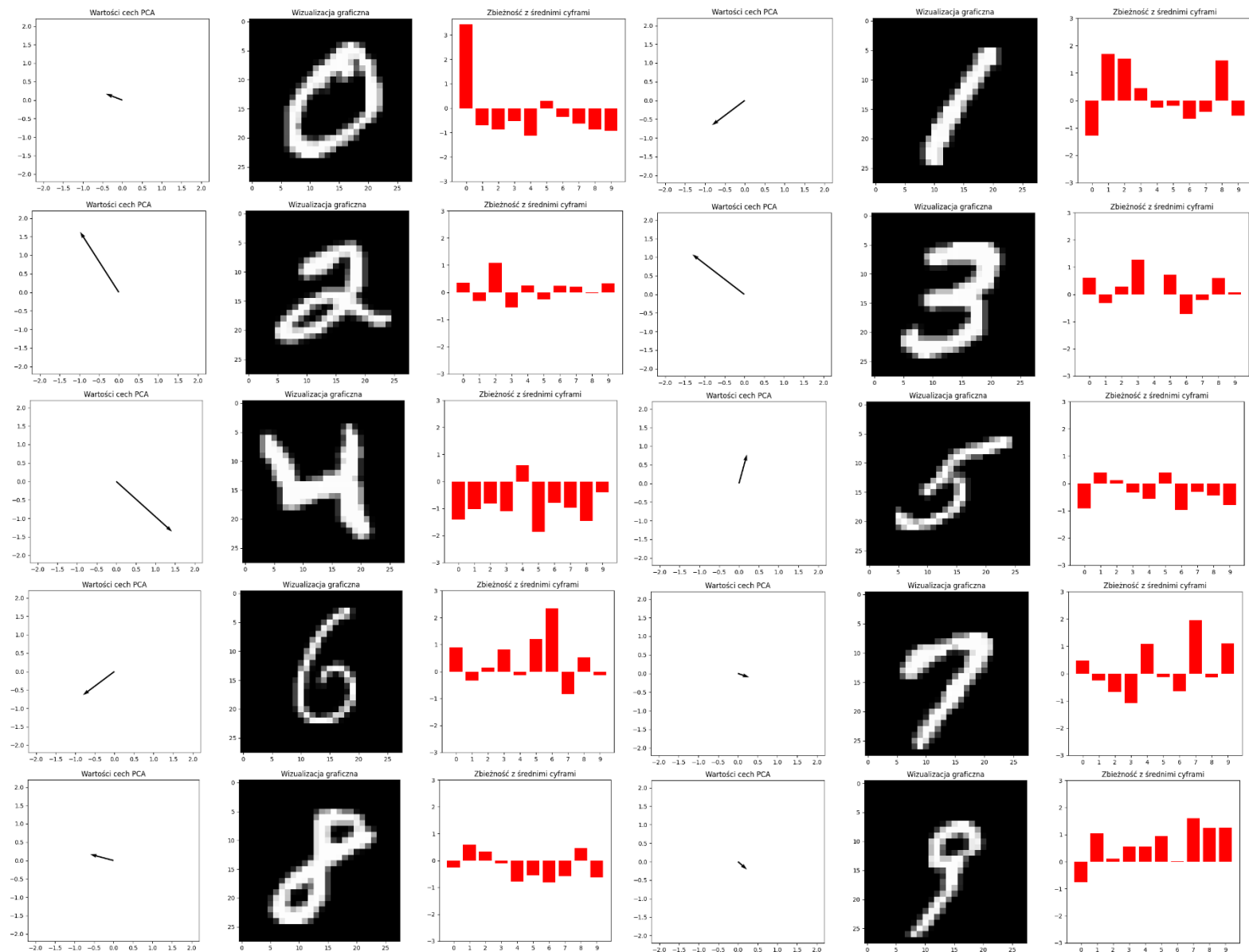


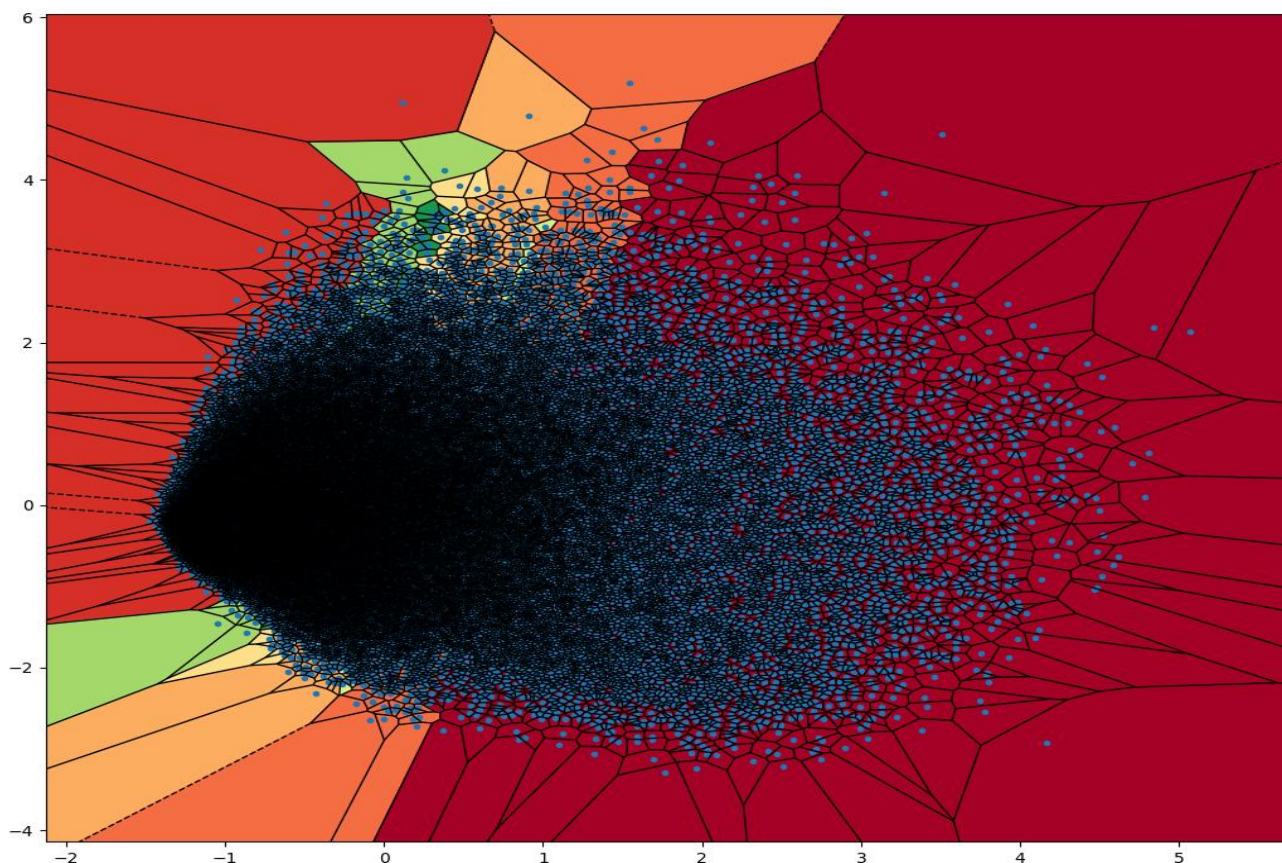
1.1 Opis dwóch sposobów ekstrakcji: algorytm PCA i konwolucja z maskami średnich cyfr



Pierwszy algorytm ekstrakcji, PCA (Principal Component Analysis) z biblioteki scikit learn przypisuje uprzednio znormalizowanemu zbiorowi danych $A_{n,d}$ zredukowany zbiór danych $B_{n,r}$, gdzie n – ilość punktów danych, d – ilość oryginalnych cech, r – żądana ilość zredukowanych cech poprzez scentralizowanie wszystkich cech względem siebie a następnie odnalezienie dekompozycji macierzy $A_{n,d} = USV^*$, gdzie U i V są ortonormalne a S diagonalna nieujemna i zdefiniowanie $B_{n,r} = U\hat{S}V^*$, gdzie \hat{S} jest równe S wszędzie, poza $(d-r)$ najmniejszymi wartościami macierzy S , którym przypisane jest 0. Algorytm ten realizuje zadanie minimalizacji normy Frobeniusa różnicy macierzy $A_{n,d}$ i $B_{n,r}$, przez co wynikowa macierz może być uważana jako aproksymacja macierzy oryginalnej. Dane na koniec są normalizowane dla każdej cechy. Metoda ta została wybrana ze względu na zdolności aproksymacyjne i dowolność wymiaru zredukowanej macierzy dzięki czemu może posłużyć do wizualizacji danych.

Drugi algorytm ekstrakcji wpierv generuje 10 mask na podstawie zbioru treningowego, poprzez sumowanie obrazów o zgodnej etykiecie dla całego zbioru, a następnie znormalizowanie ich przez podzielenie każdego elementu przez sumę wszystkich elementów maski do której jest przypisany. Następnie każdy punkt danych zostaje przypisany do wektora 10 cech, będących konwolucją kolejnych mask z punktem podzielonym przez sumę elementów punktu. Dane na koniec są normalizowane dla każdej cechy. Metoda ta została wybrana poprzez heurystyczne podejście do problemu klasyfikacji z myślą że cyfra powinna być najbardziej podobna do tych samych cyfr.

1.2 Wyniki eksperymentu pierwszego: algorytm PCA i konwolucja z maskami średnich cyfr



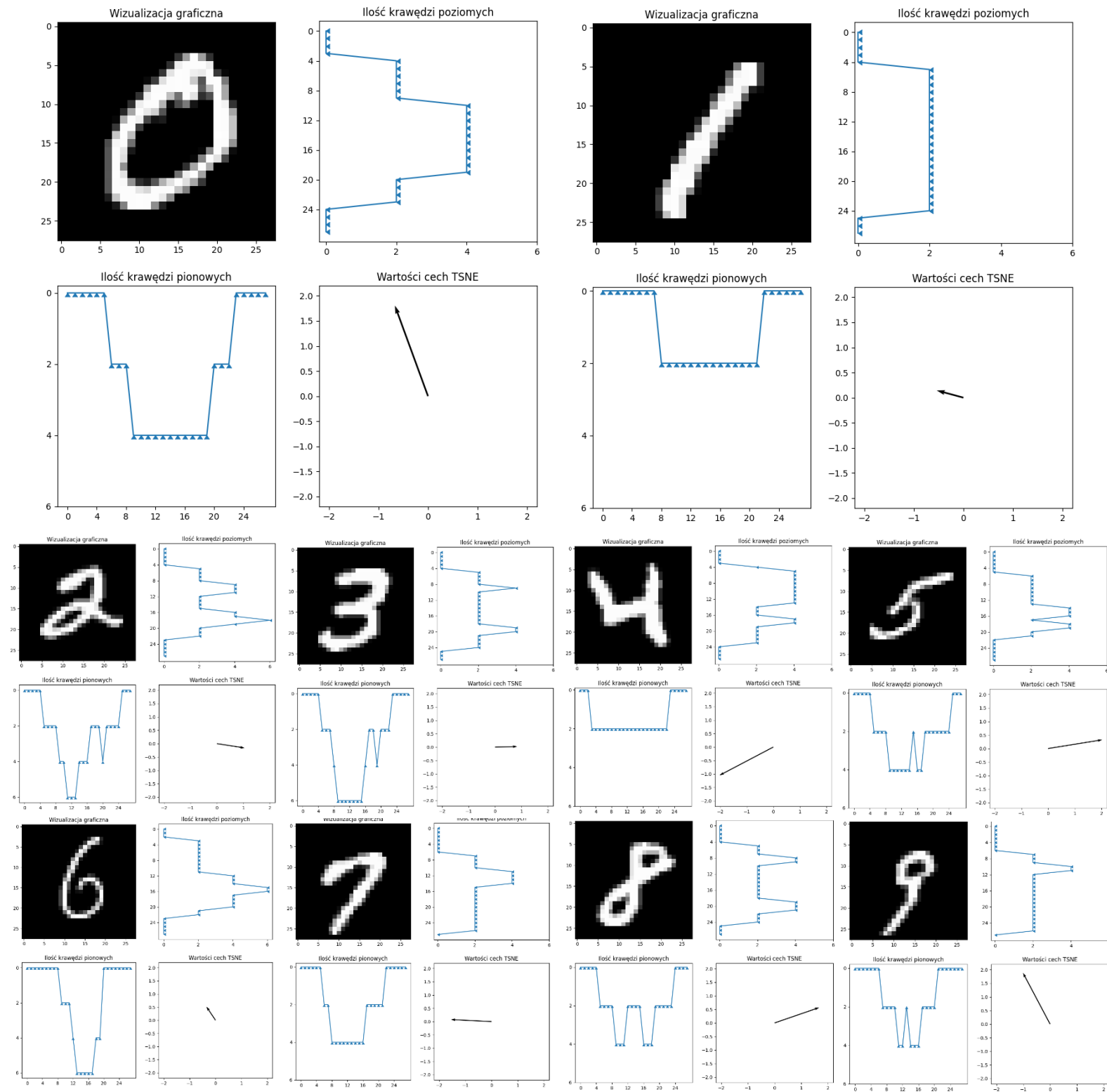
Na diagramie Voronoi'a przestrzeni ekstrahowanych cech metodą PCA i prawdziwych etykiet obserwujemy, że punkty skupione są w centrum o znacznej gęstości, która wraz z odległością spada z prędkością zależącą od kierunku. Po znaczącym stopniu "przemieszania" danych i ogólnym ich rozkładzie w przestrzeni wnioskujemy bardzo niską separowalność danych, co kwantyfikuje niska miara Silhouette Score równa -0.07.

Dla ekstrakcji poprzez zbieżność z średnimi cyframi dla zbioru treningowego wartość miary Silhouette (z wykorzystaniem oryginalnych etykiet) wyniosła 0.11. Sugeruje ona, na podstawie prawdziwych etykiet dla poszczególnych wektorów cech, że są one średnio separowalne – jednakże dobrze wyuczony i przygotowany model będzie w stanie sobie z takim problemem poradzić.

W przypadku metody PCA najczęściej mogą być mylone pary (7,8), (7,9), (2,3) - patrząc na kierunek zobrazowanych wektorów.

W przypadku poprzez zbieżność z średnimi cyframi najczęściej mogą być mylone pary (3,8), (7,9) - patrząc na zbieżności zobrazowane na wykresach słupkowych.

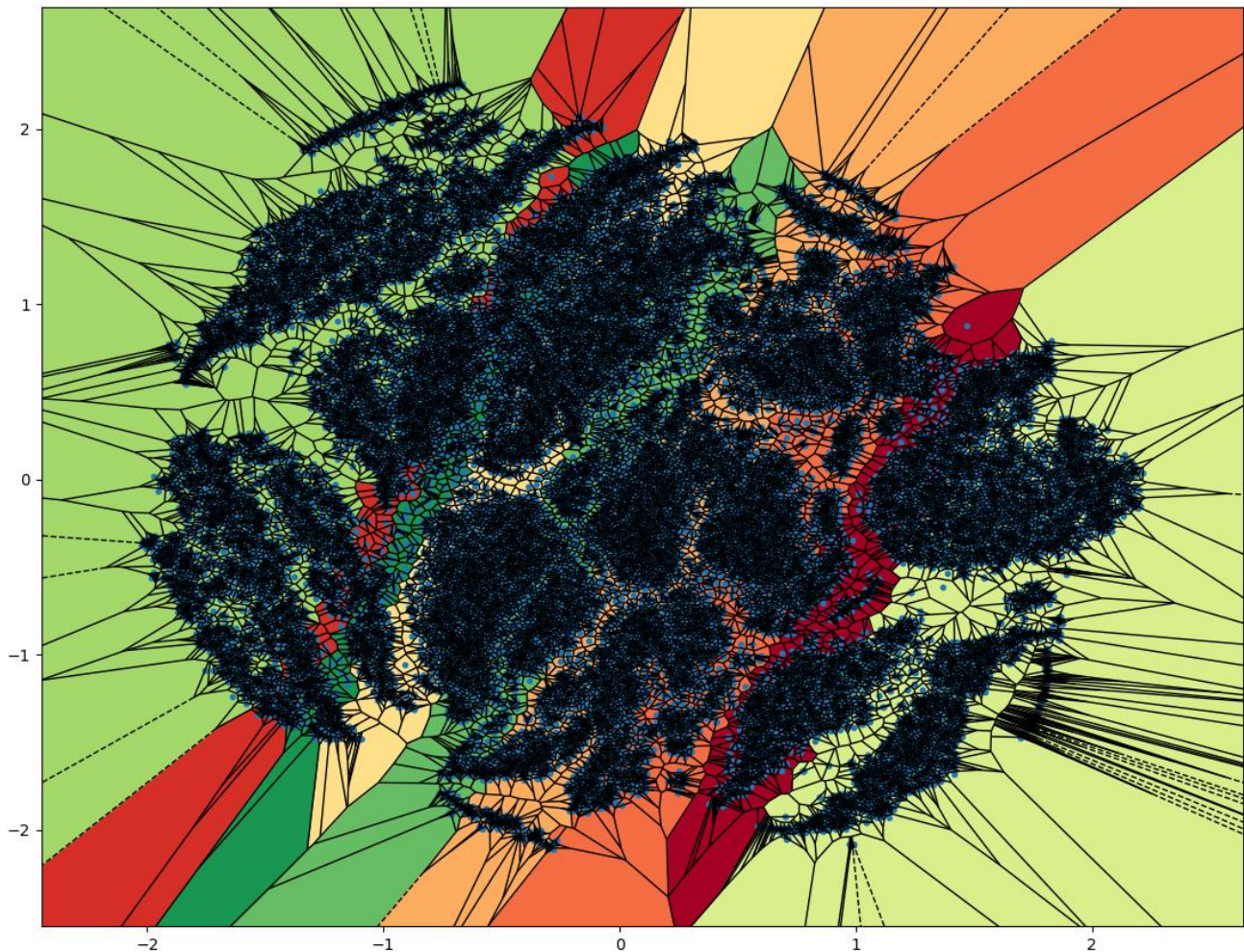
1.3 Opis dwóch sposobów ekstrakcji: wykrywanie krawędzi oraz algorytm TSNE



Pierwszą metodą ekstrakcji jest po uprzedniej normalizacji danych zastosowanie algorytmu t-SNE z biblioteki scikit learn który metodami probabilistycznymi generuje rozkłady w przestrzeni o zredukowanych wymiarach i wybiera ten, który maksymalizuje ilość punktów które zachowają swoje otoczenie z niezredukowanego wymiaru, tj jak najlepiej odzwierciedli odległości pomiędzy punktami. Metoda ta została wybrana ze względu na możliwość reprezentacji względnego grupowania danych.

Drugi algorytm ekstrakcji dla każdego punktu danych przeprowadza progowanie, a następnie zlicza zmiany koloru wzdłuż wszystkich odcinków szerokości i wysokości. Otrzymane 2x28 cechy następnie są normalizowane dla wszystkich danych. Metoda ta została wybrana ponieważ odzwierciedla istotne informacje geometryczne jednocześnie będące relatywnie łatwymi do odnalezienia.

1.4 Wyniki eksperymentu pierwszego: wykrywanie krawędzi oraz algorytm TSNE



W odróżnieniu od diagramu przy metodzie PCA, jesteśmy w stanie wizualnie ocenić kształty potencjalnych klastrów co wraz z rozmieszczeniem etykiet świadczy o obiecującej separowalności biorąc pod uwagę znaczący stopień utraty informacji związany z redukcją 784 wymiarów do 2. Odzwierciedla to wartość Silhouette Score równa 0.23 która nie jest wysoka, jednak znacząco wyższa od 0 reprezentującego bardzo niską separowalność.

Dla sposobu poprzez wykrywanie krawędzi dla zbioru treningowego wartość miary Silhouette (z wykorzystaniem oryginalnych etykiet) wyniosła 0.081. Wartość bliska zero sugeruje możliwość nakładania się na siebie klastrów. Separowalność dla tej metody oceniona została zatem na niską.

W przypadku metody t-SNE najczęściej mogą być mylone pary (3,5), (2,3) oraz (1,7) - patrząc na kierunek zobrazowanych wektorów.

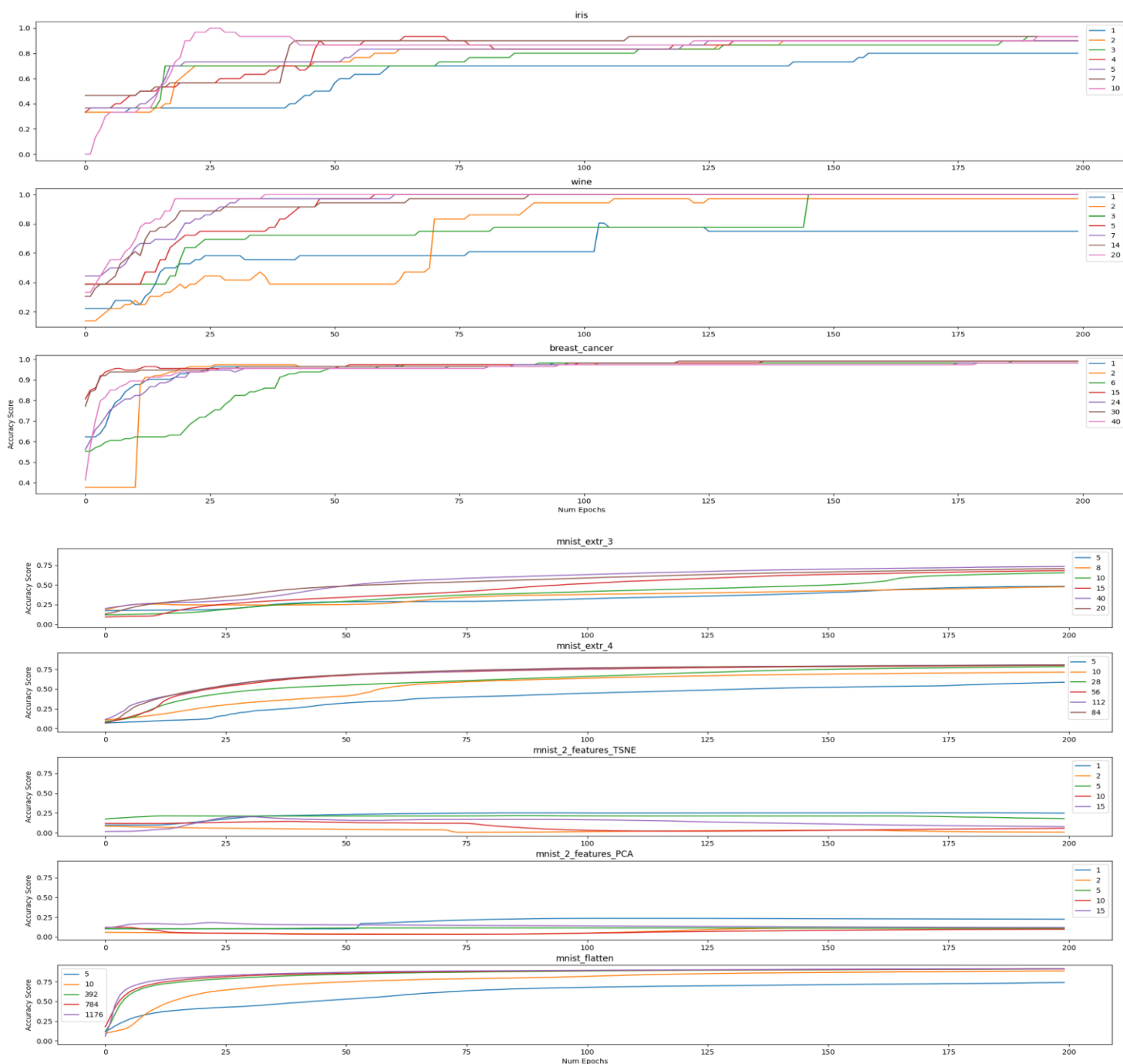
W przypadku metody ekstrakcji poprzez wykrywanie krawędzi najczęściej mogą być mylone pary (2,5), (7,9), (2,3) - patrząc na ilości wykrytych krawędzi w danym miejscu na obrazie.

1.5 Opis sposobu wyboru optymalnego modelu

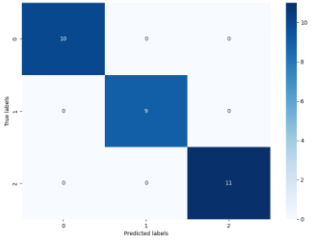
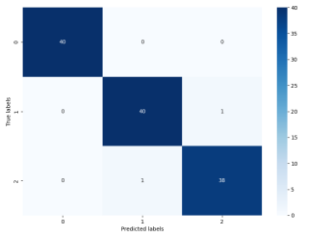
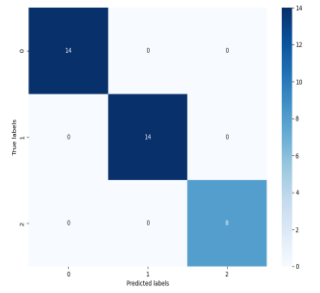
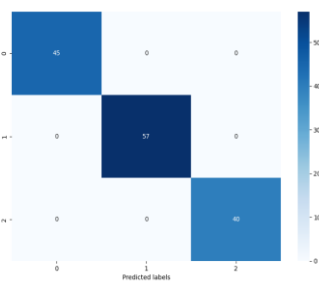
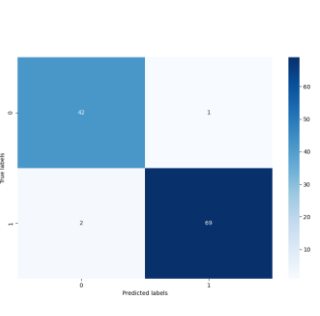
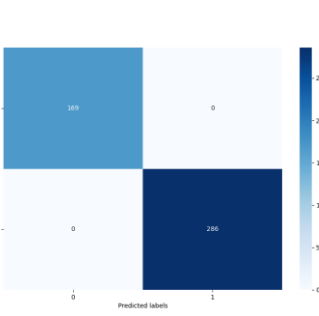
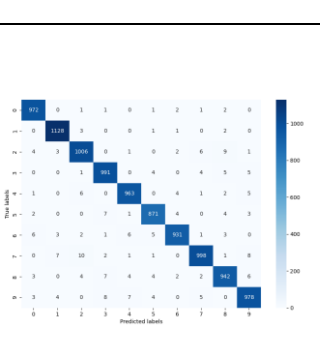
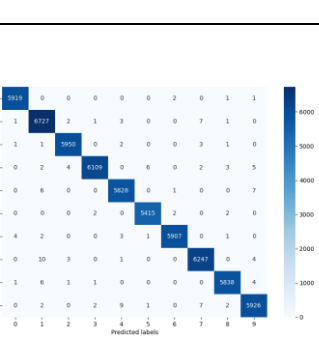
Model został zaprojektowany jako sieć neuronowa z dwoma warstwami liniowymi rozdzielonymi funkcją aktywacji ReLU. Dla warstw liniowych ilość neuronów ukrytych wyznaczono na podstawie trudności zadania klasyfikacji – wzięto pod uwagę liczbę klas do uzyskania oraz liczbę cech dla poszczególnego obiektu.

Do wyboru najlepszego modelu wykorzystano wartość *accuracy* wyliczaną na zbiorze treningowym dla różnych ilości neuronów ukrytych.

Finalnie liczba neuronów ukrytych zbiorów Iris, Wine oraz Breast Cancer została wyznaczona na wartość połowy ilości cech, jednakże z minimalną wartością nie mniejszą niż ilość klas możliwych do uzyskania. W przypadku metod PCA oraz t-SNE zastosowano kryterium maksymalizacji wartości *accuracy*, które wykazało, że najlepszą ilością neuronów jest '1'. Dla ekstrakcji cech ze zbioru MNIST polegającej na spłaszczeniu obrazu przyjęto ilość neuronów ukrytych wynoszącą rozmiar wektora (784). W przypadku ekstrakcji cech poprzez zastosowanie masek przyjęto ilość neuronów ukrytych '40' – taki model posiadał najlepszą wartość dokładności. W przypadku ekstrakcji cech poprzez sprawdzanie krawędzi przyjęto ilość neuronów wynoszącą $1.5 \times \text{ilość cech}$ - sprawdzono również ilość neuronów równą dwukrotności ilości cech, jednakże przyrost dokładności kosztem mocy obliczeniowej nie był opłacalny.



1.6 Opis wyników klasyfikacji dla trzech zbiorów danych oraz zbioru MNIST z pierwszym sposobem ekstrakcji

	TESTOWY		TRENINGOWY		Architektura		
Parametr	Acc	Confusion Matrix	Acc	Confusion Matrix	Ilość neuronów	Ilość cech	Ilość klas
IRIS	1,00		0,98		3	4	3
WINE	1,00		1,00		7	13	3
BREAST CANCER	0,97		1,00		15	30	2
MNIST (ekstrakcja poprzez spłaszczenie obrazu)	0,98		0,99		784	784	10

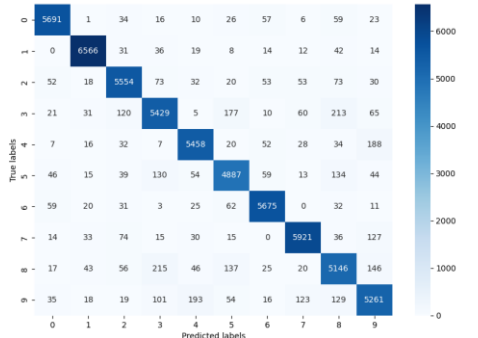
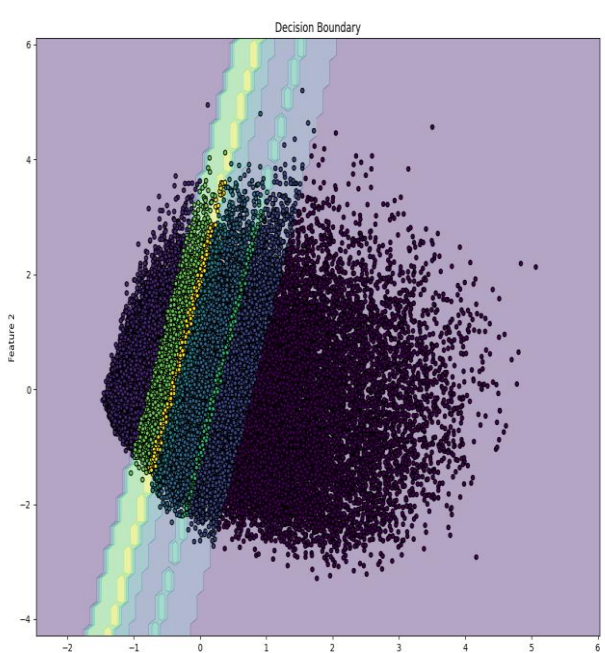
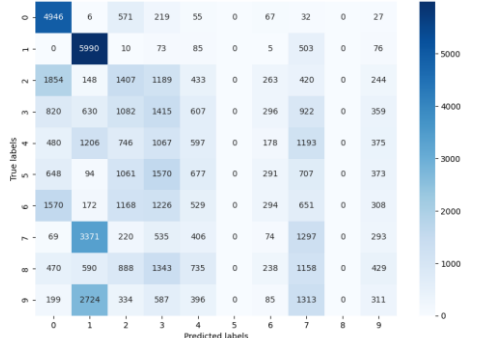
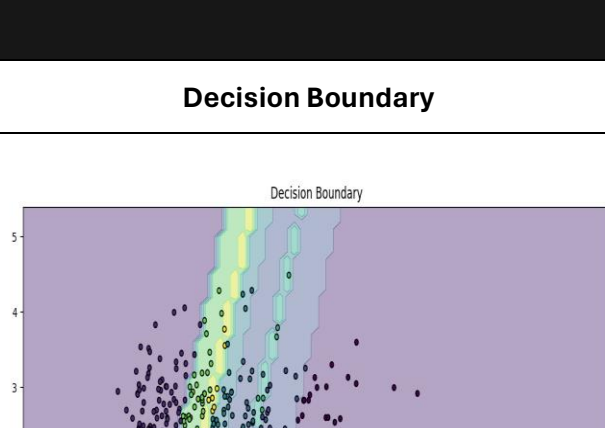
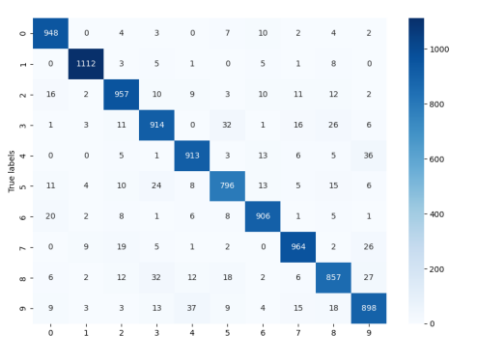
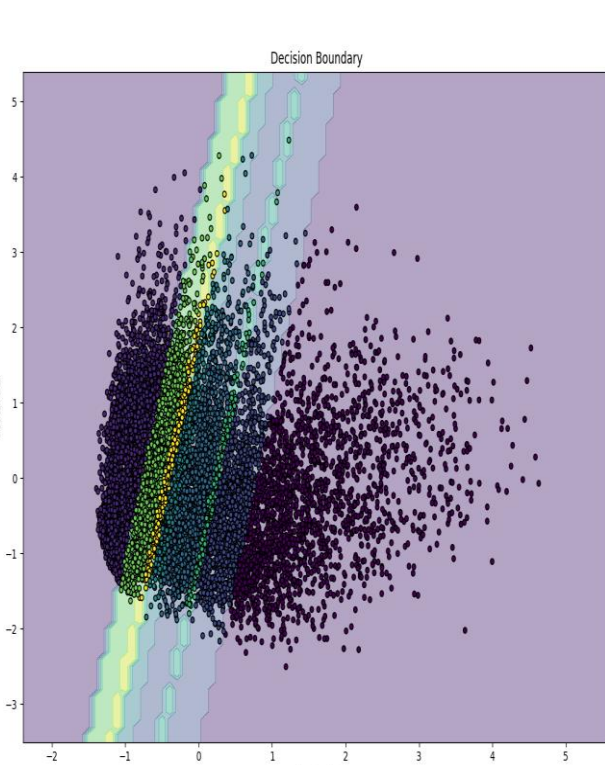
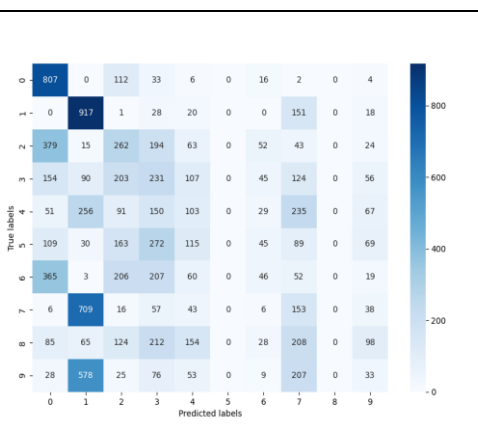
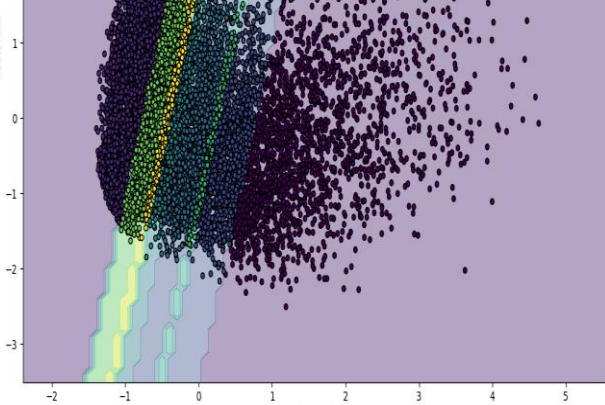
Dla zbioru Iris, posiadającego 3 cechy oraz 3 rodzaje możliwych klasyfikacji (3 różne rodzaje irysów w zbiorze) wybrano ilość neuronów wynoszącą 3.

Dla zbioru Wine, posiadającego 13 cech oraz 3 rodzaje możliwych klasyfikacji (3 rodzaje w zbiorze Wine) wybrano ilość neuronów wynoszącą 7.

Dla zbioru Breast Cancer, posiadającego 30 cech oraz 2 rodzaje możliwych klasyfikacji (2 rodzaje w zbiorze Breast Cancer) wybrano ilość neuronów wynoszącą 15.

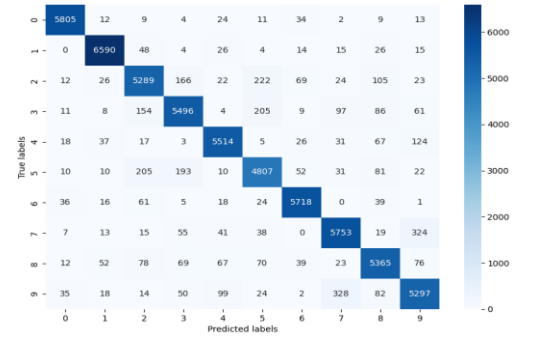
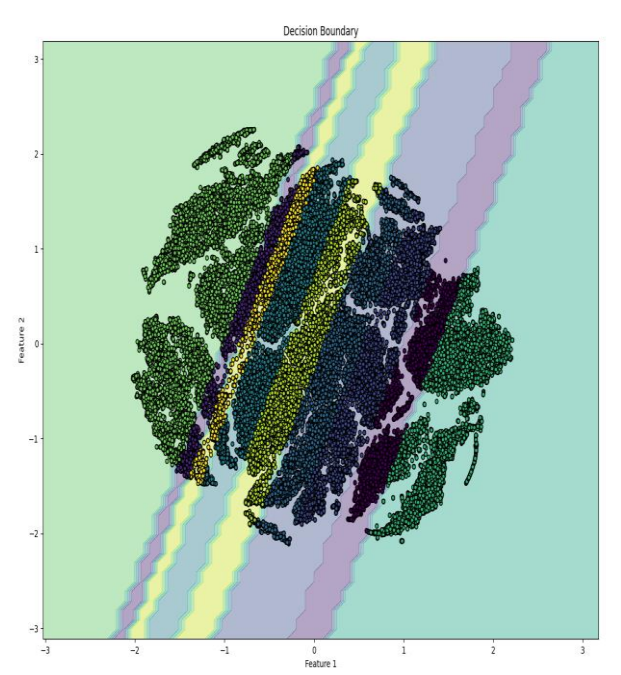
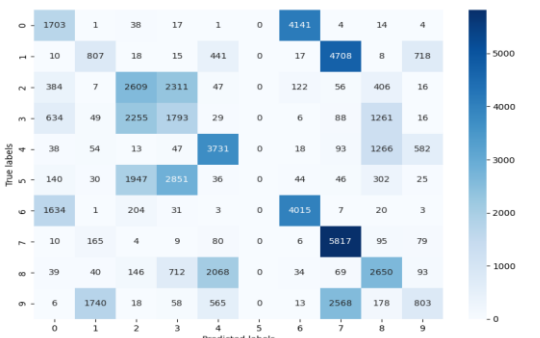
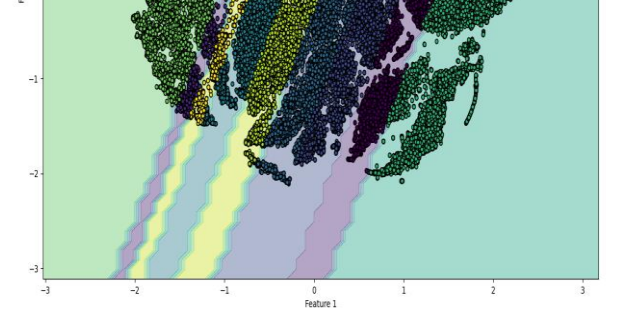
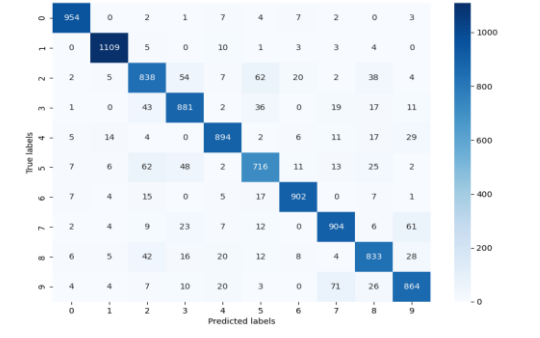
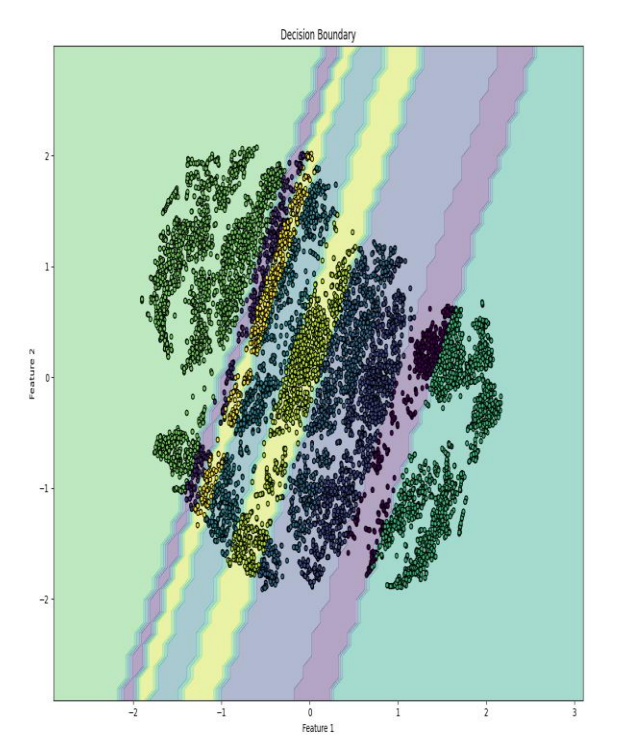
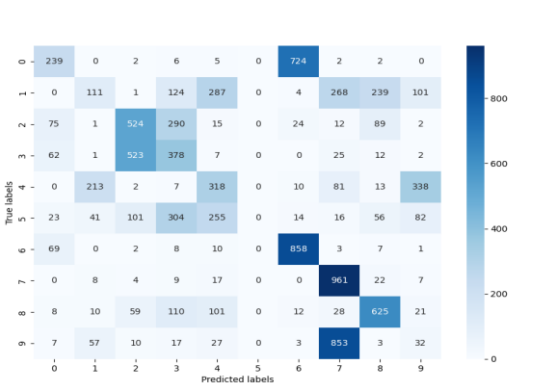
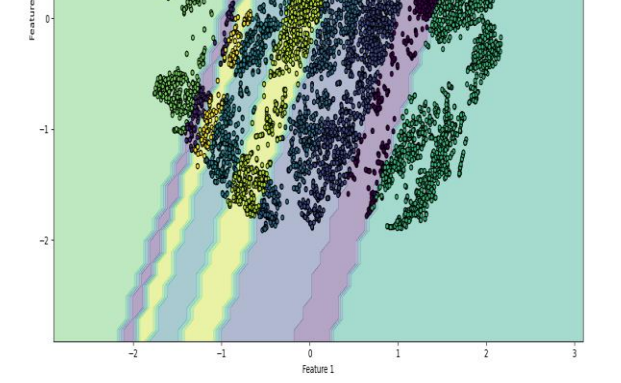
Dla zbioru Mnist, spłaszczonego za pomocą funkcji *flatten()* otrzymano 784 cechy, przy 10 możliwych klasyfikacjach (10 liczb 0-9) wybrano ilość neuronów wynoszącą 784.

1.7 Opis wyników klasyfikacji 'mnist_extr_3' oraz metody PCA

TRAIN	Accuracy Score	Confusion Matrix	Decision Boundary
konwolucja z maskami średnich cyfr	0,93		
PCA	0,27		
TEST	Accuracy Score	Confusion Matrix	Decision Boundary
konwolucja z maskami średnich cyfr	0,93		
PCA	0,26		

Bardzo niska separowalność ekstrakcji PCA sugerowała wyniki klasyfikacji bliskie szansie losowej, jednak wynikowa miara Accuaracy Score okazała się znacząco wyższa, o ponad 15 punktów procentowych. Wynikowe Accuracy Score dla ekstrakcji za pomocą mask okazało się bardzo dobre, jedynie kilka punktów procentowych za surowymi danymi, choć separowalność nie była wysoka. Wyniki te sugerują że nawet prosty model liniowy potrafi radzić sobie z bardzo nachodzącymi na siebie danymi.

1.8 Opis wyników klasyfikacji ekstrakcji 'mnist Extr_4' oraz metody TSNE

TRAIN	Accuracy Score	Confusion Matrix	Decision Boundary
wykrywanie krawędzi	0,9272		
TSNE	0,40		
TEST	Accuracy Score	Confusion Matrix	Decision Boundary
wykrywanie krawędzi	0,89		
TSNE	0,40		

Zgodnie z oczekiwaniami odnośnie separowalności klas dla metody uwzględniającej algorytm t-SNE, dokładność przypisanych etykiet jest bardzo dobra, biorąc pod uwagę stopień redukcji cech (z 784 na 2), bo wynosząca 0.4. W przypadku metody ekstrakcji cech poprzez wykrywanie krawędzi, dokładność przypisania etykiet jest bardzo wysoka, mimo wcześniej nisko ocenionej separowalności klas, poprzez miarę Silhouette sugerującą możliwość nakładania się na siebie klastrów.