Rozpoznawanie chorób serca na podstawie danych medycznych - raport

1. DANE I PROBLEM

Dataset: UCI Heart Disease Dataset - 303 próbki pacjentów z 13 cechami medycznymi (wiek, płeć, typ bólu w klatce piersiowej, ciśnienie krwi, cholesterol, cukier na czczo, EKG, maksymalne tętno, dusznica wysiłkowa, depresja ST, nachylenie ST, liczba głównych naczyń, defekt talasemii).

Target: Klasyfikacja binarna - konwersja z oryginalnej wieloklasowej (0-4) na binarną: 0: brak choroby, 1: choroba obecna (stopnie 1-4). Dane są lekko niezbalansowane, co jest kompensowane przez pos\_weight w funkcji straty.

Podział danych: Stratified split 70/15/15 (trening/walidacja/test) z normalizacją StandardScaler i automatycznym obliczaniem wag klas.

2. ARCHITEKTURA MODELU

Sieć neuronowa feed-forward:

* Input: 13 cech medycznych po normalizacji StandardScaler
* Warstwy ukryte: 64 → 32 → 16 neuronów z ReLU i Batch Normalization
* Output: 1 neuron z raw logits (bez aktywacji)
* Regularyzacja: Dropout adaptacyjny (0.5 → 0.5 → 0.25), Batch Normalization, Weight Decay (0.001)
* Inicjalizacja: Kaiming Uniform dla warstw liniowych

Optymalizacja:

* Optimizer: AdamW z learning rate 0.001 i weight decay 0.001
* Scheduler: ReduceLROnPlateau (factor=0.5, patience=10, min\_lr=1e-6)
* Early Stopping: patience=15 epok z przywracaniem najlepszych wag
* Loss: BCEWithLogitsLoss z automatycznym pos\_weight dla niezbalansowanych danych
* Gradient Clipping: max\_norm=1.0 dla stabilności uczenia

3. UZASADNIENIE ARCHITEKTURY

Dane medyczne charakteryzują się złożonymi nieliniowymi zależnościami między cechami. Kombinacja wieku, cholesterolu i ciśnienia krwi może mieć synergiczny wpływ na ryzyko choroby serca, którego regresja logistyczna nie jest w stanie uchwycić ze względu na swoje liniowe ograniczenia.

Progresywne zmniejszanie neuronów (64→32→16) pozwala na hierarchiczne uczenie się - od podstawowych wzorców do abstrakcyjnych kombinacji cech. Pierwsza warstwa (13→64) umożliwia uczenie bogatszych reprezentacji, następnie stopniowa kondensacja informacji.

Kompaktowość odpowiednia dla małego datasetu (303 próbki) - większa architektura prowadziłaby do overfittingu. Trzy warstwy ukryte stanowią optymalny balans między ekspresyjnością a ryzykiem przeuczenia.

Regularyzacja wielopoziomowa:

* Dropout adaptacyjny: 50% w początkowych warstwach, 25% w ostatniej warstwie ukrytej
* Batch Normalization: stabilizuje uczenie i przyspiesza konwergencję
* Weight Decay: L2 regularyzacja w optimizerze AdamW
* Early Stopping: automatyczne zatrzymanie przed overfittingiem
* Gradient Clipping: zapobiega eksplodującym gradientom

BCEWithLogitsLoss z pos\_weight automatycznie kompensuje niezbalansowanie klas, obliczając wagę jako stosunek próbek negatywnych do pozytywnych, co poprawia wykrywanie choroby serca.

Przewaga nad metodami ensemble - sieć neuronowa lepiej generalizuje na nowych danych i zapewnia spójne prawdopodobieństwa przez funkcję sigmoid zastosowaną do raw logits, co jest kluczowe w kontekście medycznym gdzie potrzebujemy interpretowalne wyniki wspierające decyzje diagnostyczne.

4. WYNIKI I WNIOSKI

Typowe metryki osiągane przez model:

* Test Accuracy: ~87-88%
* Precision: ~0.87-0.89 (klasa "choroba")
* Recall: ~0.86-0.90 (klasa "choroba")
* F1-Score: ~0.87-0.88
* Balanced performance: podobne wyniki dla obu klas dzięki pos\_weight

Charakterystyka uczenia:

* Stabilna konwergencja dzięki AdamW i learning rate scheduler
* Effective regularization zapobiegająca overfittingowi
* Automatyczne zatrzymanie przez early stopping (typowo ~100 epok)

Kluczowe zalety:

* Automatyczny feature engineering przez warstwy ukryte z batch normalization
* Skuteczna regularyzacja przeciwko overfittingowi przez multiple techniques
* Obsługa niezbalansowanych danych przez pos\_weight w funkcji straty
* Probabilistic output umożliwiający elastyczne progi diagnostyczne
* Numeryczna stabilność przez BCEWithLogitsLoss

Ograniczenia:

* Mały dataset ogranicza złożoność modelu i możliwość głębszej architektury

Wnioski: Model osiąga competitive wyniki (~88% accuracy). Kluczowym sukcesem jest uzyskanie interpretowalnych prawdopodobieństw przez sigmoid(logits) umożliwiających lekarzom kontekstowe dostosowywanie progów decyzyjnych w zależności od tolerancji ryzyka.

5. DALSZY ROZWÓJ

Ulepszenia obecnego modelu:

* K-fold cross-validation: bardziej wiarygodna ocena generalizacji
* Ensemble methods: połączenie z Random Forest/XGBoost przez voting/averaging

Ulepszenie rozszerzenia do klasyfikacji wieloklasowej (powrót do oryginalnego problemu 0-4):

* Zmiana architektury: output 5 neuronów z softmax (klasy 0-4 stopnia zaawansowania)
* Loss function: CrossEntropyLoss z class\_weight dla niezbalansowanych klas oryginalnych
* Metryki: Macro/Micro F1, per-class precision/recall, confusion matrix 5x5