Podstawy Sieci Neuronowych

Projekt 2

Sieci konwolucyjne

Politechnika Wrocławska

Informatyczne Systemy Automatyki
Politechnika Wrocławska
Polska

Prowadzący kursu mgr inż. Michał Zmonarski

272545 Adrian Goral
272568 Mateusz Zubrzycki
272592 Paulina Szulc

Spis treści

1 Cel projektu

W projekcie należało opracować model zdolny do stwierdzenia nowotworu piersi na podstawie zdjęć oraz dostarczonych danych. Następnie porównać wyniki modeli hybrydowych z **Early-Stopping** i z **pełnymi epokami**.

2 Dane

2.1 Charakterystyka danych

Zbiorem danych wykorzystanym w projekcie jest **CBIS-DDSM (Curated Breast Imaging Subset of DDSM)**. Jest to zbiór obrazów mammograficznych wykorzystywany w analizie i wykrywaniu zmian patologicznych w piersiach. Dane zawierają obrazy w formacie **JPEG** oraz powiązane metadane opisujące cechy diagnostyczne oraz patologię zmian.

- **Rodzaj danych:** Obrazy mammograficzne wraz z informacjami diagnostycznymi.
- **Rozmiar zbioru:**
  - Tysiące obrazów mammograficznych o różnych rozdzielczościach.
  - Dodatkowe cechy diagnostyczne w formacie CSV.

2.2 Struktura danych

Dane składają się z kilku plików CSV oraz folderu z obrazami:

Pliki CSV:

- **calc_case_description_train_set.csv** i **calc_case_description_test_set.csv**: Zawierają opisy zmian związanych z mikrozwapnieniami (calcification).
- **mass_case_description_train_set.csv** i **mass_case_description_test_set.csv**: Zawierają opisy zmian masowych (mass).
- **dicom_info.csv**: Zawiera informacje techniczne na temat obrazów w formacie DICOM.
- **meta.csv**: Informacje dodatkowe o obrazach, takie jak liczba dostępnych klatek czy widoki obrazów.

Cechy diagnostyczne (w plikach CSV):

- **Numer identyfikacyjny pacjenta** (patient_id).
- **Gęstość piersi** (breast density): Liczbowa wartość odzwierciedlająca budowę piersi.
- **Typy zmian**:
  - mass shape: Kształt mas (np. IRREGULAR, LOBULATED).
  - mass margins: Granice mas (np. SPICULATED, CIRCUMSCRIBED).
  - calc type: Typ zwapnień (np. PUNCTATE, AMORPHOUS).
  - calc distribution: Rozkład zwapnień (np. CLUSTERED, SEGMENTAL).

- **Patologia zmian** (pathology): Informacja o charakterze zmiany (łagodna lub złośliwa).
- **Ścieżka do obrazu** (image file path): Lokalizacja obrazu mammograficznego w folderze.

Obrazy mammograficzne:

- **Format**: JPEG.
- **Rozdzielczość**: Różna, maksymalnie 4081x6496 pikseli.
- **Widoki**:
  - CC (Cranio-Caudal): Widok od góry.
  - MLO (Mediolateral Oblique): Widok ukośny.

2.3 Przygotowanie danych

Przetwarzanie obrazów:

- Wszystkie obrazy zostały skalowane do wspólnego rozmiaru **512x512 pikseli**.
- Brakujące obrazy (2864) zostały zastąpione w całości czarnymi zdjęciami.
- Piksele obrazu zostały znormalizowane do zakresu wartości $[0, 1]$, aby umożliwić szybsze uczenie modelu.

Przetwarzanie danych diagnostycznych:

- Dla cech kategorycznych zastosowano O**ne-Hot Encoding**.
- Cechy liczbowe (np. breast density) zostały znormalizowane przy użyciu **Min-Max Scaling**.
- Braki w danych zostały uzupełnione:
  - UNKNOWN w danych kategorycznych zostało zakodowane jako osobna kategoria.
  - Wartości brakujące w danych liczbowych zostały uzupełnione medianą kolumny.

Podział na zbiory treningowy i testowy:

- Dane zostały podzielone na zbiory:
  - **Treningowy**: 80% danych.
  - **Testowy**: 20% danych.
- Przy podziale uwzględniono zrównoważenie klas (benign/malignant).

3 Wykonanie sieci

W tym punkcie znajduje się opis stworzonych sieci neuronowych. W ramach projektu stworzono dwa modele hybrydowe, które łączą analizę obrazów mammograficznych z dodatkowymi cechami diagnostycznymi. Różnica między modelami polega na zastosowanej metodologii treningu:

1. **Model hybrydowy z pełnymi epokami** - sieć trenowana przez ustaloną liczbę epok (np. 50), niezależnie od wyników walidacji.
2. **Model hybrydowy z EarlyStopping** - sieć trenowana z mechanizmem wczesnego zatrzymania, który kończy trening, gdy wyniki walidacyjne przestają się poprawiać.

3.1 Architektura modelu hybrydowego

Każdy z modeli wykorzystuje dwie równoległe ścieżki przetwarzania:

- **Ścieżka obrazu:** konwolucyjna sieć neuronowa (CNN), która wydobywa kluczowe cechy wizualne z obrazów mammograficznych.
- **Ścieżka diagnostyczna:** warstwy gęste (Dense), które analizują dane dodatkowe (np. gęstość piersi, typ mas, czy rozkład zwapnień).

3.1.1 Szczegóły architektury

Ścieżka obrazu (CNN):

- Wejście: Obrazy o rozmiarze 512x512x3
- Warstwy konwolucyjne:
  - 3 warstwy konwolucyjne:
    - 32, 64, 128 filtrów o rozmiarze 3x3.
    - Aktywacja ReLU (Rectified Linear Unit).
    - Padding SAME.
  - Po każdej warstwie zastosowano **MaxPooling** 2x2, który zmniejsza wymiary map cech.
- Flatten: Spłaszczenie map cech po ostatniej warstwie konwolucyjnej.
- Warstwy gęste:
  - 128 neronów, aktywacja ReLU (Rectified Linear Unit).
  - Dropout z wartością 0.5 dla regularyzacji.

Ścieżka diagnostyczna (Dense):

- Wejście: Dane o rozmiarze 96 cech (po przetworzeniu - One-Hot Encoding i normalizacja).
- Warstwy gęste:
  - Pierwsza warstwa: 16 neronów, aktywacja ReLU (Rectified Linear Unit).
  - Druga warstwa: 8 neuronów, aktywacja ReLU (Rectified Linear Unit).

Łączenie ścieżek:

- Połączenie cech z obu ścieżek za pomocą warstwy concatenate.
- Dalsze przetwarzanie:
  - Warstwa gęsta: 64 neurony, aktywacja ReLU (Rectified Linear Unit).
  - Dropout z wartościa 0.5.
  - Wyjście: 1 neuron z aktywacją sigmoid (klasyfikacja binarna: łagodna/złośliwa zmiana).

3.1.2 Kompilacja modelu

- Loss: Binary Crossentropy (strata krzyżowo-entropijna).
- Optymalizator: Adam (Adaptive Moment Estimation) (learning rate = 0.001).
- Metryki:
  - Dokładność (accuracy),
  - Średni błąd kwadratowy (Mean Squared Error - MSE).

3.1.3 Różnice między modelami

1. **Model hybrydowy z pełnymi epokami**
   - Trening sieci odbywa się przez ustaloną liczbę epok (np. 50), niezależnie od tego, czy wyniki walidacyjne ulegają pogorszeniu.
   - Jest bardziej narażony na przeuczenie (overfitting), ponieważ może kontynuować trening po osiągnięciu najlepszego wyniku walidacji.
2. **Model hybrydowy z EarlyStopping**
   - Zastosowano mechanizm wczesnego zatrzymania (EarlyStopping):
     - Monitorowanie strat walidacyjnych (val_loss).
     - Trening kończy się, gdy przez 5 kolejnych epok strata walidacyjna nie poprawia się o więcej niż 0.001.
   - Chroni model przed przeuczeniem, ale może zakończyć trening z mniejszą liczbą epok.

4 Przeuczenie sieci (overfitting)

Przeuczenie sieci neuronowej występuje, gdy model uczy się zbyt dobrze na zbiorze treningowym, zapamiętując specyficzne wzorce, a nie uogólniając ich. W efekcie sieć osiąga wysoką skuteczność na danych treningowych, ale jej wyniki na zbiorze testowym i w realnych zastosowaniach są znacznie gorsze.

4.1 Objawy przeuczenia

- Duża różnica między dokładnością (accuracy) na zbiorze treningowym i walidacyjnym:
  - Wysoka dokładność na treningu, niska na walidacji.
- Niższa jakość predykcji na zbiorze testowym (test loss).
- Brak poprawy metryk walidacyjnych mimo dalszego treningu:
  - Krzywa strat walidacyjnych (val_loss) przestaje maleć, a czasem nawet rośnie.

4.2 Zagrożenia przeuczenia

- Niska użyteczność modelu w rzeczywistych zadaniach:
  - Model, który zapamiętał dane treningowe, może nie rozpoznać nowych, nieznanych wzorców.
- Większa podatność na szum:
  - Sieć uczy się specyficznych detali i błędów w danych treningowych, które nie mają znaczenia w praktyce.
- Strata zasobów i czasu:
  - Zbyt długi trening przy braku poprawy wyników to marnowanie zasobów obliczeniowych.

4.3 Zabezpieczenia przed przeuczeniem

W projekcie zastosowano kilka mechanizmów redukujących ryzyko przeuczenia:

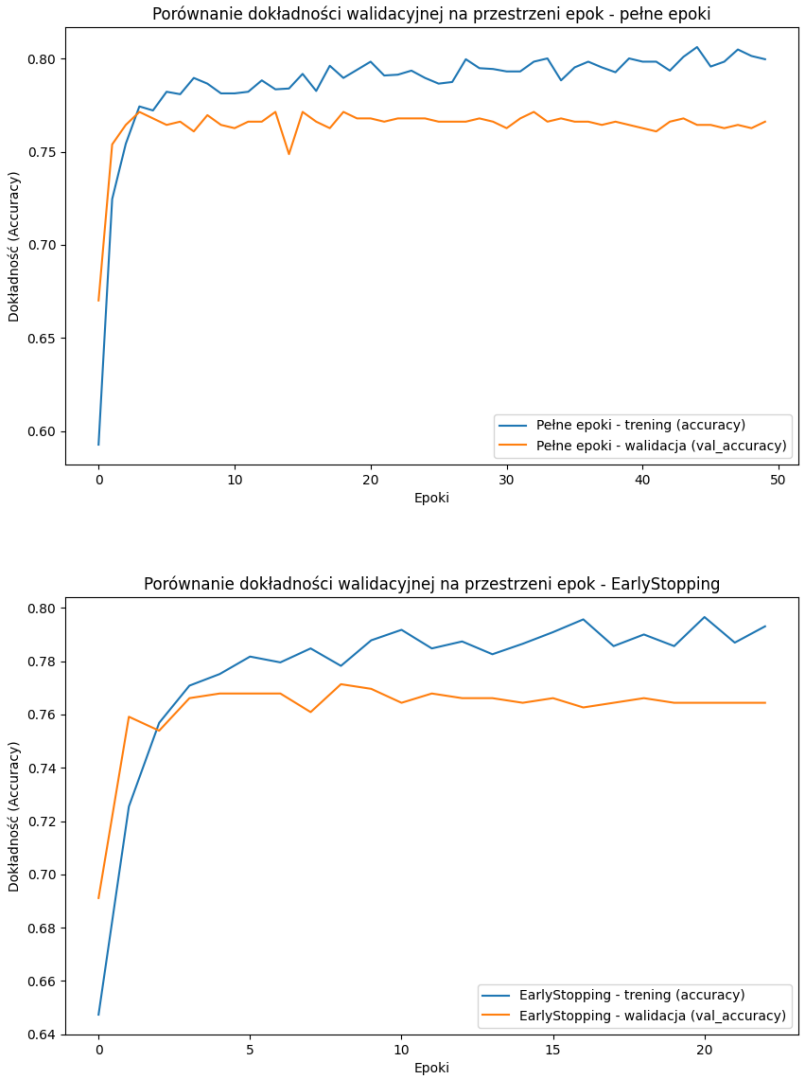
- **EarlyStopping** (dla jednego z modeli):
  - Trening kończy się automatycznie, jeśli strata walidacyjna (val_loss) przestaje się poprawiać przez 5 kolejnych epok.
  - Zapobiega zbędnemu dalszemu treningowi, który mógłby prowadzić do przeuczenia.
- **Dropout:**
  - Mechanizm regularyzacji, który w każdej iteracji losowo wyłącza część neuronów w warstwach gęstych (Dense).
  - W naszym modelu Dropout wynosi:
    - 0.5 w ścieżce obrazu,

    * 0.5 w warstwie połączonych cech (concatenate).

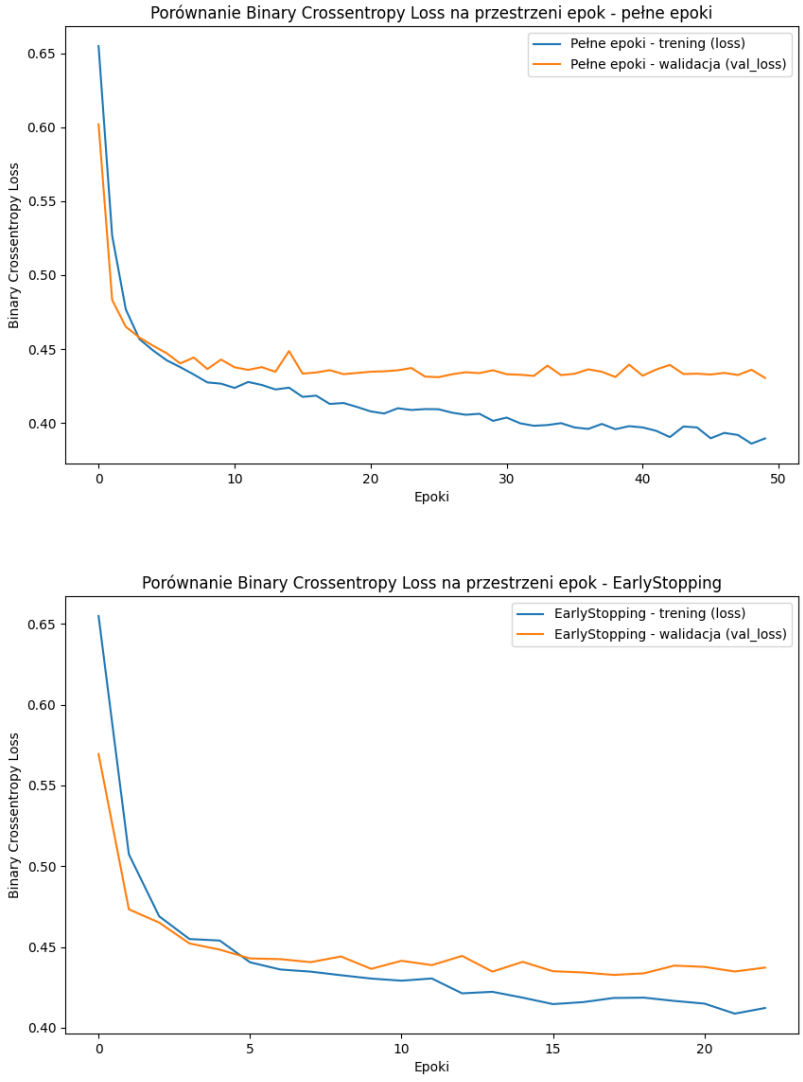
- **Podział zbioru danych:**
  - Dane zostały podzielone na niezależne zbiory treningowy i testowy, co pozwala ocenić zdolność generalizacji sieci.
- **Zastosowanie regularyzacji w strukturze modelu:**
  - Wybór optymalizatora Adam (Adaptive Moment Estimation) z kontrolowaną wartością learning rate=0.001, co zmniejsza ryzyko zbyt szybkiego zapamiętywania danych.

5 Przedstawienie wyników

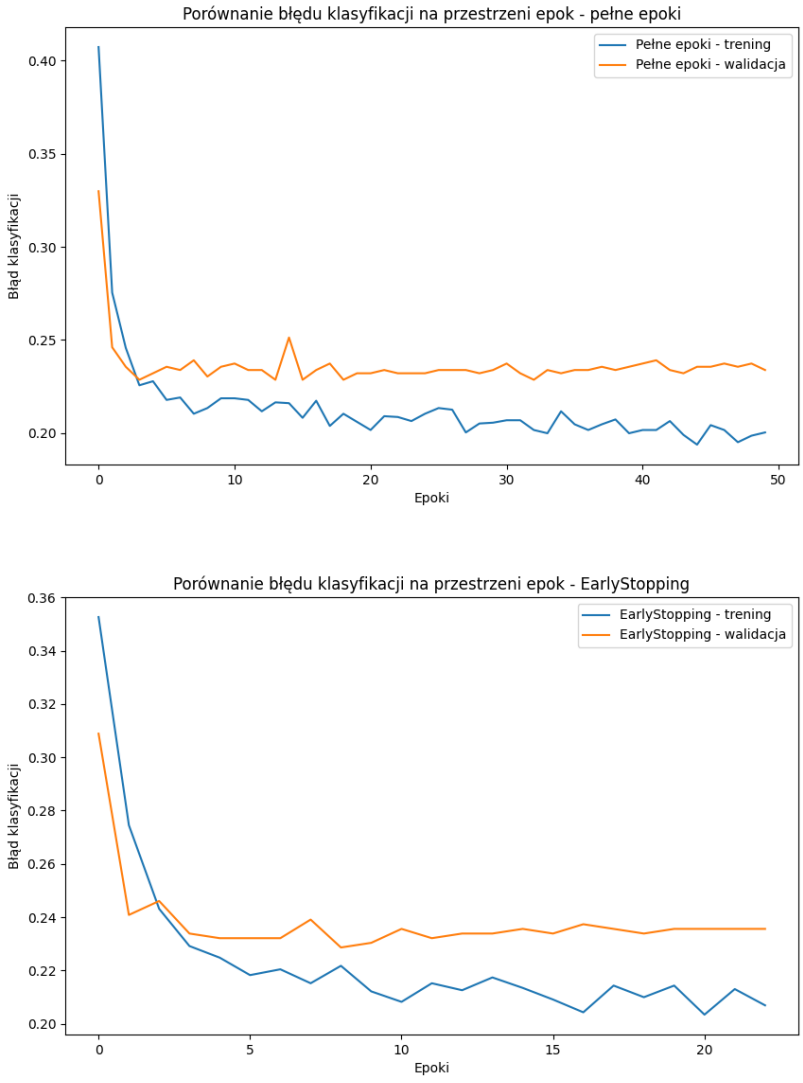
Poniżej znajdują się wykresy ukazujące wyniki badań dla dwóch sieci - uczącej się na pełnej liczbie epok oraz wykorzystującej metodę Early Stopping. Przedstawione wyniki to kolejno: dokładności, Binary Crossentropy, błędy klasyfikacji, MSE - wszystkie na przestrzeni epok, a także macierz pomyłek.



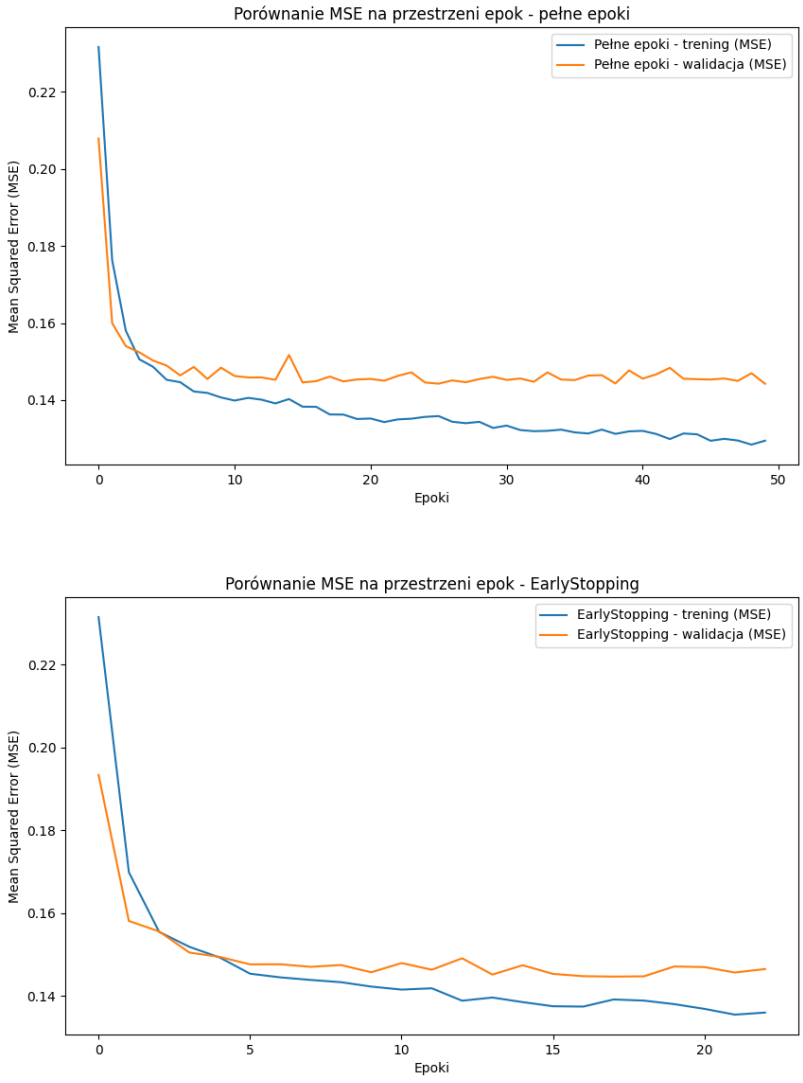
Rysunek 1: Wykresy dokładności na przestrzeni epok



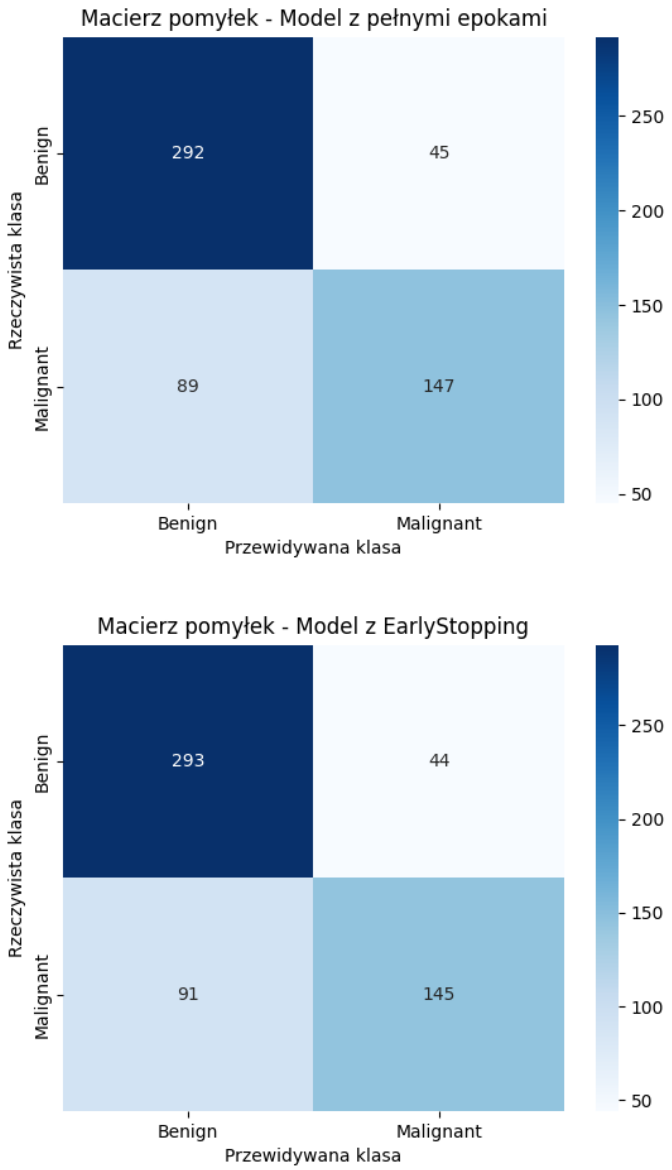
Rysunek 2: Wykresy Binary Crossentropy na przestrzeni epok



Rysunek 3: Wykresy błędu klasyfikacji na przestrzeni epok



Rysunek 4: Wykresy MSE na przestrzeni epok



Rysunek 5: Macierze pomyłek

6 Wnioski

Wyniki są zadowalające, zważywszy na złożoność modelu - model bierze pod uwagę zarówno dane z obrazów, jak i z parametrów diagnostycznych, czyniąc go modelem hybrydowym. Ze względu na relatywnie duży rozmiar badanych zdjęć (512x512) czas trenowania sieci ograniczał się **do 5 godzin** - tyle wyniosło w przypadku sieci uczącej się na pełnej liczbie epok.

Czas uczenia sieci dałoby się skrócić, jednakże należałoby zmniejszyć rozdzielczość wykorzystywanych zdjęć, co mogłoby wpłynąć negatywnie na dokładność sieci.

Przy tworzeniu sieci pojawiły się następujące problemy, związane z niepełnością danych:

- **Brakujące zdjęcia** - część danych nie posiadała zdjęcia, co zostało rozwiązane poprzez wygenerowanie czarnych obrazów na ich miejsce.
- **Brakujące dane diagnostyczne** - część etykiet była pozbawiona niektórych wartości, zostały one zastąpione medianą wartości pozostałych odpowiadających etykiet (metoda imputacji średniej)

Na wykresie Binary Crossentropy Loss widoczne jest **skuteczne działanie metody Early Stopping w zapobieganiu zjawiska przeuczenia**. Na pierwszym z wykresów - z pełną liczbą epok - wyniki dla zestawu treningowego z każdą kolejną epoką są coraz bardziej oddalone od wyników dla zestawu walidacyjnego. Metoda EarlyStopping, która w odpowiednim momencie zatrzymała uczenie się sieci, zapobiegła zwiększaniu się różnicy pomiędzy zestawami danych, co byłoby oznaką przeuczenia sieci.

Stworzona sieć neuronowa pozwalająca wykryć występowanie raka piersi na podstawie zdjęć mammograficznych **ma ogromny potencjał w życiu codziennym**. Dobrze opracowana sieć, wytrenowana na dużym zbiorze danych, ma szansę być wykorzystywana w szpitalach i punktach diagnostycznych tym samym ułatwiając prace lekarzy i potencjalnie zwiększając ilość pacjentów ze zdiagnozowanym nowotworem.

Na sieci neuronowej nie powinno się polegać w 100%, natomiast może ona służyć personelowi medycznemu, celem wspomagania procesu decyzyjnego, na podstawie którego można zlecić pacjentowi udanie się na bardziej zaawansowane badania.

Spis rysunków