa versicolor z nieznaczną przewagą nad długością kielicha. Wyniki eksperymentu przedstawiono na Rysunek 2.



Rysunek 2 Zcałkowane gradienty dla zbioru danych Iris, model MLP.

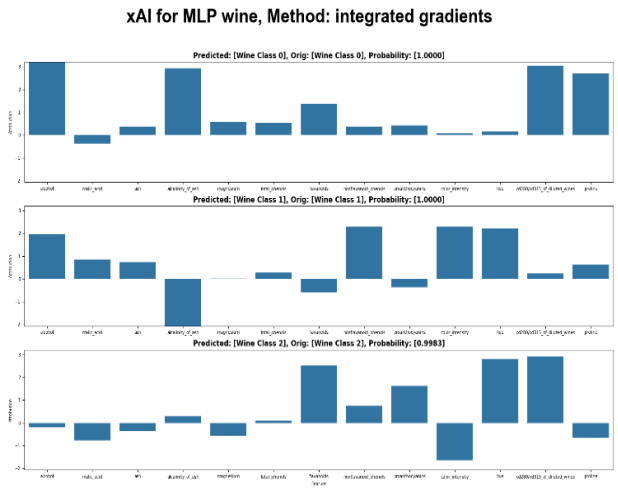
C. *Model MLP dla zbioru danych Wine*

Mapa ważności dla zbioru danych Wine wykazała, że wyróżniającymi się spośród wszystkich cech pod względem wpływu na decyzję modelu są {prolina, alkohol, flawonoidy}, które przyczyniają się do zwiększenia wartości predykcji oraz zasadowość, która znacząco wpływa na obniżenie wartości predykcji modelu w przypadku pierwszego rodzaju wina, {zasadowość, barwa}, które wpływały pozytywnie na wartość predykcji modelu oraz {intensywność koloru, prolina}, które obniżały wartość predykcji modelu w przypadku drugiego rodzaju wina oraz {intensywność koloru, zasadowość}, które zwiększały wartość predykcji i {barwa, flawonoidy}, które obniżały wartość predykcji modelu w przypadku trzeciego rodzaju wina. Zauważyć również na Rysunek 3 można fakt, że w przypadku trzeciego rodzaju wina model zwraca szczególną uwagę na jego intensywność koloru, flawonoidy i barwę, gdzie w pozostałych dwóch rodzajach nie zauważono tak dominujących cech.



Rysunek 3 Mapa ważności dla zbioru danych Wine, model MLP.

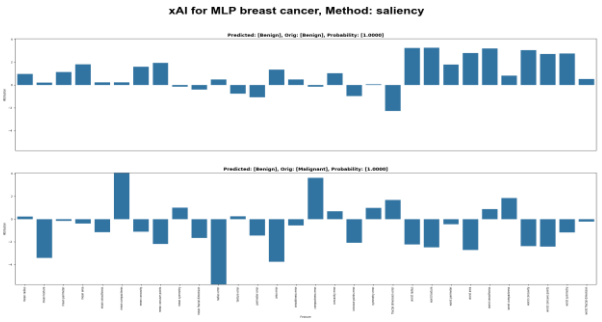
W przypadku zastosowania metody scałkowanych gradientów ważnymi na podobnym poziomie cechami dla pierwszego rodzaju wina okazały się {alkohol, zasadowość, absorbancja światła, prolina}, dla drugiego rodzaju wina ważnymi cechami okazały się być {intensywność koloru, barwa, związki nie należące do grupy flawonoidów, alkohol}, a cechą wyraźnie przyczyniającą się do obniżenia wartości predykcji modelu była zasadowość. Dla trzeciego rodzaju wina najważniejszymi cechami okazały się być {barwa, absorbancja światła, flawonoidy}. Co ciekawe, cechą znacznie obniżającą wartość predykcji modelu okazała się być intensywność koloru, która w metodzie map ważności była cechą dominującą.



Rysunek 4 Zcałkowane gradienty dla zbioru danych Wine, model MLP.

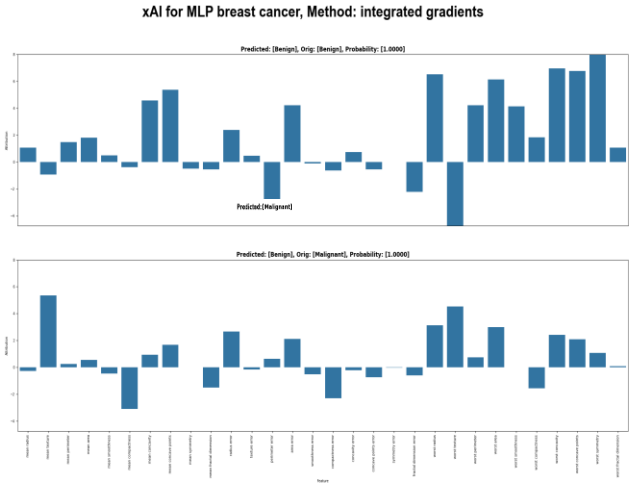
D. *Model MLP dla zbioru danych Breast Cancer*

Mapa ważności dla zbioru danych Breast Cancer wykazała, że wyróżniającymi się spośród wszystkich cech pod względem dodatniego wpływu na predykcję modelu dla łagodnego raka piersi były największy promień oraz cechy odpowiadające za najgorsze wartości promienia, tekstury, obwodu, powierzchni, gładkości, zwartości, wypukłości, wypukłości punktowej, symetrii, wymiaru fraktalnego, a wpływającymi negatywnie na wartość predykcji modelu były błędy standardowe tekstury, obwodu i wymiaru fraktalnego oraz wymiar fraktalny guza. W przypadku złośliwego raka piersi, wyróżniającymi się cechami pod względem dodatniego wpływu na predykcję modelu okazały się być {zwartość, błąd standardowy powierzchni}, a pod względem negatywnego wpływu na predykcję modelu {symetria guza, średnica guza, błąd standardowy tekstury}. Analizę przedstawiono na Rysunek 5.



Rysunek 5 Mapa ważności dla zbioru danych Breast Cancer, model MLP.

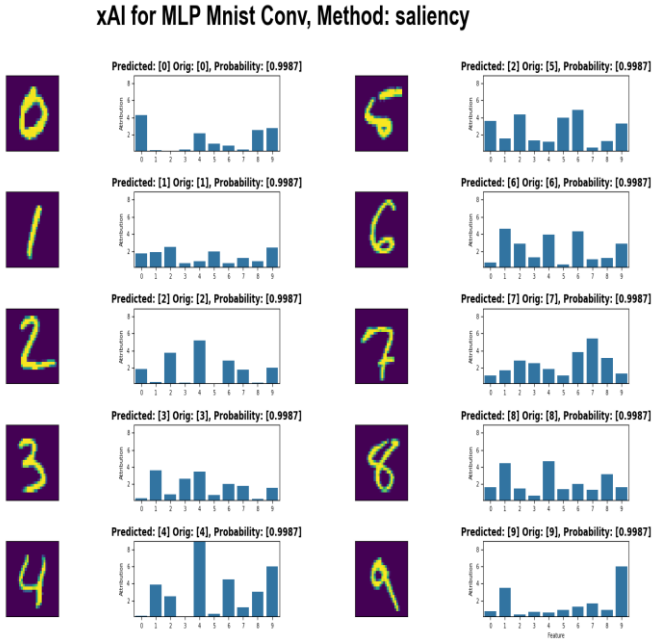
W przypadku zastosowania metody scałkowanych gradientów, dla łagodnego raka piersi cechami mającymi największy dodatni wpływ na predykcję modelu okazały się być te same co dla metody map ważności, za wyjątkiem największego promienia, który w znaczący sposób wpływał na ujemną wartość predykcji modelu, a w przypadku złośliwego raka piersi cechami pozytywnie wpływającymi na predykcję były {różnica guza, najgorsza tekstura, największy promień}, a negatywnie wpływającymi na wartość predykcji modelu były {zwartość, błąd standardowy zwartości}. Można zauważyć, że w przypadku map ważności różnica guza dla złośliwego raka piersi wpływa pozytywnie na wartość predykcji modelu, a w przypadku scałkowanych gradientów negatywnie. Analizę przedstawiono na Rysunek 6.



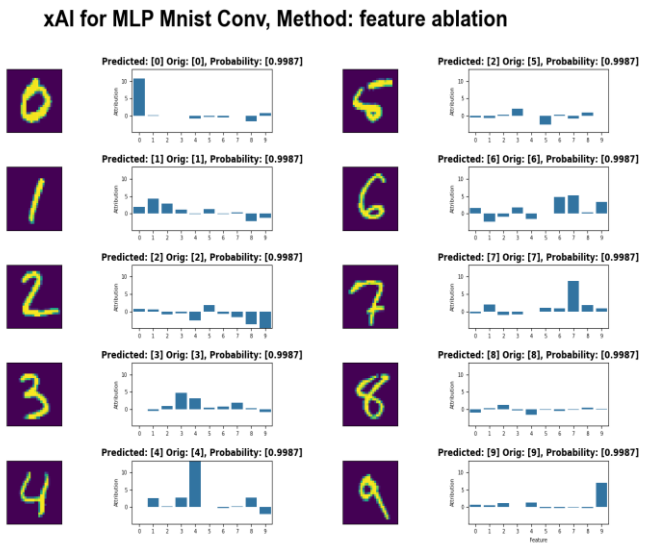
Rysunek 6 Zcałkowane gradienty dla zbioru danych Breast Cancer, model MLP.

E. *Model MLP dla zbioru danych MNIST z ekstrakcją cech w postaci masek konwolucyjnych średnich cyfr*

Zbiór danych treningowych MNIST został przefiltrowany względem rzeczywistych klas a następnie wygenerowano 10 masek poprzez sumowanie obrazów o zgodnej etykiecie i normalizację wyników, tak by elementy każdej maski sumowały się do jedności. Następnie każdy punkt danych całego zbioru MNIST zostaje przypisany do wektora 10 nowych cech, poprzez normalizację jasności pikseli by sumowały się do jedności i następnie konwolucję z wektorem masek. Dane na koniec są normalizowane dla każdej cechy poprzez przeskalowanie na rozkład normalny. Analizę przeprowadzonego eksperymentu objaśnienia metodami ablacji cech i mapy ważności przedstawiono na Rysunkach 7 i 8.



Rysunek 7 Mapa ważności dla zbioru danych MNIST, model MLP.

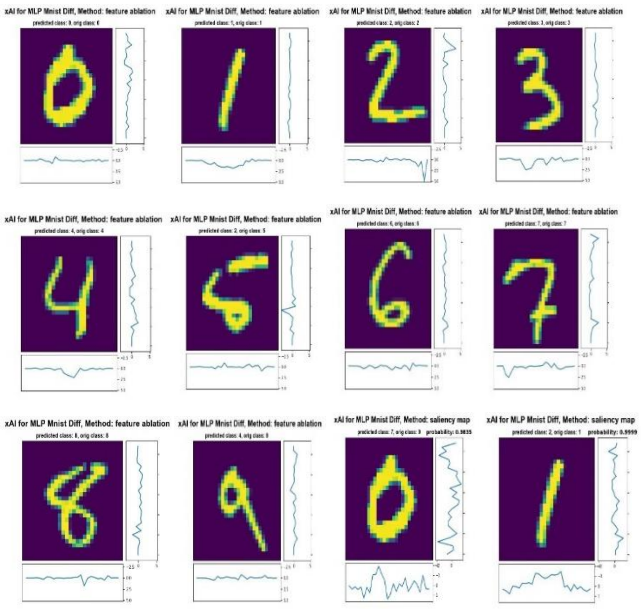


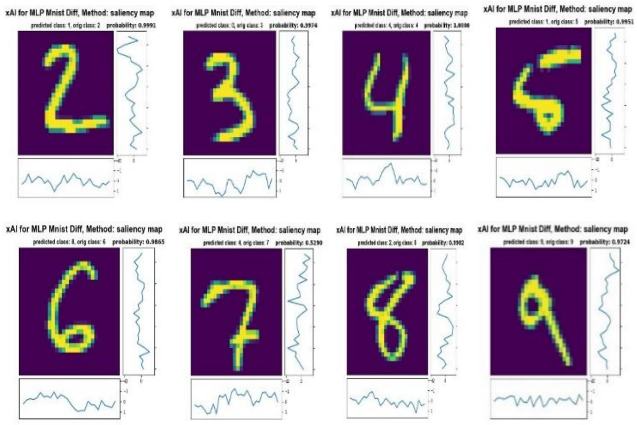
Rysunek 8 Ablacja cech dla zbioru danych MNIST, model MLP.

Powyższa metoda ekstrakcji przygotowana została by sprawdzać podobieństwo testowanego obrazu z jego średnią reprezentacją, więc wynik gdzie ponad połowa cyfr nie wykazuje widocznej preferencji dla swojej średniej jest zaskakujący. W sygnaturach cyfr dostrzec można mimo tego charakterystyczne różnice, a ich istnieniu dowodzi niemal perfekcyjna dokładność badanego klasyfikatora. Jest to dobra ilustracja tego, że model potrafi dostrzec powiązania nie przewidziane przez projektanta.

F. Model MLP dla zbioru danych MNIST z ekstrakcją cech w postaci zliczania krawędzi

Zbiór danych MNIST poddany został progowaniu, a następnie zliczono zmiany kolorów wzdłuż wszystkich odcinków pikseli szerokości i wysokości. Otrzymano 2x28 cech, które następnie są normalizowane dla wszystkich danych. Dzięki normalizacji wektory odniesienia do metod je wykorzystujących mogą przyjąć postać wektorów zerowych, co odzwierciedla średnią ilość krawędzi na danym odcinku. Do analizy zastosowano metody ablacji cech i mapy ważności.



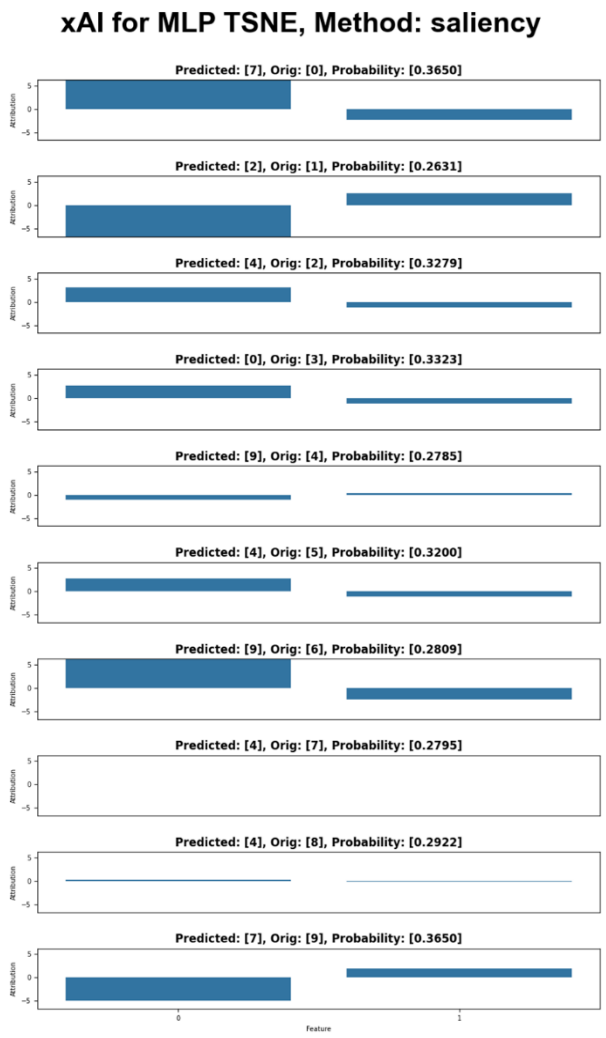


Rysunek 9 Wartości map ważności oraz ablacji cech dla zbioru danych MNIST z ekstrakcją cech w postaci zliczania krawędzi, model MLP.

Powyższe wykresy ilustrują sąsiadujące atrybucje dla kolumn i wierszy obrazu, np. pierwszy punkt od lewej na wykresie dołączonym pod obrazem odpowiada atrybucji znormalizowanej ilości krawędzi w pierwszej od lewej kolumnie obrazu. Do wychodzących wbrew intuicji zachowań modelu należą między innymi wyniki dla cyfr 1 i 8. Ósemka jest charakterystyczna, ponieważ jako jedyna posiada dwie zamknięte pętle, jednak definiujący je środek wpływa negatywnie na klasyfikację. Podobnie jedynka, która jako jedyna posiada średnio dwie krawędzie na całym zakresie otrzymuje negatywny wpływ ich ważności, choć wskaźnik ablacji cech daje wyniki pokrywające się z intuicją. Trend ten jest kontynuowany w innych cyfrach, np. rozpoznanie niedomkniętej pętli na szycie dwójki i wystającą do prawej podstawę, gdy mapa ważności przedstawia dość chaotyczne relacje.

G. Model MLP dla zbioru danych MNIST z ekstrakcją cech wykorzystującą algorytm redukcji wymiarów t-SNE

Do danych treningowych zbioru MNIST po uprzednim spłaszczeniu oraz normalizacji zastosowano algorytm t-SNE z biblioteki scikit-learn, który metodami probabilistycznymi generuje mapy do przestrzeni o zredukowanych wymiarach. Następnie wybierana jest ta, która maksymalizuje ilość obrazów punktów zawierających w nowym otoczeniu obrazy punktów z pierwotnego otocznia, by jak najlepiej odzwierciedlić rozłożenie punktów w przestrzeni. Uzyskaną mapą ekstrakcji poddawane są dane testowe zbioru MNIST.

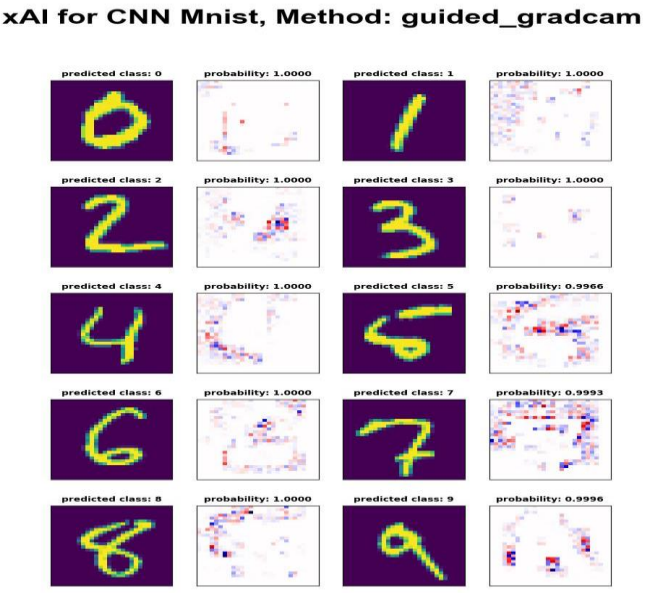


Rysunek 10 Mapa ważności dla zbioru danych MNIST z ekstrakcją cech za pomocą algorytmu t-SNE, model MLP.

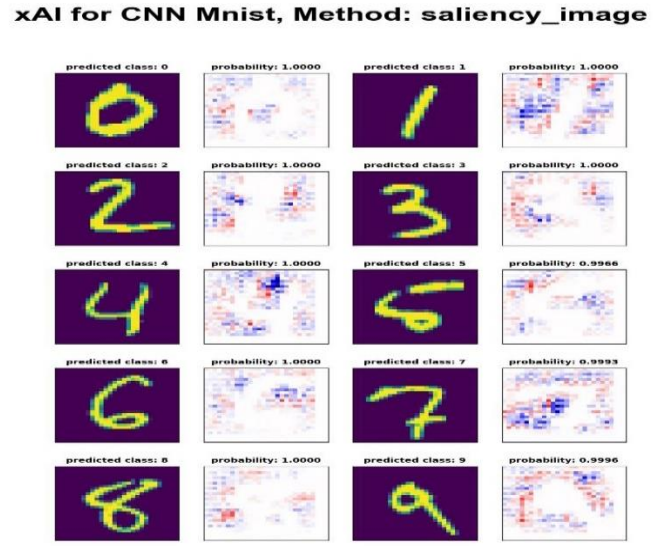
W przypadku zastosowania algorytmu t-SNE odpowiadającego za redukcję wymiaru w celu uzyskania dwóch cech dla zbioru danych MNIST, z których to następnie model MLP miał przewidzieć cyfrę. Cecha 0 wpływa, z różną intensywnością, na dodatnią wartość predykcji modelu w przypadku cyfr {0, 2, 3, 6,}, a na ujemną wartość predykcji w przypadku cyfr {1, 9}. Cecha 1 wpływa, z różną intensywnością, na dodatnią wartość predykcji modelu w przypadku cyfr {1,9}, a na ujemną w przypadku {0, 2, 3, 6}. W przypadku cyfr {4, 7, 8} metoda map ważności nie wskazała istotnie wpływających na predykcję modelu cech, co oznacza, że trudno jest ocenić, dlaczego model dokonuje klasyfikacji tych cyfr na podstawie cech 0 i 1. Uogólniając, dwie cechy w przypadku klasyfikacji do 10 klas, nie są wystarczające do poprawnego działania modelu, o czym świadczy dokładność na poziomie 40%.

H. Model CNN dla zbioru danych MNIST

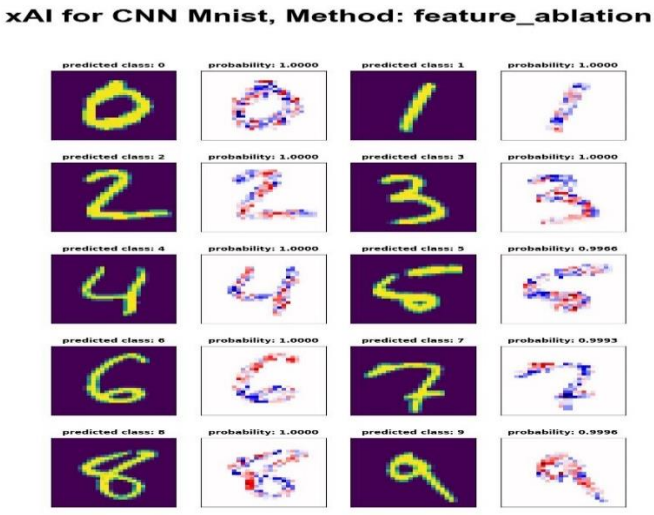
Zastosowanymi metodami przy objaśnianiu działania modelu CNN wytrenowanego na zbiorze danych MNIST były: mapa ważności, GG-CAM, ablacja cech oraz wartości Shapley'a. Rezultaty przeprowadzonych eksperymentów przedstawiono na obrazach Rysunek 10, Rysunek 11, Rysunek 12 oraz Rysunek 13.



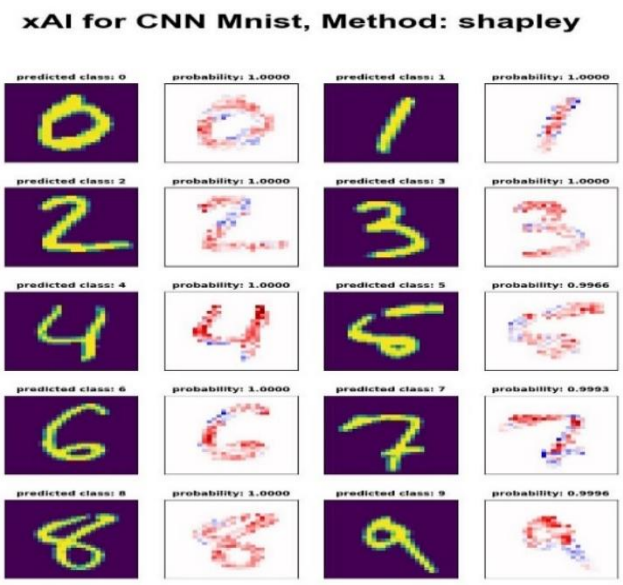
Rysunek 10 11 GradCAM dla zbioru danych MNIST, model CNN.



Rysunek 11 12 Mapa ważności dla zbioru danych MNIST, model CNN.



Rysunek 12 13 Ablacja cech dla zbioru danych MNIST, model CNN.

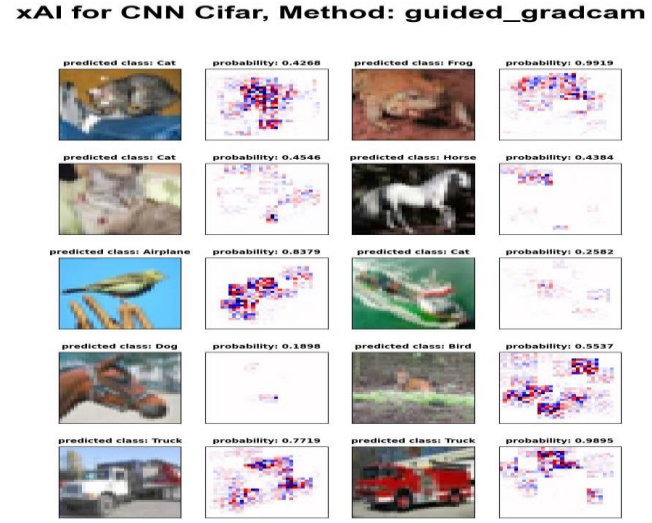


Rysunek 13 14 Wartości Shapley'a dla zbioru danych MNIST, model CNN.

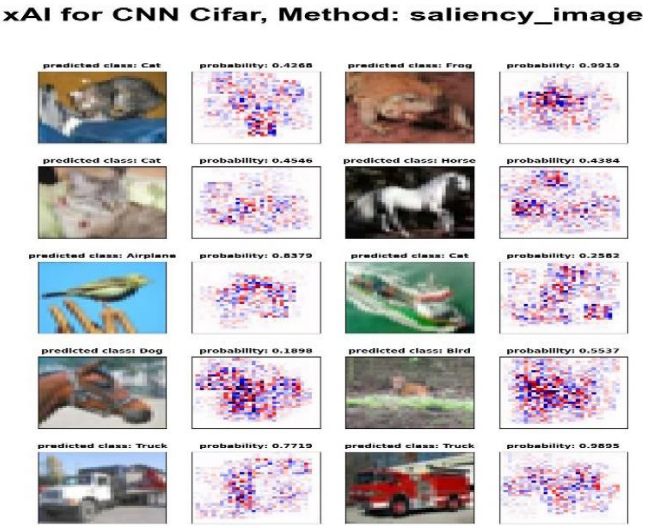
Powyższe wyniki podzielić można na dwie kategorie, mapujące piksele należące do rozpoznawanego obiektu i te które należą do otoczenia. Wartości Shapley'a i ablacja cech naturalnie nie przypisze tłu żadnej wagi, ponieważ zastosowana wartość odniesienia była obrazem o jednolitym kolorze tła – wymiana tych cech nie daje więc żadnego efektu na wynik. Przeciwny efekt zaobserwować można w GG-CAM i mapie ważności, do której szczególnie interesujących wyników należą cyfry 1 i 4. Łącząc ciemnoniebieskie obszary z białymi otrzymać można inne cyfry, kolejno 7 i 9, co jest wynikiem pożądanym, ponieważ pokazuje silny wpływ dodania cech na zmianę wyniku klasyfikacji.

I. Model CNN dla zbioru danych CIFAR10

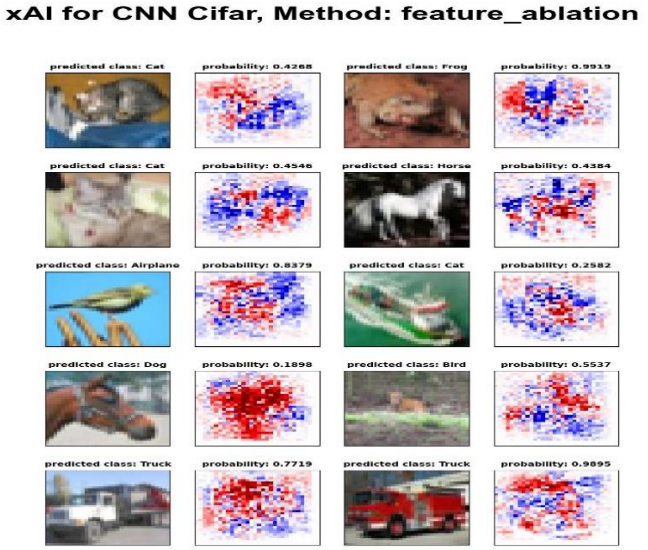
Zastosowanymi metodami modelu CNN wytrenowanego na zbiorze danych CIFAR10 były: mapa ważności, naprowadzany GG-CAM, ablacja cech oraz wartości Shapley'a. Rezultaty przeprowadzonych eksperymentów przedstawiono na obrazach Rysunek 14, Rysunek 15, Rysunek 16, Rysunek 17.



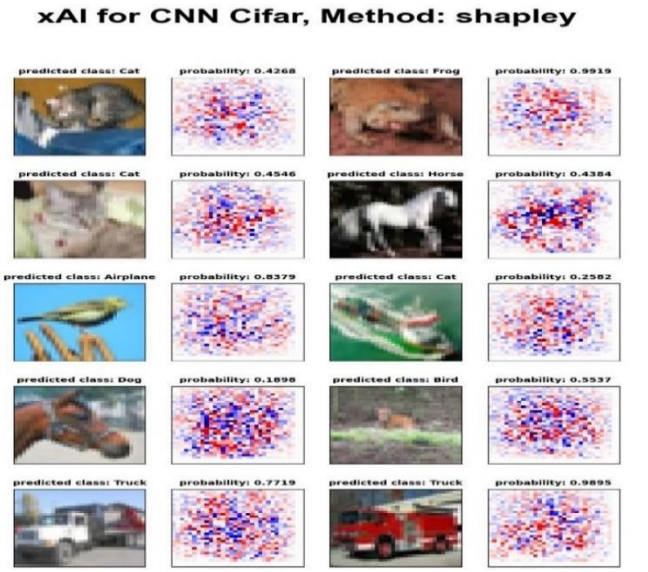
Rysunek 14 15 GG-CAM dla zbioru danych CIFAR10, model CNN.



Rysunek 15 16 Mapa ważności dla zbioru danych CIFAR10, model CNN.



Rysunek 16 17 Mapa ważności dla zbioru danych CIFAR10, model CNN.



Rysunek 17 18 Wartości Shapley'a dla zbioru danych CIFAR10, model CNN.

W przeciwieństwie do modelu wytrenowanego na zbiorze danych MNIST, zastosowanie tych samych metod objaśnienia na modelu wytrenowanym na zbiorze danych CIFAR10 zwraca rozproszone po całym obrazie obszary wpływające na zmianę predykcji modelu. Ocena takiej wizualizacji jest zatem niepraktyczna, co dobrze ilustruje Rysunek 17, gdzie dla każdego przykładu widać rozproszone plamy punktów, które dla człowieka wyglądają niemal identycznie, czy przykład obrazu konia na Rysunek 16, gdzie usunięcie pikseli reprezentujących konia na obrazie spowoduje zmianę jego błędnej predykcji jako psa. Potencjalnym powodem otrzymania takich wyników jest niska dokładność modelu, a ich chaotyczność może odzwierciedlać brak umiejętności modelu do generalizacji.

PODSUMOWANIE

Zastosowane w przeprowadzonych eksperymentach metody objaśniania modeli uczenia maszynowego z modułu Captum, realizowane przy dziedzinie Interpretowalnej Sztucznej Inteligencji, pomagają nam zrozumieć wpływy konkretnych cech na wyniki predykcji naszych modeli. Umożliwiają nam obserwowanie sytuacji, w których model myli się w klasyfikacji, przez co możemy wyciągnąć odpowiednie wnioski w celu jego ewentualnej poprawy, czego dobrym przykładem jest model CNN wytrenowany do rozpoznawania obrazów ze zbioru CIFAR 10.

Porównanie dokładności analizowanych modeli uczenia maszynowego przedstawiono na tabeli TABLE 1.

TABELA 1. DOKŁADNOŚCI ANALIZOWANYCH MODELI

Dane	**Model i sposób ekstrakcji cech**	**Dokładność [%]**
Iris	MLP, cechy ze zbioru	100
Wine	MLP, cechy ze zbioru	100
Breast Cancer Wisconsin	MLP, cechy ze zbioru	97
MNIST	MLP, konwolucyjne maski ze średnimi cyframi	93
MNIST	MLP, wykrywanie krawędzi	89
MNIST	MLP, redukcja wymiaru spłaszczonego obrazu do 2 cech za pomocą algorytmu t-SNE	40
MNIST	CNN	100
CIFAR10	CNN	72

SPIS LITERATURY

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, i D. Batra, „Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, Georgia Institute of Technology, Atlanta, GA, USA, ostatni dostęp: czerwiec 2024.

[2] K. Simonyan, A. Vedaldi, i A. Zisserman, „Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, Visual Geometry Group, University of Oxford, ostatni dostęp: czerwiec 2024.

[3] S. M. Lundberg i S. Lee, „A Unified Approach to Interpreting Model Predictions”, University of Washington, Seattle, WA, USA, ostatni dostęp: czerwiec 2024.

[4] Captum, [Online]. Dostępne: https://captum.ai. [Dostęp: czerwiec 2024].

[5] K. Simonyan, A. Vedaldi, i A. Zisserman, „Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, Visual Geometry Group, University of Oxford, ostatni dostęp: czerwiec 2024.

[6] M. Sundararajan, A. Taly, i Q. Yan, "Axiomatic Attribution for Deep Networks," p. 3.

[7] Captum, „FeatureAblation”, [Online]. Dostępne: https://captum.ai/api/_modules/captum/attr/_core/feature_ablation.html#FeatureAblation. [Dostęp: czerwiec 2024].

[8] Scikit-learn, „sklearn.datasets”, [Online]. Dostępne: https://scikit-learn.org/stable/api/sklearn.datasets.html#module-sklearn.datasets. [Dostęp: czerwiec 2024].

[9] Wikipedia, „List of datasets for machine-learning research”, [Online]. Dostępne: https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research. [Dostęp: czerwiec 2024].

[10] Kaggle, „MNIST Dataset”, [Online]. Dostępne: https://www.kaggle.com/datasets/hojjatk/mnist-dataset. [Dostęp: czerwiec 2024].

[11] PyTorch, „Datasets”, [Online]. Dostępne: https://pytorch.org/vision/main/datasets.html. [Dostęp: czerwiec 2024].

[12] W. Cukierski, „CIFAR-10 - Object Recognition in Images”, Kaggle, 2013. [Online]. Dostępne: https://kaggle.com/competitions/cifar-10. [Dostęp: czerwiec 2024].

SPIS ILUSTRACJI