Projekt 8
Konwolucyjne sieci neuronowe

Piotr Szczygieł, Tomasz Zawadzki

Temat projektu **2**

Wstęp **2**

Klasyfikator SVM **3**
MNIST 3
FMNIST 3

Konwolucyjne sieci neuronowe (CNN) **4**

Ocena klasyfikatorów w zależności od liczby przykładów uczących **7**
MNIST 7
FMNIST 12

Ocena klasyfikatorów w zależności od czasu uczenia **17**
MNIST 17
FMNIST 20

Klasyfikator zespołowy **23**
MNIST 23
FMNIST 24

Wnioski **25**

Temat projektu

1. Zbiory danych:
    a. MNIST obrazki cyfr 28x28, 70000 próbek, 10 klas
    b. FMNIST obrazki odzieży 28x28, 70000 próbek, 10 klas
2. Celem projektu jest porównanie klasyfikatorów: klasycznego klasyfikatora SVM trenowanego dla pełnych danych i wszystkich cechach oraz kilku sieci CNN składających się z tej samej liczby parametrów (wag) lecz różnej ilości warstw konwolucyjnych (1, 2, 5, 10). Zrobić to dla dwóch różnych ilości wag - stosunkowo małej oraz dużej.
3. Ocenić jakość klasyfikatorów w zależności od:
    a. ilości przykładów uczących, zakładając zbiór testowy złożony z reszty nieużytych do uczenia przykładów.
    b. od czasów uczenia (założyć 3 budżety czasowe T1, T2, T3, krótki, średni i długi). Można założyć że klasyfikatory wchodzące w skład zespołowego klasyfikatora uczone są równolegle.
4. Wyniki jakości klasyfikatorów oceniać na bazie krzyżowej walidacji, accuracy, loss, krzywa ROC, krzywa precision-recall, pola pod krzywymi, F1.
   Dokonać porównania wyników accuracy i loss.

Wstęp

Celem projektu jest dostarczenie odpowiedzi na następujące pytania:

- Czy konwolucyjne sieci neuronowe (CNN) są rzeczywiście lepsze od klasyfikatora SVM?
- Czy wydłużanie sieci jest rzeczywiście korzystne? Jeśli tak, to do jakiego momentu?
- Czy połączenie kilku wytrenowanych sieci ostatnią warstwą daje lepsze rezultaty niż każda ze składowych sieci osobno?

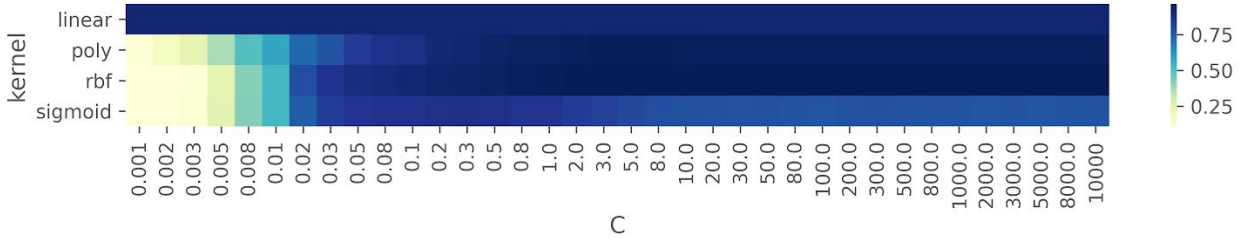
Klasyfikator SVM

W celu znalezienia najlepszego klasyfikatora SVM sprawdzono różne kombinacje hiperparametrów:

- kernel (jądro),
- stała ucząca (parametr regulujący).

Dla każdego zbioru danych (MNIST, FMNIST) dokładności klasyfikatorów zostały zwizualizowane na heatmapach. Do dalszej analizy w celu porównania klasyfikatora SVM z sieciami CNN został wybrany klasyfikator o najwyższej dokładności (accuracy) dla każdego zbioru danych.

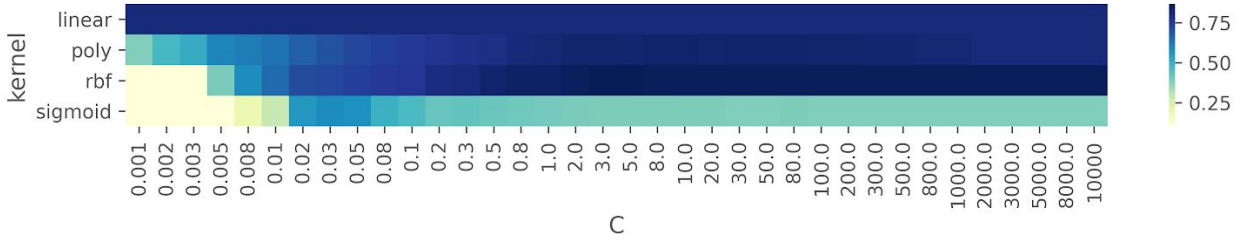
MNIST



Rys. 1. Heatmapa prezentująca dokładność klasyfikatora SVM dla zbioru MNIST dla różnych kerneli w zależności od stałej uczącej

Najlepszy klasyfikator SVM z kernelem rbf i stałą uczącą 3.0 osiąga na zbiorze testowym wysoką dokładność 96,4%. W praktyce oznacza to, że na 100 obrazków jedynie 3-4 z nich zostaną zaklasyfikowane nieprawidłowo. Jest to najprawdopodobniej spowodowane faktem, że zbiór MNIST jest relatywnie prostym zbiorem danych.

FMNIST



Rys. 2. Analogiczna heatmapa dla zbioru FMNIST

Najlepszy klasyfikator SVM z kernelem rbf i stałą uczącą 5.0 osiąga na zbiorze testowym dokładność 86,7%. Zbiór danych FMNIST jest zauważalnie trudniejszy niż MNIST, co znajduje odzwierciedlenie w niższej dokładności osiąganej przez klasyfikator SVM na tym zbiorze.

Konwolucyjne sieci neuronowe (CNN)

Sieci CNN zostały skonstruowane dla 2 różnych liczb wag (w przybliżeniu) oraz 4 różnych liczb warstw konwolucyjnych.

- warstwy konwolucyjne
    - kernel (jądro): 3x3
    - padding (wyrównanie): 0
    - funkcja aktywacji: ReLU

- warstwy poolingu
    - max pooling 2x2

- warstwa sieci neuronowej
    - 1 warstwa ukryta
        - 16, 32 albo 64 neuronów
        - funkcja aktywacji: ReLU

    - warstwa końcowa
        - 10 neuronów
        - funkcja aktywacji: softmax

Liczba warstw konwolucyjnych	Nazwa modelu	Liczba parametrów (wag)
1	CNN 1 S	~12,000
2	CNN 2 S	
5	CNN 5 S	
10	CNN 10 S	
1	CNN 1 L	~55,000
2	CNN 2 L	
5	CNN 5 L	
10	CNN 10 L	

Tabela 1. Wybrane parametry wykorzystanych konwolucyjnych sieci neuronowych

Aby zmniejszyć wpływ losowości na wyniki eksperymentu, sieci były zawsze inicjowane tym samym zestawem wag początkowych. Przykładowo sieć “CNN 1 S” trenowana na 7000 i 8000 przykładach miała te same wagi początkowe. Natomiast sieci “CNN 1 S” i “CNN 1 L” miały różne wagi początkowe z uwagi na różnicę w budowie sieci (inna liczba wag).

Sieci były trenowane przy użyciu solvera (optymalizatora) adam.

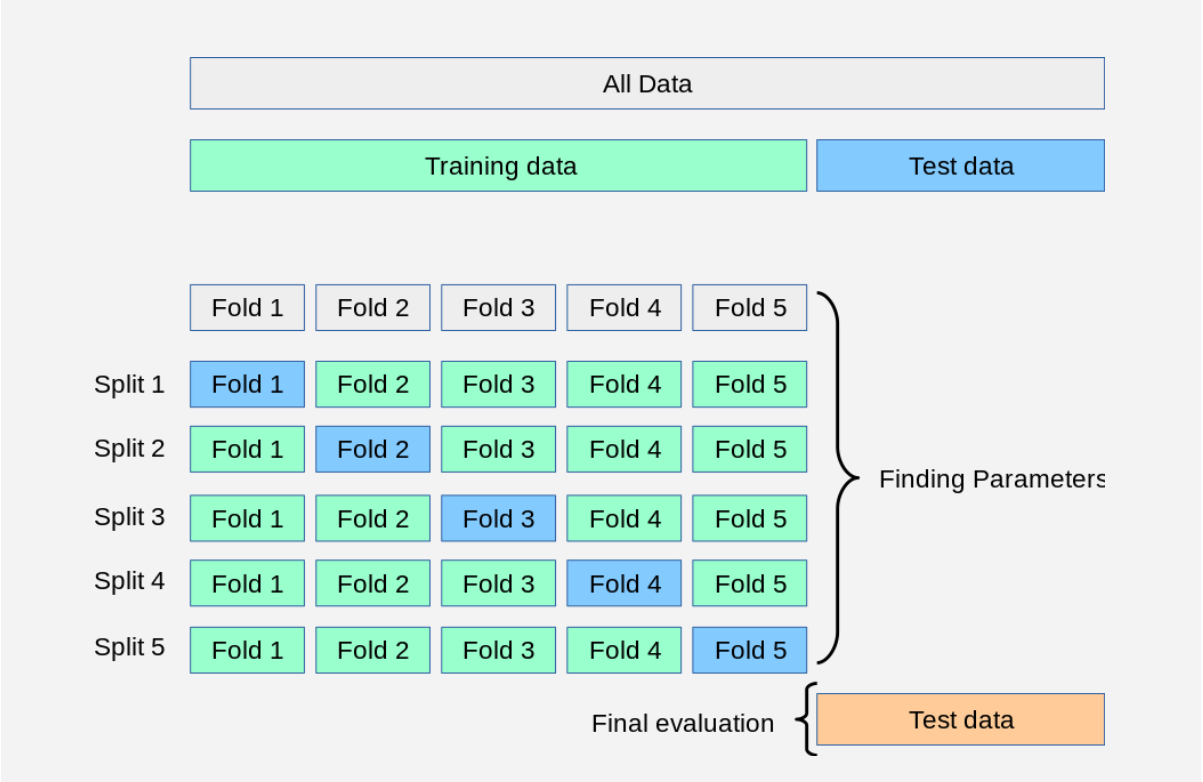
W celu zachowania zbliżonych liczności poszczególnych klas zastosowano taki resampling, który dodatkowo spełnia relację zawierania, tzn. zbiór treningowy o rozmiarze 8000 zawiera wszystkie przykłady z wykorzystanego wcześniej zbioru o rozmiarze 7000 oraz 1000 nowych sampli.

Jakości klasyfikatorów były analizowane na podstawie następujących metryk:

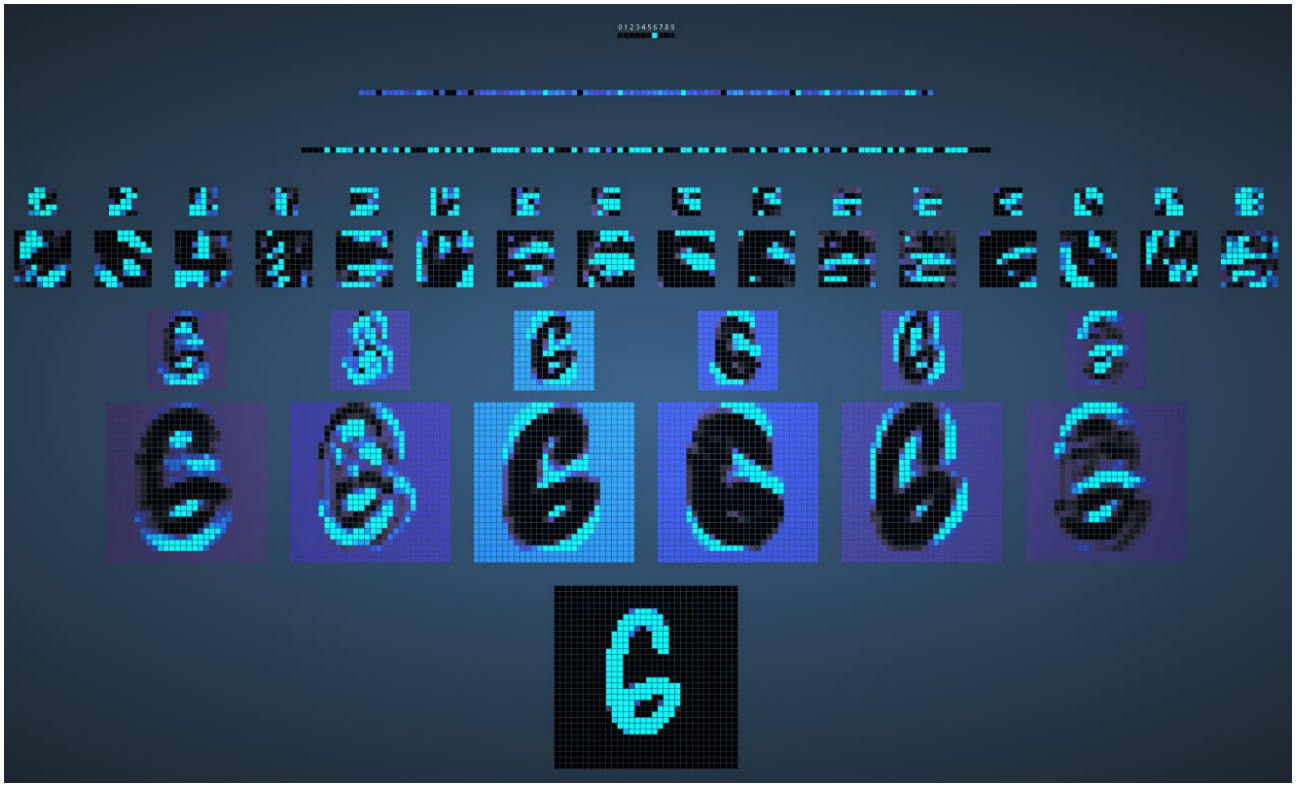
- walidacja krzyżowa (5-fold cross-validation):
    - accuracy
    - loss
    - F1 score

- walidacja zbiorem testowym:
    - accuracy (*val_accuracy*)
    - loss (*val_loss*)
    - F1 score (*val_f1*)

    - confusion matrix
    - krzywa ROC + pole pod krzywą (AUC)
    - krzywa precision-recall + pole pod krzywą (AUC)



Rys. 3. Schemat podziału zbioru danych na 5 podzbiorów do walidacji krzyżowej (https://scikit-learn.org/stable/modules/cross_validation.html)



Rys. 4. Interaktywny przykład działania konwolucyjnej sieci neuronowej uczonej zbiorem MNIST (https://www.cs.ryerson.ca/~aharley/vis/conv/flat.html)

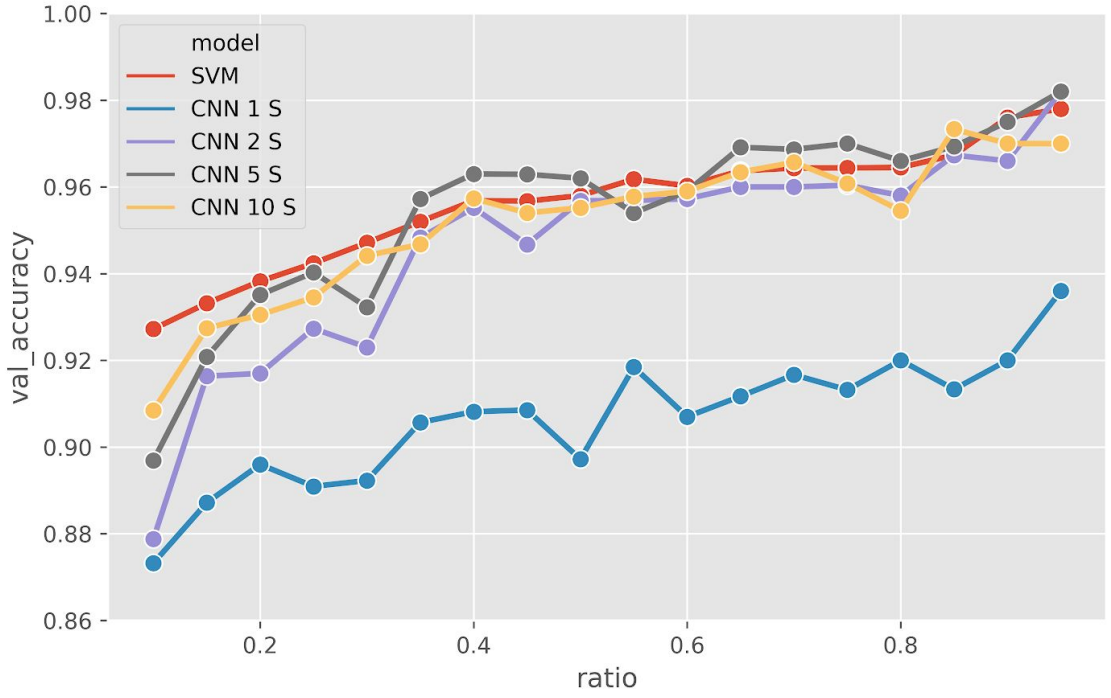
Ocena klasyfikatorów w zależności od liczby przykładów uczących

W pierwszym zadaniu badaliśmy klasyfikatory w zależności od liczby próbek treningowych i testowych, z zachowaniem ich stałej sumy. Przykładowo testowaliśmy jakość klasyfikatora dla 1000 próbek treningowych i 9000 testowych oraz dla 8000 próbek treningowych i 2000 testowych. Wszystkie sieci CNN były trenowane przez 10 epok.

Ponieważ złożoność obliczeniowa sieci neuronowych zależy bezpośrednio od liczby wag, performance sieci można porównywać tylko gdy mają one zbliżoną liczbę parametrów. Z tego powodu analiza wyników została przeprowadzona w dwóch kategoriach, osobno dla małych i dużych sieci.

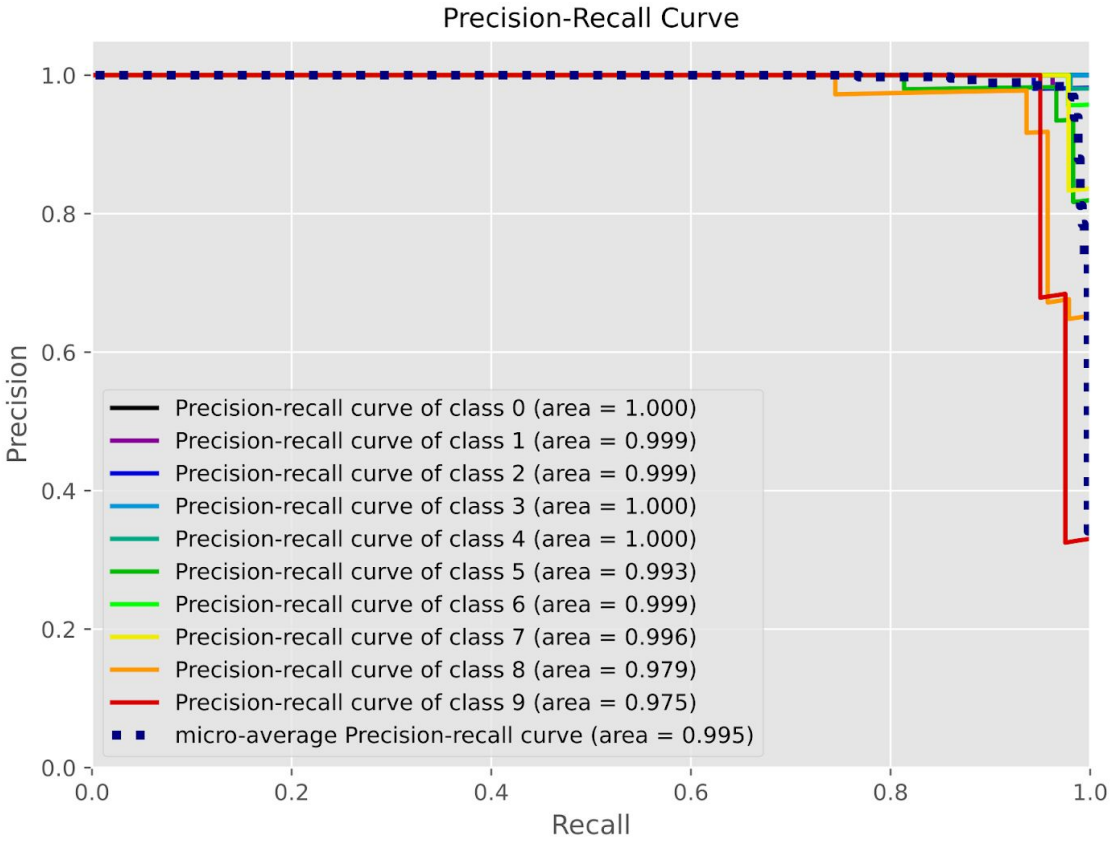
MNIST

Dla zbioru MNIST łączna liczba próbek wynosiła 10000.

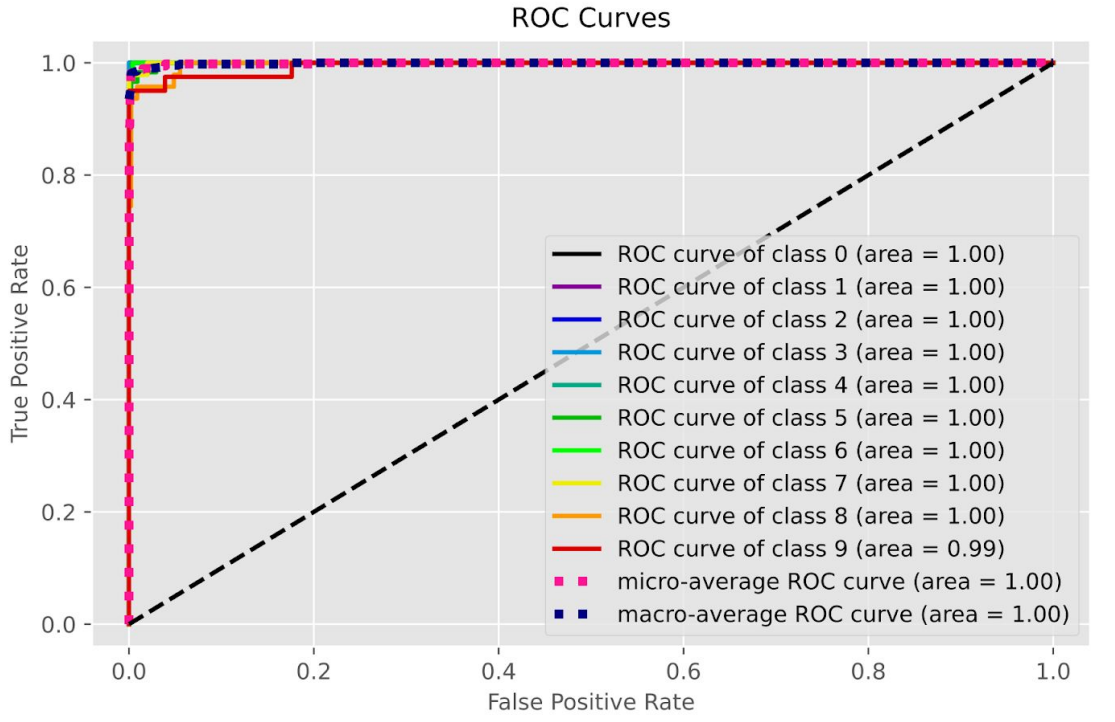


Rys. 5. Zależność dokładności konwolucyjnych sieci neuronowych z małą liczbą parametrów (S) od stosunku liczby przykładów treningowych i testowych przy zachowaniu ich stałej sumy

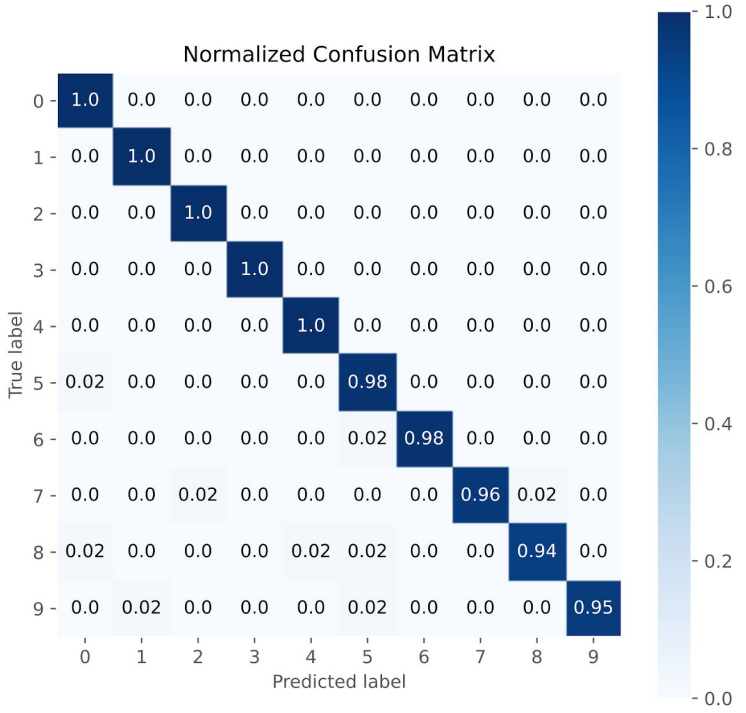
Na wykresie widać oczywistą zależność, że im więcej próbek treningowych, tym lepsza dokładność modelu. Widać też, że sieć z pojedynczą warstwą konwolucyjną osiągnęła najsłabsze wyniki, natomiast reszta sieci osiągnęła wyniki podobne do klasyfikatora SVM. Najlepsza sieć CNN 5 S osiąga nieznacznie lepsze rezultaty od SVM.



Rys. 6. Krzywa precision-recall dla najlepszej spośród wytrenowanych sieci (CNN 5 S, ratio 0.95)
Przebieg krzywych jest zbliżony do wzorcowego wykresu dla klasyfikatora idealnego.
Najmniejsze pole pod krzywą ma klasa cyfry 9.

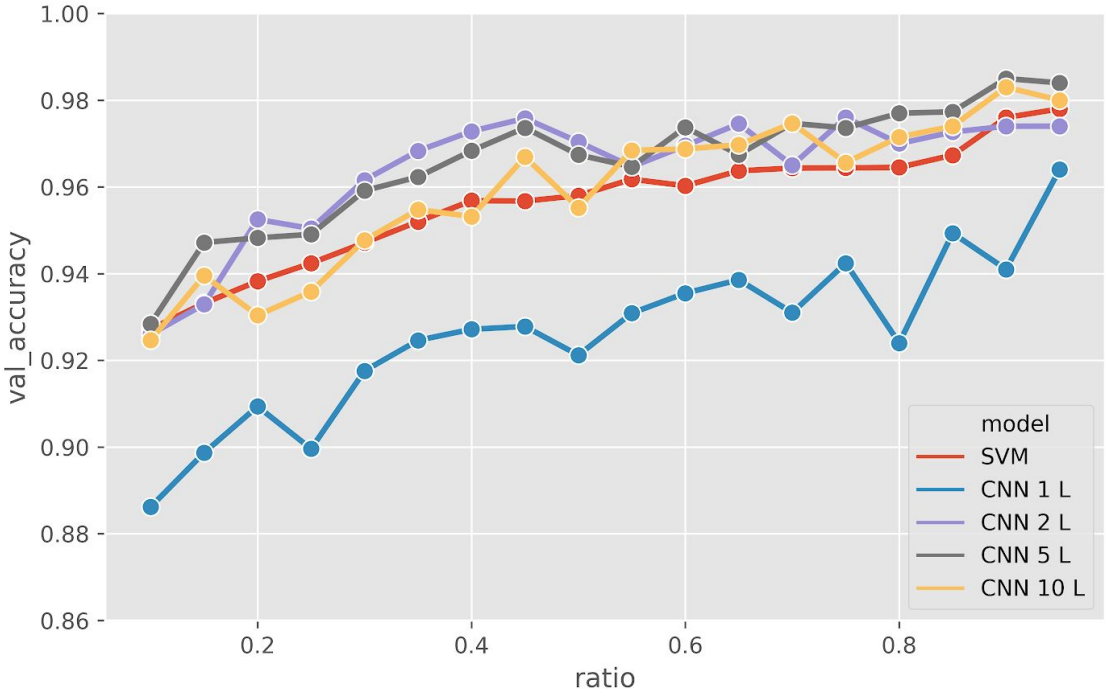


Rys. 7. Krzywa ROC dla najlepszej spośród wytrenowanych sieci (CNN 5 S, ratio 0.95)



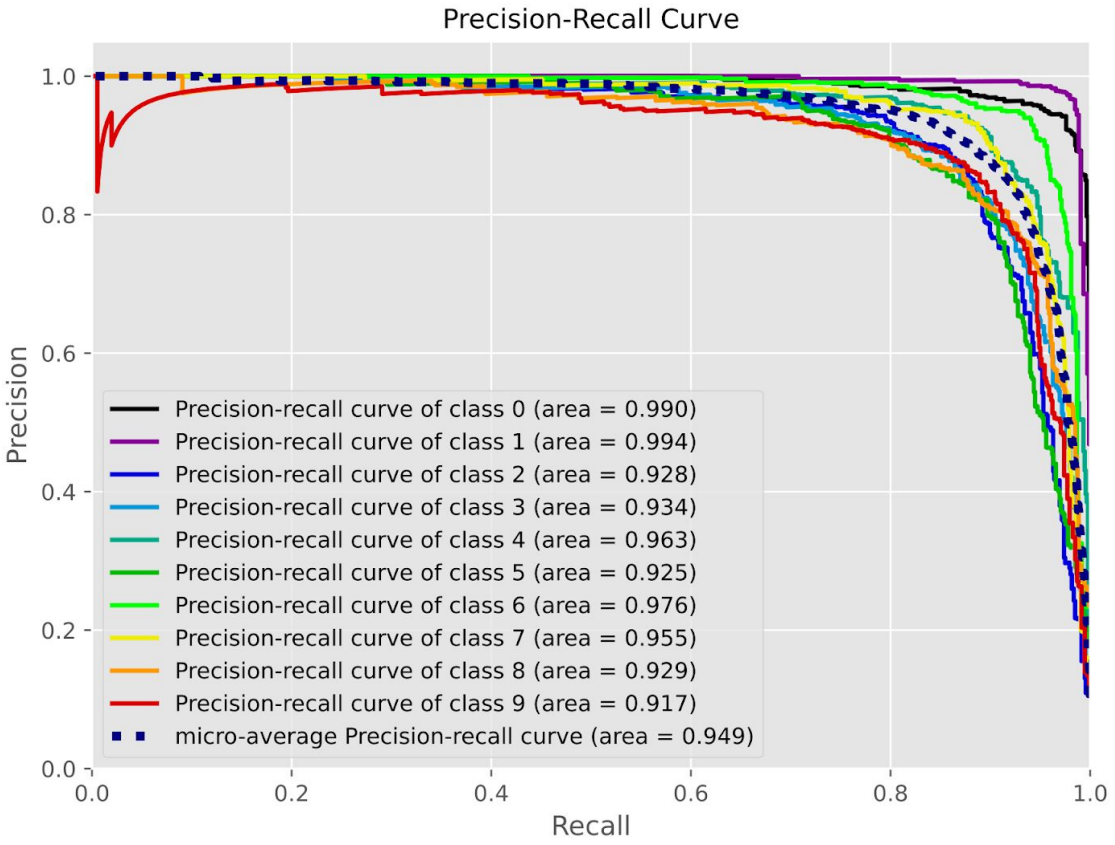
Rys. 8. Tablica pomyłek (confusion matrix) dla najlepszej spośród wytrenowanych sieci.

Wartości 1.0 na głównej przekątnej macierzy pomyłek oznaczają bezbłędne rozpoznanie tych cyfr. Sieć pomyliła się rozpoznając niektóre cyfry 8 i 9, które błędnie zaklasyfikowała jako 0, 5, 6 oraz 1, 5 odpowiednio.

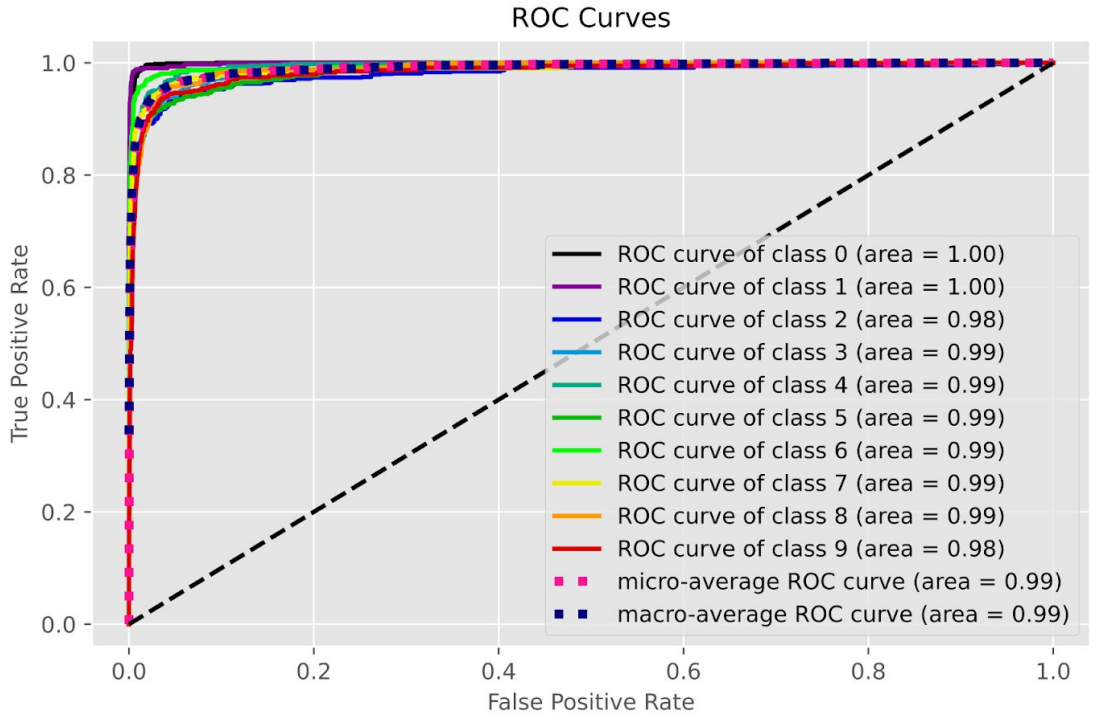


Rys. 9. Zależność dokładności konwolucyjnych sieci neuronowych z dużą liczbą parametrów (L) od stosunku liczby przykładów treningowych i testowych przy zachowaniu ich stałej sumy

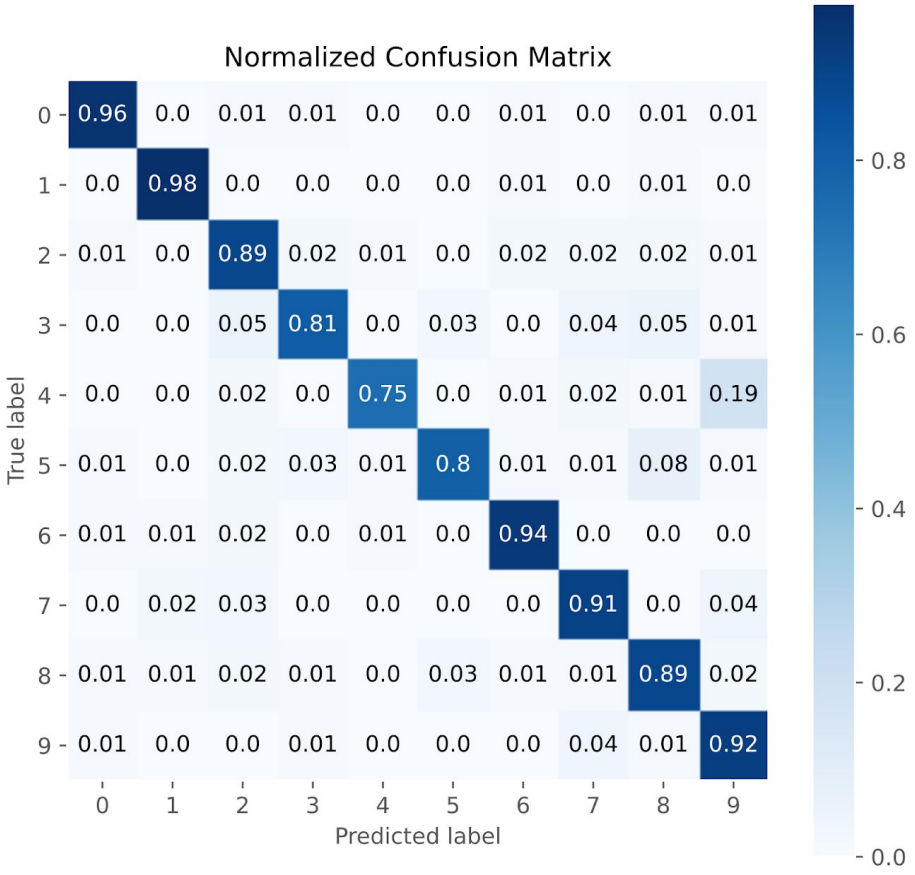
Na tym wykresie możemy zaobserwować taką samą zależność jak dla małych sieci CNN, z tą różnicą, że sieci które mają więcej niż jedną warstwę konwolucyjną osiągają lepsze wyniki niż klasyfikator SVM. Jest to spowodowane większą ilością parametrów w każdej z tych sieci.



Rys. 10. Krzywa precision-recall dla najgorszej spośród wytrenowanych sieci (CNN 1 L, ratio 0.1)
Przebieg krzywych oraz pole powierzchni pod nimi (AUC) wskazuje,
że sieć najgorzej radzi sobie z klasyfikacją cyfr 3, 5 i 9.

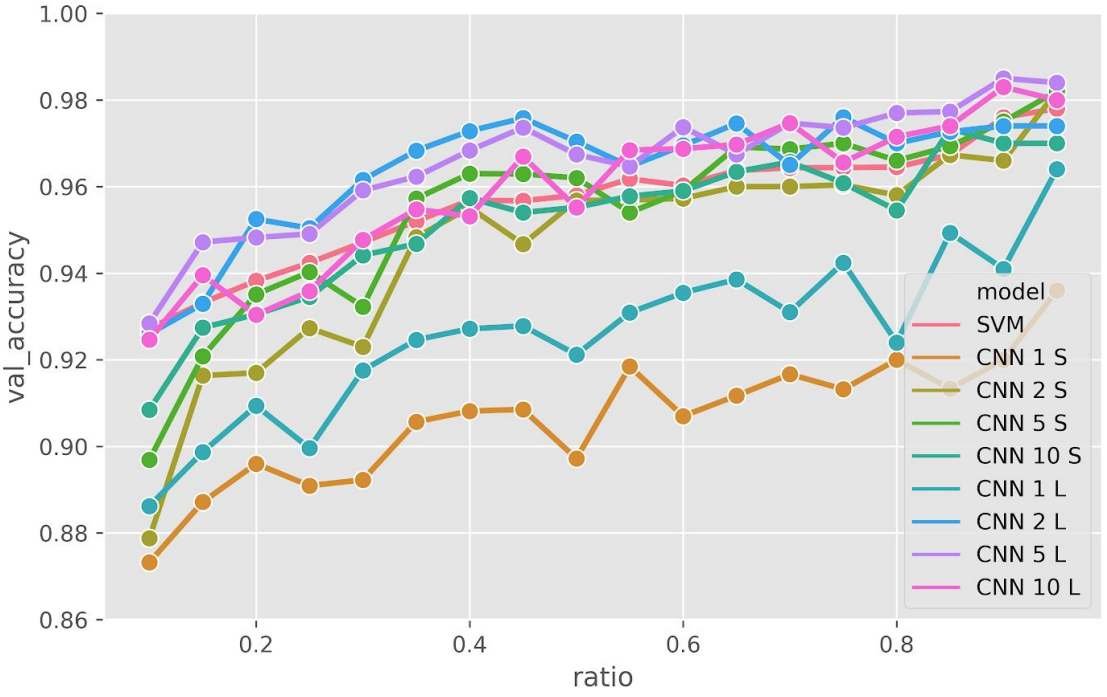


Rys. 11. Krzywa ROC dla najgorszej spośród wytrenowanych sieci (CNN 1 L, ratio 0.1)



Rys. 12. Tablica pomyłek (confusion matrix) dla najgorszej spośród wytrenowanych sieci.

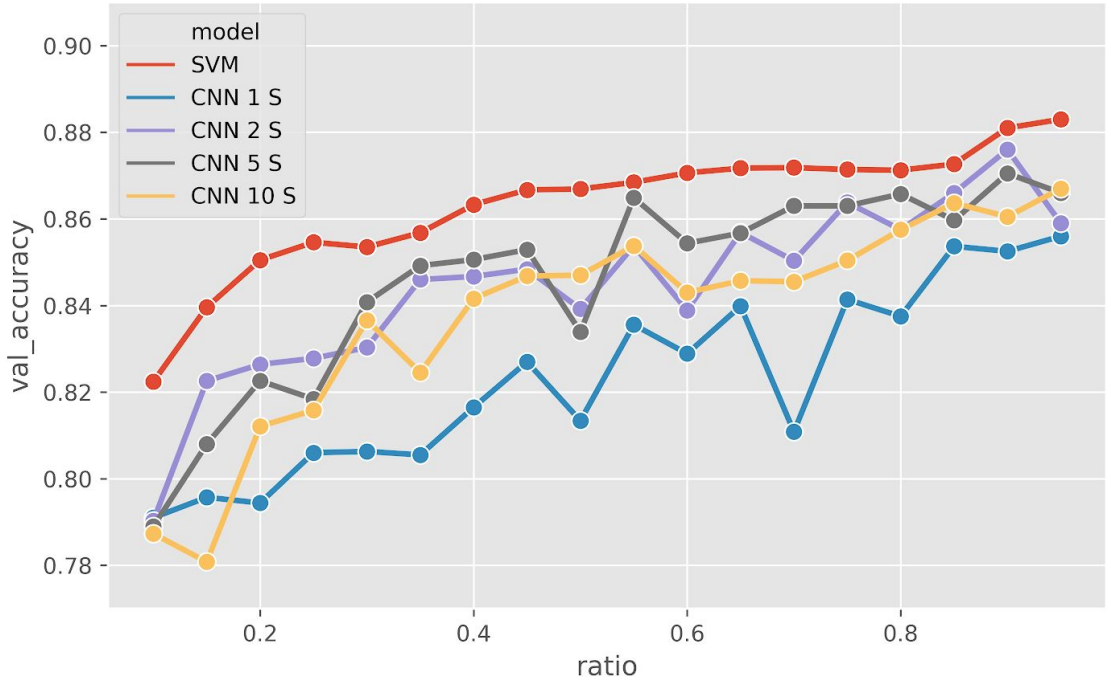
Na podstawie macierzy pomyłek można stwierdzić, że większość przykładów jest klasyfikowana poprawnie. Najczęściej popełnianym błędem jest zaklasyfikowanie cyfry 4 jako 9, a na drugim miejscu cyfry 5 jako 8.



Rys. 13. Zestawienie wyników dla klasyfikatora SVM oraz małych i dużych sieci CNN

FMNIST

Dla zbioru FMNIST łączna liczba próbek wynosiła 20000.

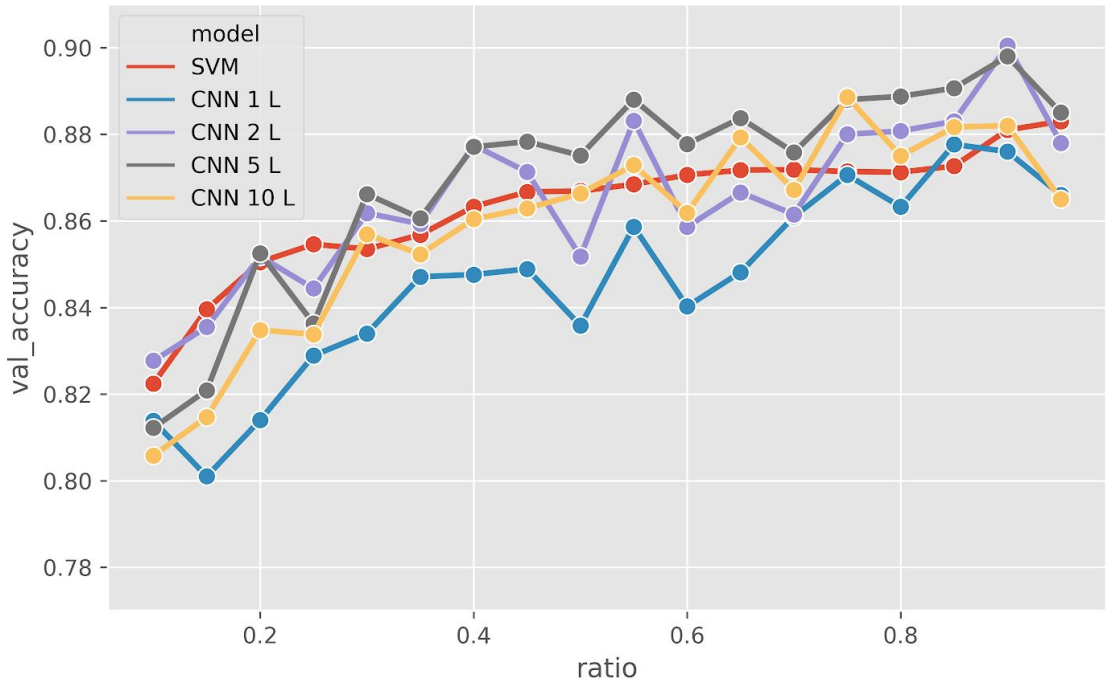


Rys. 14. Zależność dokładności konwolucyjnych sieci neuronowych z małą liczbą parametrów (S) od stosunku liczby przykładów treningowych i testowych przy zachowaniu ich stałej sumy

Analiza wykresu pozwala stwierdzić, że sieci CNN z mniejszą ilością wag uzyskały gorsze wyniki od klasyfikatora SVM. Jest to spowodowane faktem, że sieci posiadały jedynie około 12000 parametrów trenowalnych, co okazało się zbyt małą ilością na zbiór danych FMNIST.

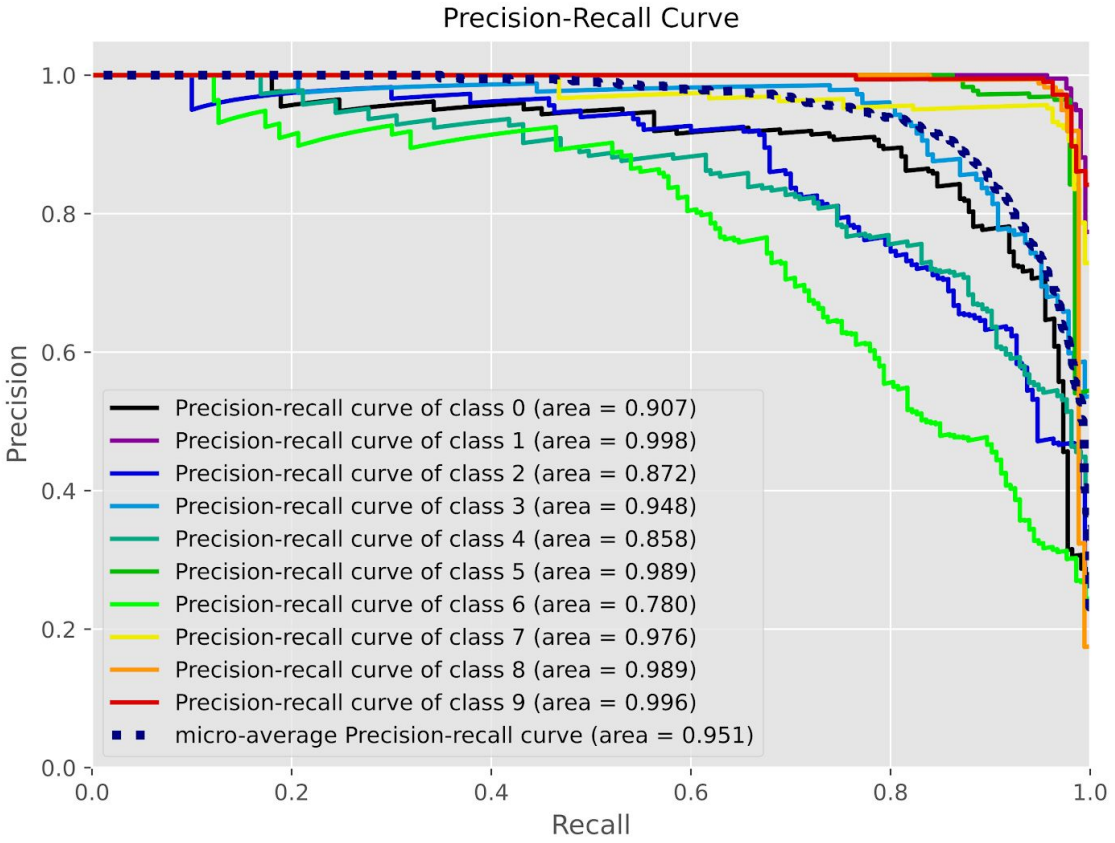
Różnica w osiąganych dokładnościach wynika również z faktu, że wytrenowanie klasyfikatora SVM trwało kilkadziesiąt razy dłużej niż wytrenowanie małej sieci CNN.

Przy zwiększaniu zbioru treningowego klasyfikator SVM zachowuje się bardziej stabilnie, podczas gdy sieci CNN osiągają bardziej zróżnicowane rezultaty.

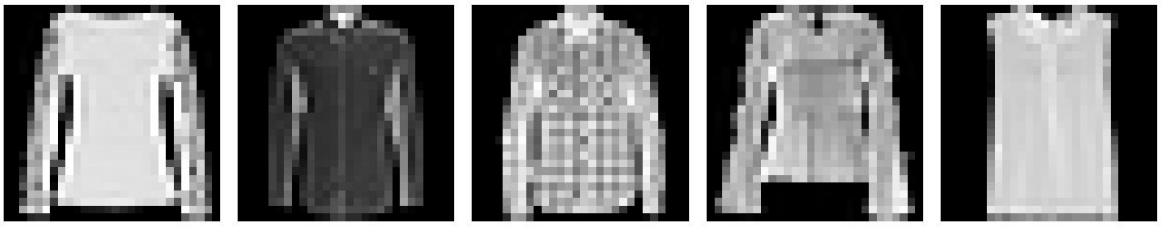


Rys. 15. Zależność dokładności konwolucyjnych sieci neuronowych z dużą liczbą parametrów (L) od stosunku liczby przykładów treningowych i testowych przy zachowaniu ich stałej sumy

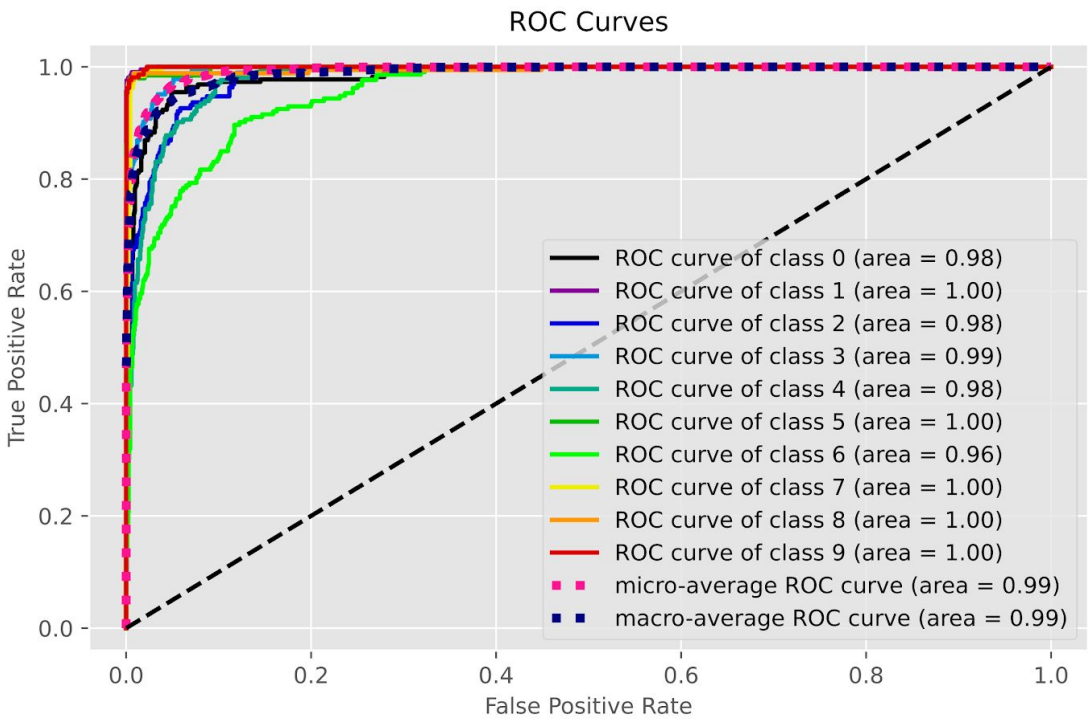
Widzimy, że jeśli zwiększymy liczbę parametrów sieci CNN do ponad 50000, potrafi ona osiągnąć lepsze wyniki niż klasyfikator SVM, przy dużo krótszym czasie trenowania. Ponownie najlepsze wyniki osiągane są dla sieci posiadających 2 oraz 5 warstw konwolucyjnych, a najgorsze dla pojedynczej warstwy konwolucyjnej.



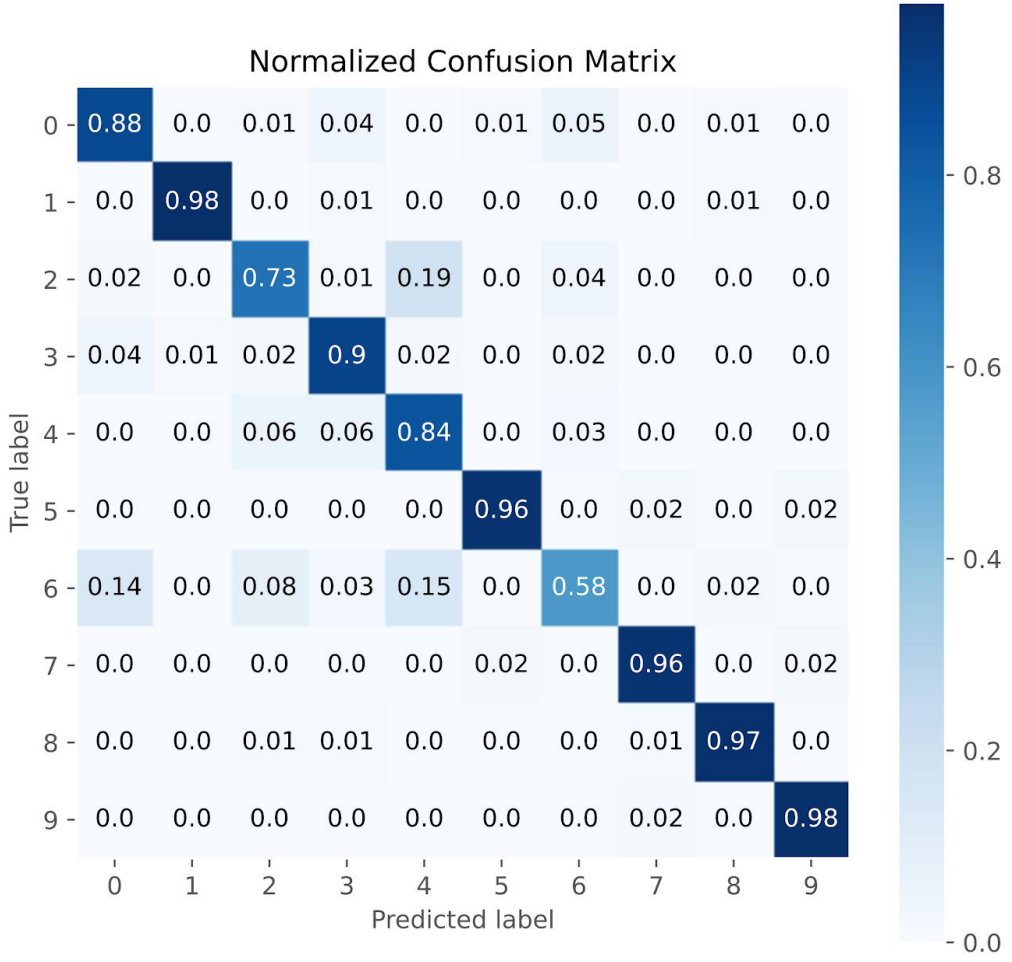
Rys. 16. Krzywa precision-recall dla najlepszej spośród wytrenowanych sieci o mniejszej liczbie parametrów (CNN 2 S, ratio 0.9). Przebieg krzywych oraz pole powierzchni pod nimi (AUC) wskazuje, że sieć miała największe problemy z klasyfikacją klasy 6 (koszule)



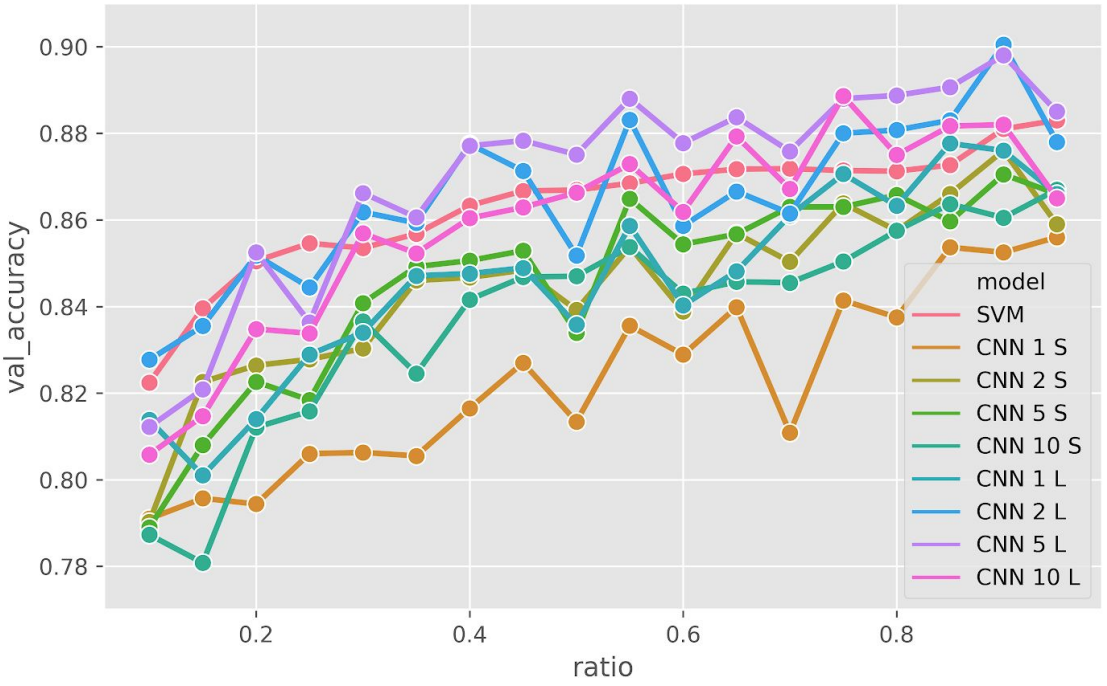
Rys. 17. Przykładowe obrazki z klasy 6 (koszule), z którymi największe problemy miały sieci CNN



Rys. 18. Krzywa ROC dla najlepszej spośród wytrenowanych sieci o mniejszej liczbie parametrów (CNN 2 S, ratio 0.9)



Rys. 19. Tablica pomyłek (confusion matrix) dla najlepszej spośród wytrenowanych sieci o mniejszej liczbie parametrów (CNN 2 S, ratio 0.9).



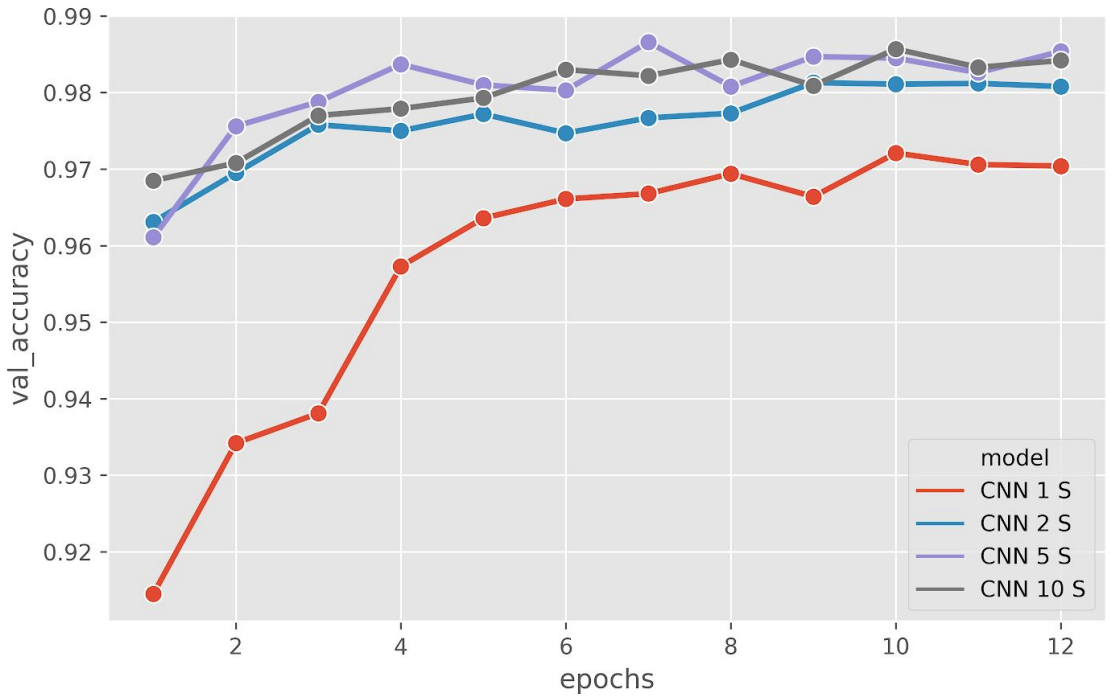
Rys. 20. Zestawienie wyników dla małych i dużych sieci CNN

Ocena klasyfikatorów w zależności od czasu uczenia

W drugim zadaniu analizowaliśmy dokładność różnych konwolucyjnych sieci neuronowych w zależności od długości czasu uczenia. Zrealizowaliśmy to trenując sieci dla różnej ilości epok z zakresu od 1 do 12. Wszystkie sieci trenowaliśmy na większej ilości próbek niż w poprzednim zadaniu. W tym celu użyliśmy zbioru 48000 próbek testowych oraz 10000 treningowych.

Na wykresach pominęliśmy klasyfikator SVM, ponieważ czas uczenia tego klasyfikatora zależy od liczby przykładów treningowych (nie jest możliwe jego trenowanie w zależności od ilości epok).

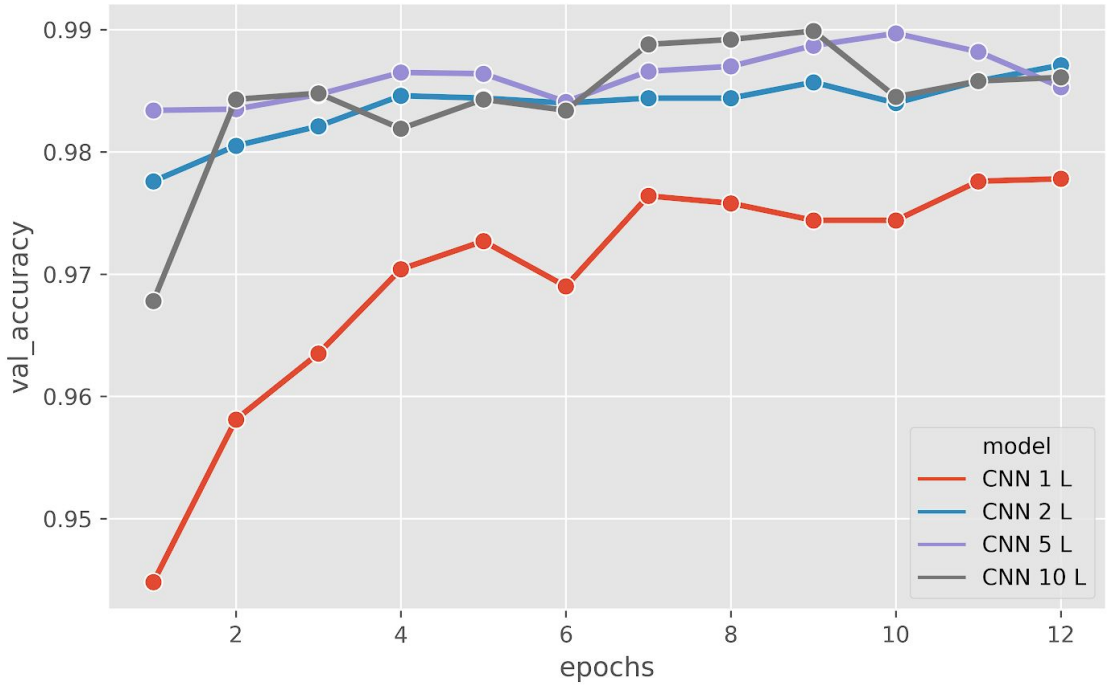
MNIST



Rys. 21. Zależność dokładności konwolucyjnych sieci neuronowych z małą liczbą parametrów (S) od liczby epok (MNIST)

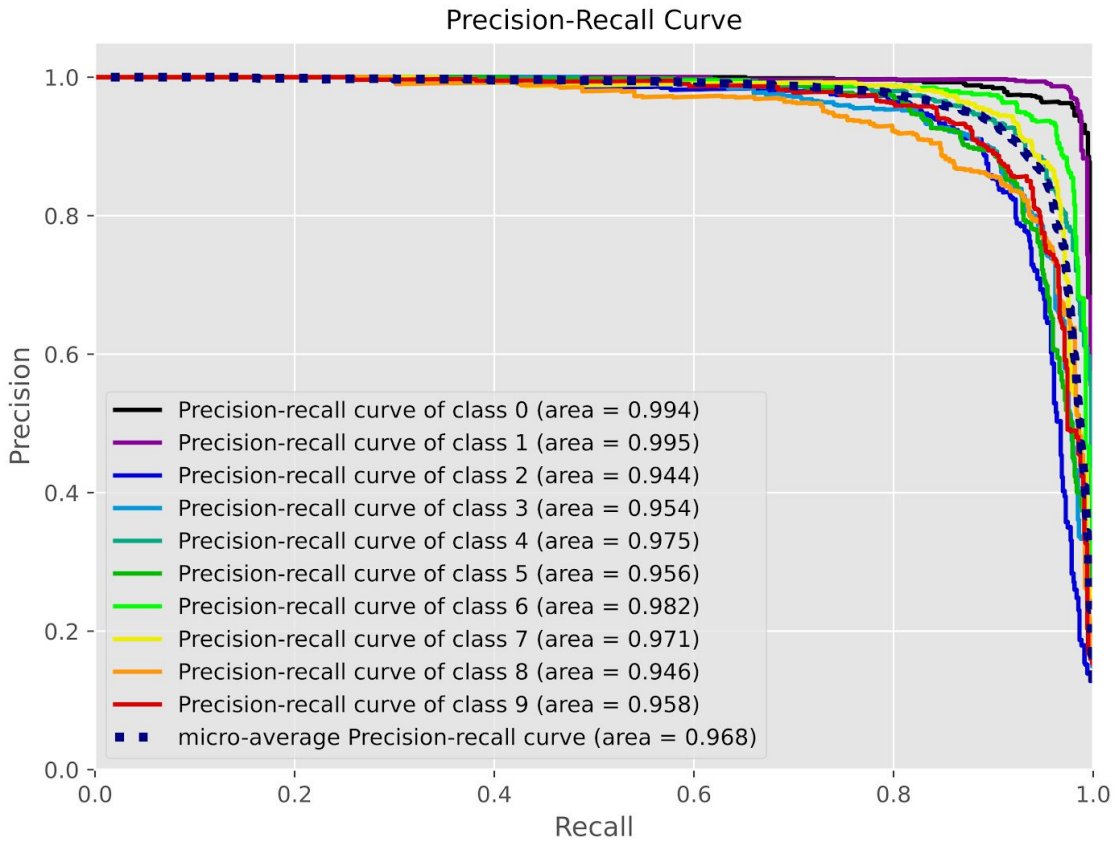
Analiza wykresu pozwala stwierdzić, że wraz ze wzrostem ilości epok, na których trenowana jest sieć, ich dokładność wzrasta. Warto jednak zaznaczyć, że po kilku epokach krzywe zaczynają się spłaszczać, a po 8 epokach dokładność sieci nie ulega już zauważalnej poprawie.

W tym wypadku najlepszymi sieciami okazały się sieci z 5 oraz 10 warstwami konwolucyjnymi. Widać również, że sieć z pojedynczą warstwą konwolucyjną osiągnęła zdecydowanie gorsze rezultaty od pozostałych.

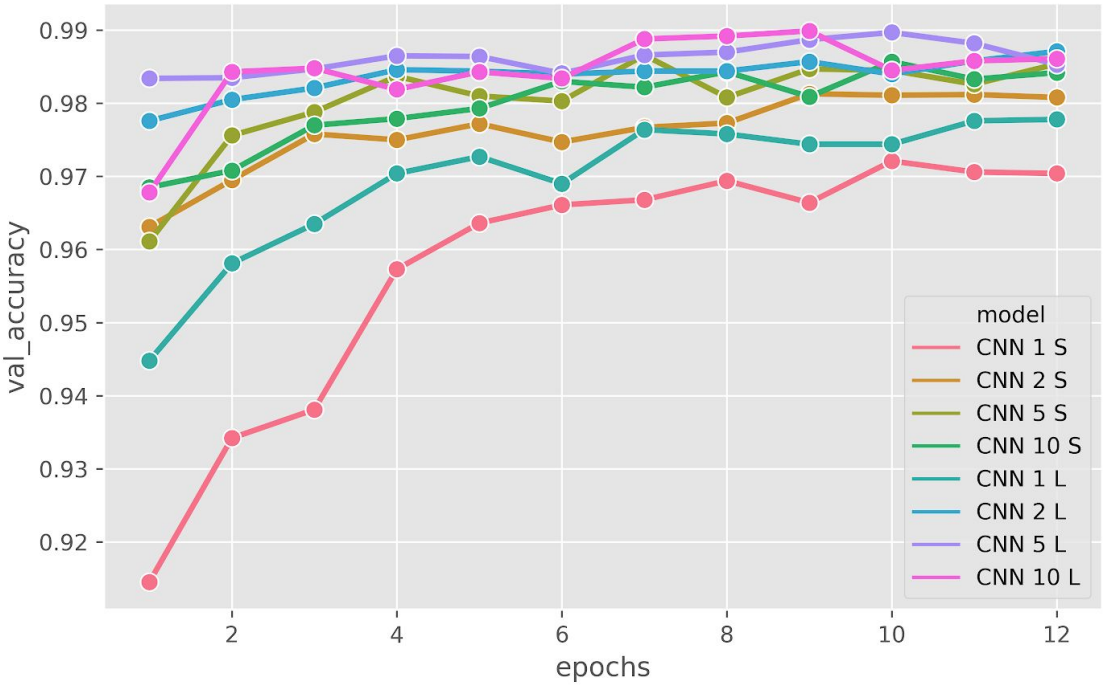


Rys. 22. Zależność dokładności konwolucyjnych sieci neuronowych z dużą liczbą parametrów (L) od liczby epok (MNIST)

Na wykresie, dla sieci z większą ilością parametrów można zaobserwować prawie to samo co na poprzednim wykresie. Sugeruje to, że dla zbioru danych MNIST, sieci CNN o mniejszej liczbie parametrów są wystarczające do osiągnięcia zadowalającej dokładności.

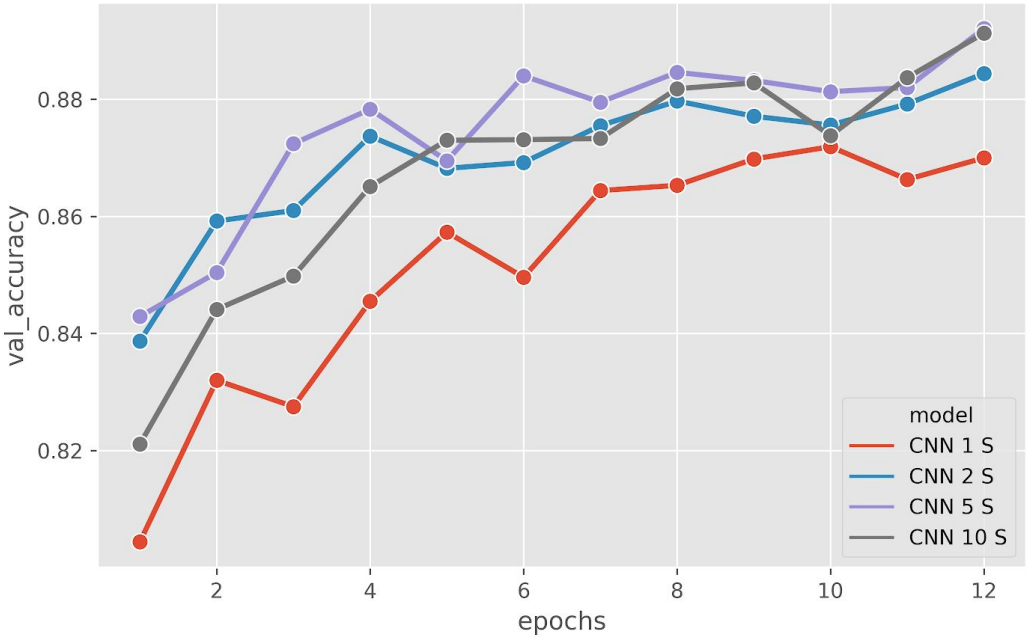


Rys. 23. Krzywa precision-recall dla najsłabszej spośród sieci o mniejszej liczbie parametrów.



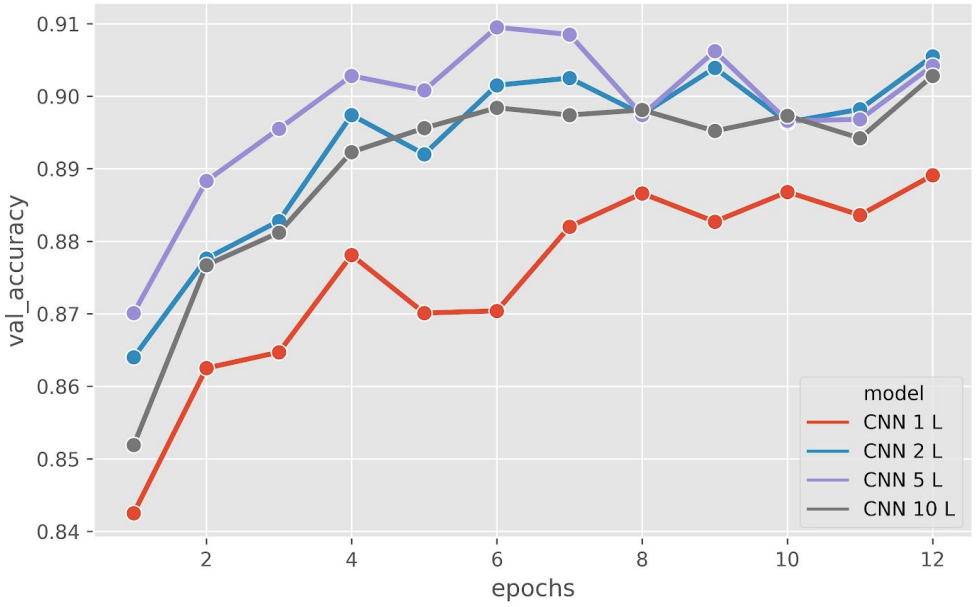
Rys. 24. Zestawienie wyników z dwóch poprzednich wykresów

FMNIST



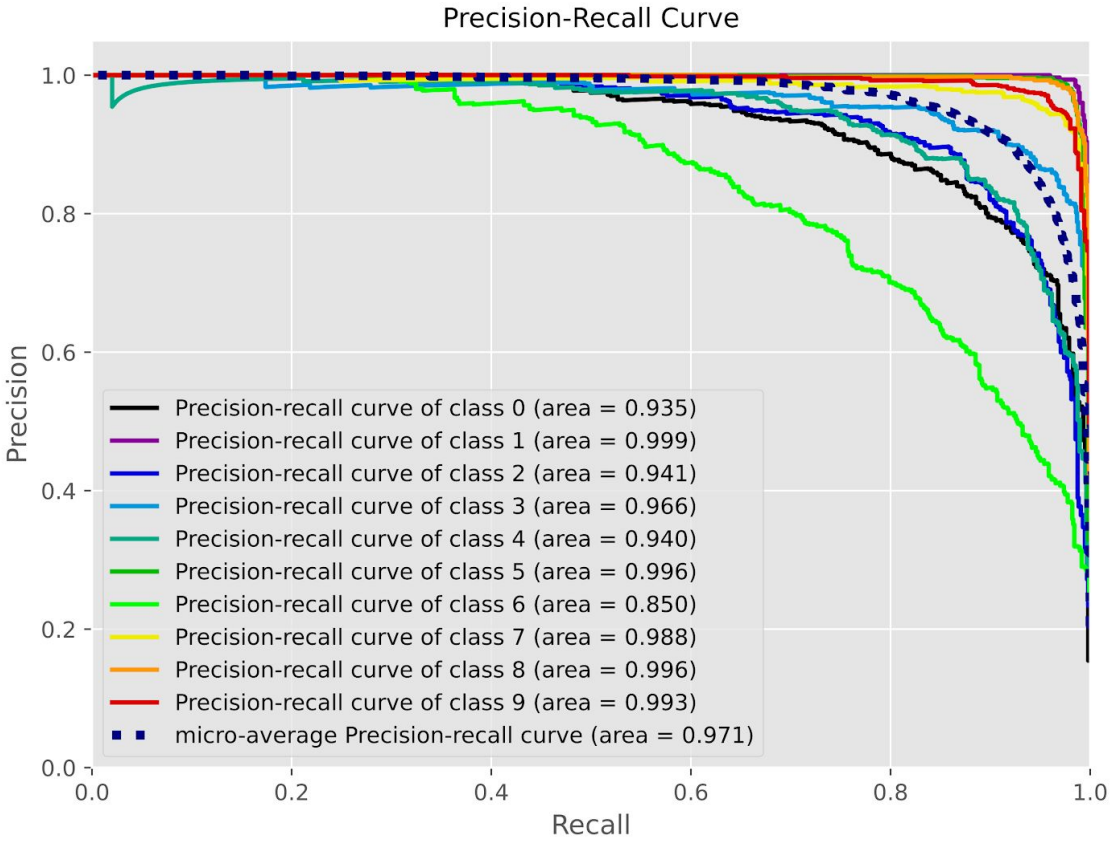
Rys. 25. Zależność dokładności konwolucyjnych sieci neuronowych z małą liczbą parametrów (S) od liczby epok (FMNIST)

Dla trudniejszego zbioru danych jakim jest FMNIST, widzimy, że ilość epok ma dużo większy wpływ na dokładność klasyfikatorów. Znowu widzimy podobne zależności jak w poprzednich zadaniach – sieć z pojedynczą warstwą konwolucyjną miała zauważalnie niższą dokładność od pozostałych sieci z większa ilością warstw konwolucyjnych.

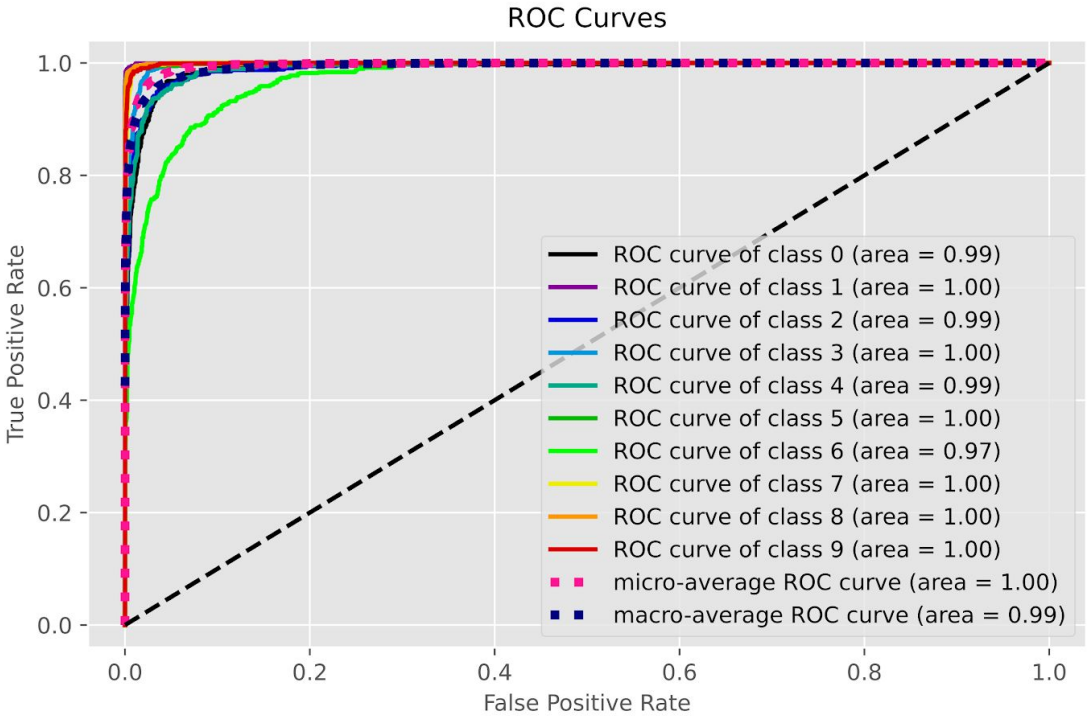


Rys. 26. Zależność dokładności konwolucyjnych sieci neuronowych z dużą liczbą parametrów (L) od liczby epok (FMNIST)

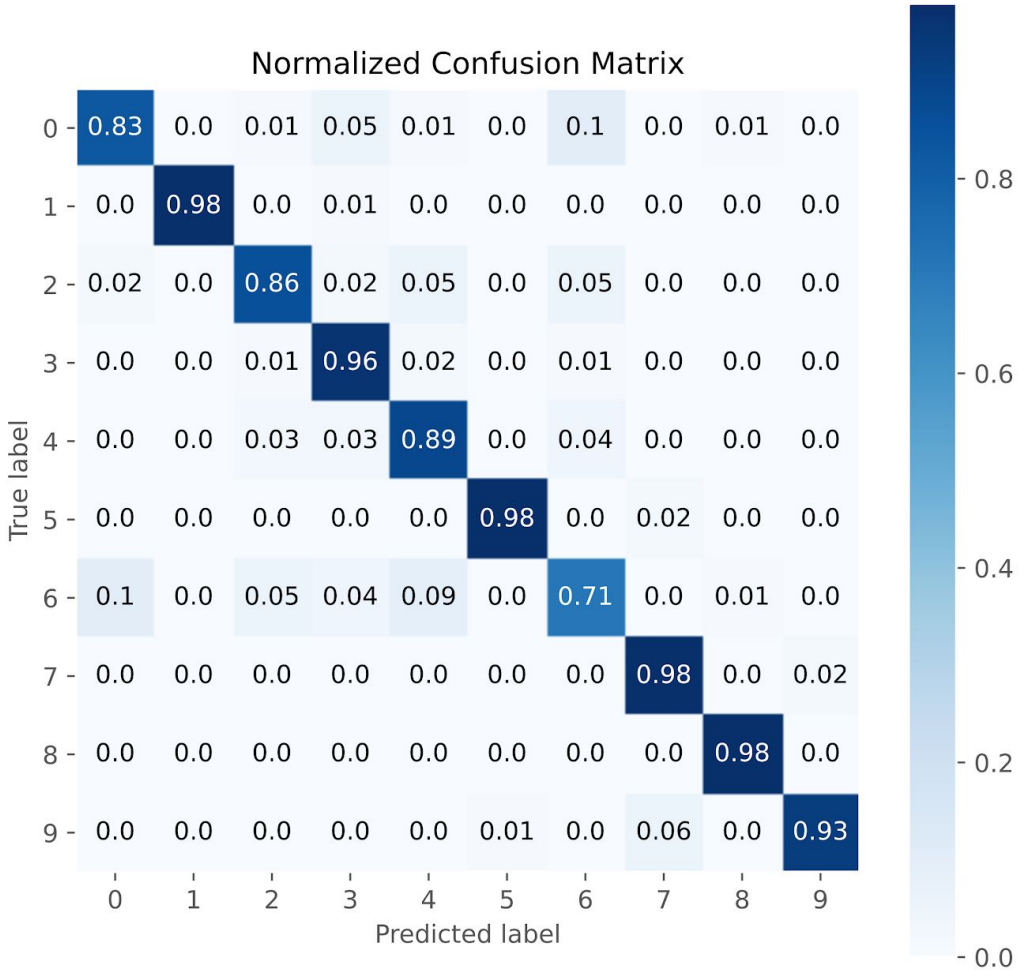
Na wykresie widzimy, że wykres jest bardzo zbliżony do tego z poprzedniego zadania, z tą różnicą, że każda z sieci ma większą dokładność.



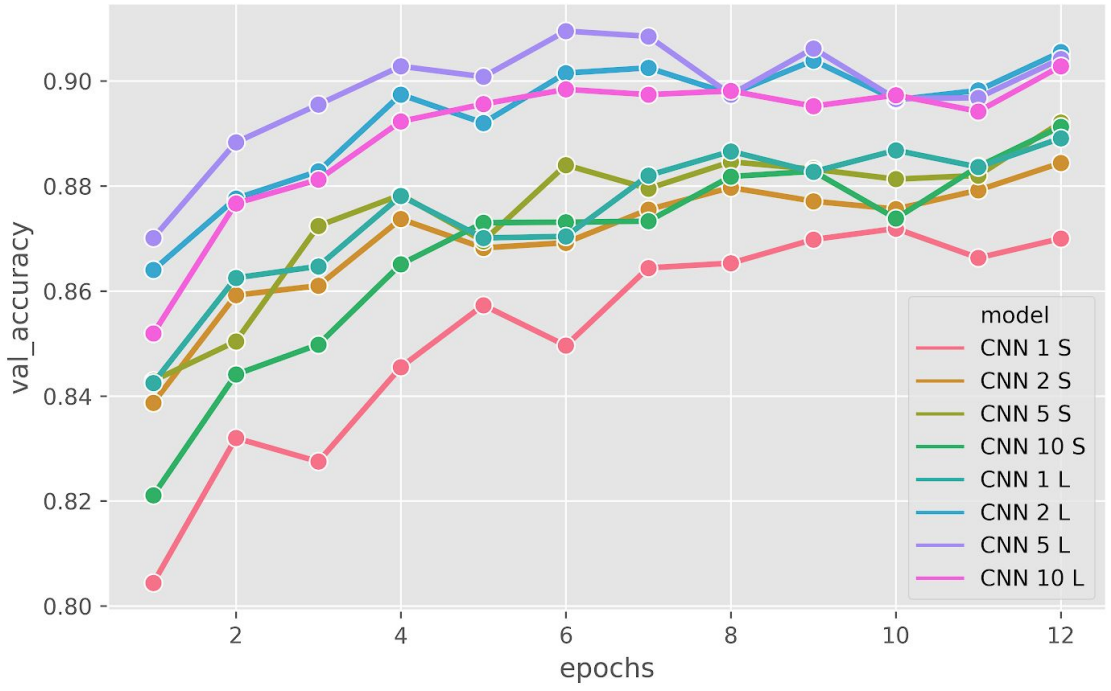
Rys. 27. Krzywa precision-recall dla najlepszej spośród wytrenowanych sieci (CNN 5 L, 6 epok). Sieć ponownie miała problem z klasą 6, czyli koszulami



Rys. 28. Krzywa ROC dla najlepszej spośród wytrenowanych sieci (CNN 5 L, 6 epok)



Rys. 29. Tablica pomyłek (confusion matrix) dla najlepszej spośród wytrenowanych sieci. Można zauważyć, że klasa 6 jest często mylona z klasą 1 (spodnie) oraz 4 (płaszcze)



Rys. 30. Zestawienie wyników z dwóch poprzednich wykresów

Klasyfikator zespołowy

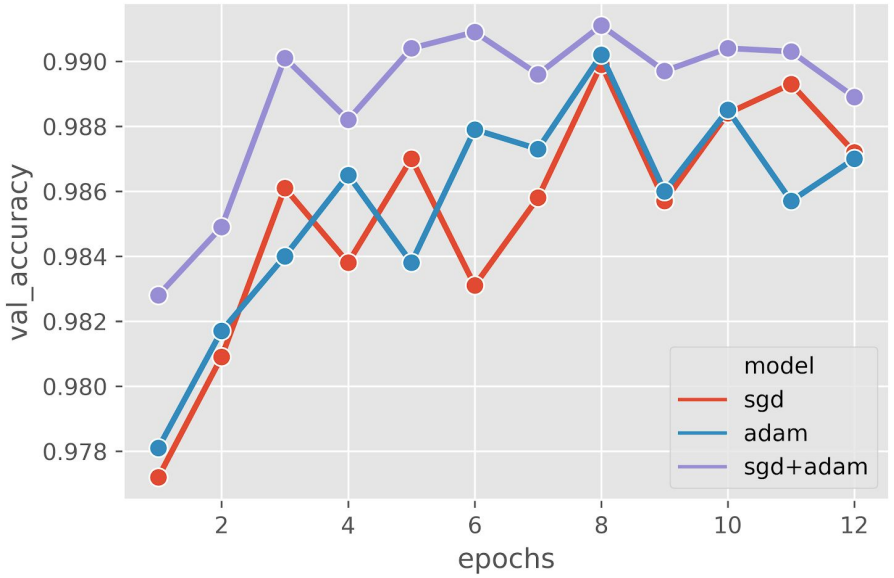
Dotychczas najlepsze rezultaty uzyskiwała sieć “CNN 2 L” składająca się z dwóch warstw konwolucyjnych i posiadająca około 55,000 parametrów trenowalnych (wag), dlatego to ona zostanie wykorzystana do zbudowania klasyfikatora zespołowego.

Klasyfikator zespołowy został skonstruowany z dwóch sieci o identycznej budowie, ale różniących się wagami z powodu wykorzystania dwóch różnych optymalizatorów. Sieci składowe zostały wcześniej wytrenowane przy użyciu optymalizatorów: SGD(learning_rate=0.01, momentum=0.9, nesterov=True) oraz adam.

Wytrenowane sieci składowe zostały pozbawione własnych warstw końcowych, a następnie zostały połączone ostatnimi warstwami MLP (zawierającymi po 64 neuronów każda) do jednej warstwy końcowej (wspólny softmax). Nowy model wymaga wytrenowania jedynie nowej ostatniej warstwy, która posiada jedynie 1290 parametrów.

Zakładamy, że pojedyncze sieci mogą być trenowane równolegle, natomiast trenowanie ostatniej warstwy sieci zespołowej trwa krótko.

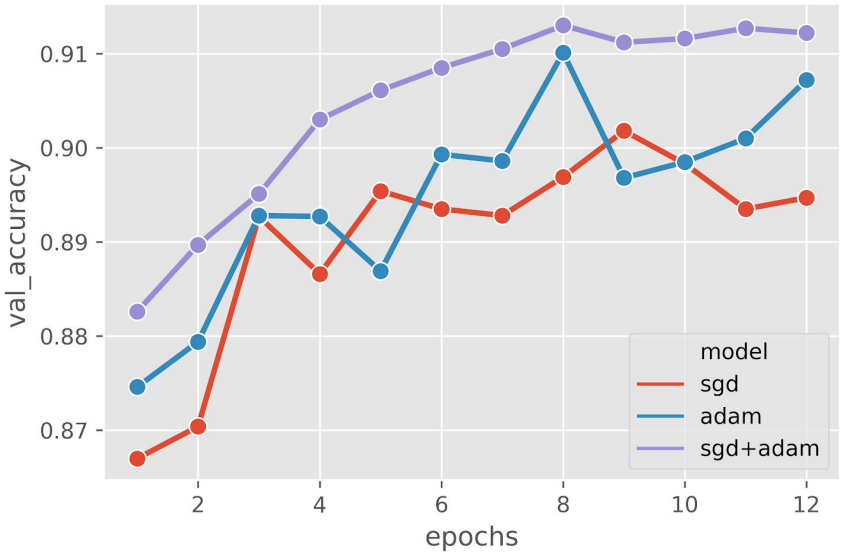
MNIST



Rys. 31. Zależność dokładności od liczby epok sieci CNN z optymalizatorami sgd i adam oraz sieci powstałej z połączenia ostatnią warstwą dwóch wytrenowanych sieci (MNIST)

Można zaobserwować, że klasyfikator zespołowy osiągnął lepsze wyniki niż pojedyncze sieci CNN. Dokładności na poziomie powyżej 99% dla zbioru MNIST nie udało się uzyskać dla żadnej innej pojedynczej sieci CNN.

FMNIST



Rys. 32. Zależność dokładności od liczby epok sieci CNN z optymalizatorami sgd i adam oraz sieci powstałej z połączenia ostatnią warstwą dwóch wytrenowanych sieci (FMNIST)

Na powyższym wykresie możemy zaobserwować podobną zależność jak w przypadku zbioru MNIST. Klasyfikator zespołowy osiąga lepsze rezultaty niż pojedyncze klasyfikatory CNN.

Wnioski

- Na podstawie zgromadzonych wyników można stwierdzić, że konwolucyjne sieci neuronowe (CNN) rzeczywiście odnoszą lepszę rezultaty niż klasyczny klasyfikator SVM, o ile zostaną odpowiednio zaprojektowane (liczba warstw konwolucyjnych, liczba wag), a także odpowiednio wytrenowane (liczba próbek treningowych, ilość epok).
- Dokładność zarówno klasyfikatora SVM jak i sieci CNN rośnie wraz ze zwiększaniem stosunku próbek treningowych do próbek testowych.
- Dla małej ilości próbek klasyfikator SVM osiągał lepsze wyniki niż sieci CNN.
- Dla większej liczby przykładów treningowych sieci CNN uzyskiwały podobne lub nieznacznie lepsze rezultaty niż klasyfikator SVM.
- Konwolucyjne sieci neuronowe (CNN) wymagają o wiele mniej czasu do wytrenowania niż osiągający podobne lub nawet nieznacznie gorsze rezultaty klasyfikator SVM.
- Dokładność sieci CNN wzrasta w zależności od liczby epok, ale po pewnym czasie zaczyna oscylować wokół pewnej wartości. Wynika to z faktu, że na zbiorze treningowym sieci osiągają dokładność blisko 100% (overfitting).
- Krzywe ROC oraz PR pozwalają zaobserwować, które klasy sprawiają sieci największe problemy. Tablica pomyłek pozwala dowiedzieć się, z jakimi klasami mylone są klasy z najmniejszą dokładnością.
- Zwiększanie liczby warstw konwolucyjnych prowadzi do poprawienia dokładności sieci. Dzieję się to jednak do pewnego momentu i podobnie jak dla ilości epok, należy znaleźć optymalną liczbę warstw dla każdego zbioru danych.
- Dla zbiorów danych użytych w projekcie najlepsze rezultaty osiągnęły sieci posiadające 2 i 5 warstw konwolucyjnych.
- Połączenie kilku sieci CNN w jedną może spowodować osiągnięcie wyższej dokładności niż którakolwiek z pojedynczych sieci byłaby w stanie osiągnąć.
- Sieci CNN wymagają większego wkładu czasu w trakcie projektowania, ale wynagradzają to lepszą dokładnością i zauważalnie krótszym czasem uczenia.