ANN for medical diagnosis

April 5, 2025

1 Introduction

This report analyzes neural network implementations for three medical classification tasks: heart disease classification, diabetes prediction, and breast cancer prediction. All implementations use TensorFlow to build and train multi-layer neural networks on datasets containing patient attributes and clinical measurements. The approach involved data preparation, model design, training, and performance evaluation, focusing on addressing key challenges including missing value handling, feature normalization, and overfitting prevention.

2 Heart Disease Classification

2.1 Dataset and Preprocessing

The UCI Heart Disease Dataset contains demographic features (age, sex), clinical measurements (blood pressure, cholesterol, maximum heart rate), and test results, with the target variable indicating heart disease severity (0-4). Preprocessing steps included:

- Removing identifier columns
- Imputing missing values using mean (numerical) and mode (categorical) by disease group.(the values are imputed after seperating the dataset as per level of heart disease since the distribution of the features is different in different levels of disease)
- One-hot encoding categorical features and standardizing numerical features.(since the neural network can't take categorical inputs the categorical columns are converted to numerical values using one-hot encoding and the distributions of the numerical columns are gaussian so the standard scaler is preferred.)

2.2 Model Architecture and Training

A feedforward neural network was implemented with: input layer matching preprocessed features, first hidden layer (32 neurons, ReLU), second hidden layer (16 neurons, ReLU), and output layer (5 neurons, softmax). The model used sparse categorical crossentropy loss and SGD optimizer. Early stopping prevented overfitting by monitoring validation loss and retraining for the optimal number of epochs.

Final performance metrics showed strong classification capability:

- Training accuracy: 88.85%
- Test accuracy: 75.00%

3 Diabetes Classification

3.1 Dataset and Preprocessing

The Pima Indians Diabetes Database includes attributes such as pregnancies, glucose, blood pressure, BMI, and age, with a binary target variable (0: no diabetes, 1: diabetes). Preprocessing involved:

- Replacing implausible zero values with NaN(since zero value is not possible for these attributes it is considered as missing value and imputed using median)
- Split-based imputation (diabetic vs. non-diabetic groups)
- Normalization using MinMaxScaler (since the distribution of the columns is highly skewed min-max scaler is preferred)

3.2 Model Architecture and Training

The binary classification network included: input layer (8 features), first hidden layer (8 neurons, ReLU), second hidden layer (4 neurons, ReLU), and output layer (1 neuron, sigmoid). The model used binary crossentropy loss and Adam optimizer, with similar early stopping optimization.

Final performance metrics showed strong classification capability:

- Training accuracy: 83.55%
- Test accuracy: 81.16%
- Training recall: 72.76%
- Test recall: 74.54%

4 Breast Cancer Classification

4.1 Dataset and Preprocessing

The Wisconsin Breast Cancer dataset contains 30 features derived from digitized breast mass images, with diagnosis classified as malignant (M=1) or benign (B=0). The dataset had 569 samples with no missing values. Preprocessing included:

- Mapping categorical diagnosis values to binary (M→1, B→0)
- Feature standardization using StandardScaler

4.2 Model Architecture

A binary classification neural network was implemented with:

- Input layer: 30 features
- First hidden layer: 32 neurons with ReLU activation
- Second hidden layer: 16 neurons with ReLU activation
- Output layer: 1 neuron with sigmoid activation

The model was compiled with binary crossentropy loss, SGD optimizer, and multiple metrics (accuracy, precision, recall).

4.3 Training and Optimization

The model was initially trained for 750 epochs with batch size 32, monitoring both training and validation loss. Early stopping optimization identified the epoch with minimum validation loss ($\min_i ndex$), $and a new model with$

Final performance metrics showed strong classification capability:

- Training accuracy: 99.12%
- Test accuracy: 98.24%
- Training recall: 98.22%
- Test recall: 97.67%

5 Conclusion

This project successfully implemented neural networks for three medical classification tasks, demonstrating:

- Domain-specific preprocessing is crucial for medical data quality
- Early stopping effectively prevents overfitting across all tasks
- Deeper architectures benefit tasks with more complex feature relationships
- the accuracy and recall score may appear slightly different after executing again due to random initialization of parameters for neural network in fit function
- all the plots are provided seperately .

6 Contributions

- Varshith(230150023,s.varshith@iitg.ac.in): data preprocessing for diabetes,breast cancer and EDA of all data sets.
- Kalyan Ram (230150005,r.bandi@iitg.ac.in): data preprocessing of heart disease dataset, designing of neural network and report.
- Pranav (230150016, r.mogali@iitg.ac.in): model building and training of diabetes, breastcancer .