Machine Vs Deep Learning Performance for Heart Disease prediction

Mordechai Mushkin, Dmitry Strizhak, Nathan Schumann

Table of Contents

1. Keywords:

ML – Machine Learning

DL – Deep Learning

LR- Logistic Regression

MES- Mean Error squared.

EDA- Exploratory Data Analysis

XGBoost - Extreme Gradient Boosting

ROC - Receiver Operating Characteristic

AUC - Area Under the Curve

SVM- Support Vector Machines

MLP - Multi-layer Perceptrons

NN- Neural Network

RF- Random Forest

2. Abstract:

This project presents a comprehensive study in predictive analytics by leveraging a Kaggle dataset to address a classification problem. The work begins with EDA to understand underlying patterns and data quality issues. A preprocessing is applied without any data cleaning. Initial results are obtained by using a baseline ML algorithms of LS, XGBoost and RF decision tree, evaluated via multiple relevant metrics. The project explores the impact of hyperparameter tuning by varying key parameters across different settings. Additional experiments involve refining the network architecture to improve convergence speed and overall performance. The model is assessed the model's by altering data balance levels and applying alternative dimensionality reduction techniques. Overall, the project provides insights of preprocessing, model selection and systematic experimentation contribute to optimizing predictive performance in the given data.

3. Introduction:

Heart disease remains one of the leading causes of mortality in the United States, making early detection of heart disease a critical public health objective. The dataset is the Cleveland Heart Disease dataset taken from the UCI repository and located at Kaggle as well, contributed to develop and evaluate ML models capable of identifying individuals at high risk.

The project begins with an extensive EDA to uncover patterns, assess feature distributions, and identify potential anomalies or missing values. For the modelling phase, several ML including the baseline LR and the advanced RF and XGBClassifier are deployed to establish performance benchmarks. ANN architectures are subsequently employed trying to improve predictive accuracy. The performance of those models is evaluated using a range of metrics. Standard classification metrics such as accuracy, precision, recall, and F1 score are used to assess the models' ability to correctly predict heart disease risks. Further experiments explore the impact of hyperparameter tuning, dataset modifications, and architectural adjustments on model performance.

We have found that the dataset is tide, and requires no modification. It turned out that for this dataset good performances were obtained by a basic ANN with one hidden layer, and various attempts to further improve the performances by modifying the network architecture did not improve the performances, as well as for this dataset, ANN does not have substantial advantage over start of the art ML algorithms, such as RF.

4. Tables and Figures:

5. Literature review

5.1 Studies revies on heart disease prediction using ML and DL

5.1.1 (Alshraideh, et al., 2024) The researchers used the JUH Heart Disease dataset from Jordan University Hospital (486 cases with 58 attributes), the study applies multiple machine learning algorithms—including SVM, RF, decision tree, naive Bayes, and KNN—with particle swarm optimization (PSO) for feature selection. Models are rigorously evaluated via 10-fold cross-validation using metrics such as accuracy, precision, recall, F1-score, and ROC AUC. Notably, SVM with PSO achieves a 94.3% accuracy. The study concludes that optimized machine learning models significantly enhance early heart disease prediction, supporting timely diagnosis and improved patient outcomes.

5.1.2 (Rojek, Kotlarz, w Kozielski, Jagodzi ´nski, & Królikowski, 2024) Using a patient dataset capturing diverse clinical parameters, this study develops an AI-based tool to predict heart attack risk for preventive medicine. Multiple machine learning models—Linear SVC, LR, KNN, and RF —were compared to determine personalized risk and identify a minimal feature set (heart rate, age, BMI, cholesterol). Evaluation revealed that LR, while moderately predictive, provided the most accurate results for initial screening. The system offers a rapid, cost-effective, and non-invasive predictive analysis, enabling early intervention and improved preclinical care.

5.1.3 (Aghamohammadi, Madan, Ki Hong, & Watson, 2019) The paper proposes a novel classification method that combines a Genetic Algorithm (GA) with an Adaptive Neural Fuzzy Inference System (ANFIS) to predict heart attack risk using the Cleveland dataset (297 patients, 14 features). The system categorizes risk from no to very high and is evaluated using sensitivity, specificity, precision, accuracy, RMSE, and 9-fold cross-validation. Training and testing results indicate satisfactory performance, with significant RMSE reduction. Overall, explainable outputs and an Importance Evaluation Function (IEF) robustly reveal key predictive features, demonstrating transparent and effective diagnosis.

5.2 Previous studies common approaches: Across these four studies, traditional machine learning methods most frequently used for heart disease prediction include LR, RF, k-nearest neighbours, and SVM. We have come upon some papers calming the use of DL approaches including CNN, RNN, LSTM, GRU and hybrid CNN-GRU architectures , but the

reliability of those papers is questionable, since non of those techniques is suitable for tabular data-sets. Popular evaluation metrics across these works typically include accuracy, precision, recall, F1-score and AUC, with additional use of RMSE and k-fold cross-validation to ensure robustness.

6. Methodology:

6.1 Dataset description:

6.1.1 Data set details

This dataset contains information about different health and lifestyle factors that may influence heart attacks in the USA. It includes details like age, cholesterol, blood pressure, and smoking habits, along with outcomes like whether a heart attack occurred. The goal is to help identify potential risks and trends that can lead to better heart health awareness and prevention. It has 372,974 rows × 32 columns.

Variable Name	**Description**	**Values/Notes**
age	Age of the patient	(years)
sex	Gender	1 = male, 0 = female
cp	Chest pain type	0: Typical angina 1: Atypical angina 2: Non-anginal pain 3: Asymptomatic
trestbps	Resting blood pressure	(mm Hg)
chol	Serum cholesterol	(mg/dl)
fbs	Fasting blood sugar	1 = true, 0 = false
restecg	Resting ECG results	0: Normal 1: ST-T wave abnormality 2: Left ventricular hypertrophy
thalach	Max heart rate achieved	(bpm)
exang	Exercise-induced angina	1 = yes, 0 = no
oldpeak	ST depression (exercise vs. rest)	(numeric)
slope	Slope of peak exercise ST segment	0: Upsloping 1: Flat 2: Downsloping

ca	Major vessels colored by fluoroscopy	(0–3)
thal	Thalassemia	0: Normal 1: Fixed defect 2: Reversible defect
target	Heart disease diagnosis	0: No significant disease 1: Significant disease

Table 1: Predictors List

Variable	**Target-variable correlation value**
exang	-0.436757
cp	0.433798
oldpeak	-0.430696
thalach	0.421741
ca	-0.391724
slope	0.345877
thal	-0.344029
sex	-0.280937
age	-0.225439
trestbps	-0.144931
restecg	0.137230
chol	-0.085239
fbs	-0.028046

Table 2: Predictors Target Correlation

6.1.2 Dataset preprocess

There are no missing values. Most columns are numeric including binary columns. The non-binary categorial predictors were encoded in a one hop encoding. Most of the categorial predictors are binary. The non-binary categorial predictors are: cp, restecg, slope, and thal. We scaled the numeric predictors to [0 1] range. Later on while trying to improve NN performance we changes the balance of the training and validation subsets by removing 50% and 90% of the positive-target observations in order to ensures that each class is sufficiently represented. We reduce dimensions by removing the predictors trestbps, restecg, chol, and fbs, which have the lowest correlation with the target, decreased the metrics.

6.1.3 Evaluation of the predictor's characteristics

The evaluation of corelation between predictors is visualized by the "Correlation Heat Map" (Figure 1.) and analytically calculated. The results indicate that exang, cp, oldpeak, and thalach have moderate correlation with the target. ca, slope, thal, sex, and age have weak correlation with the target.

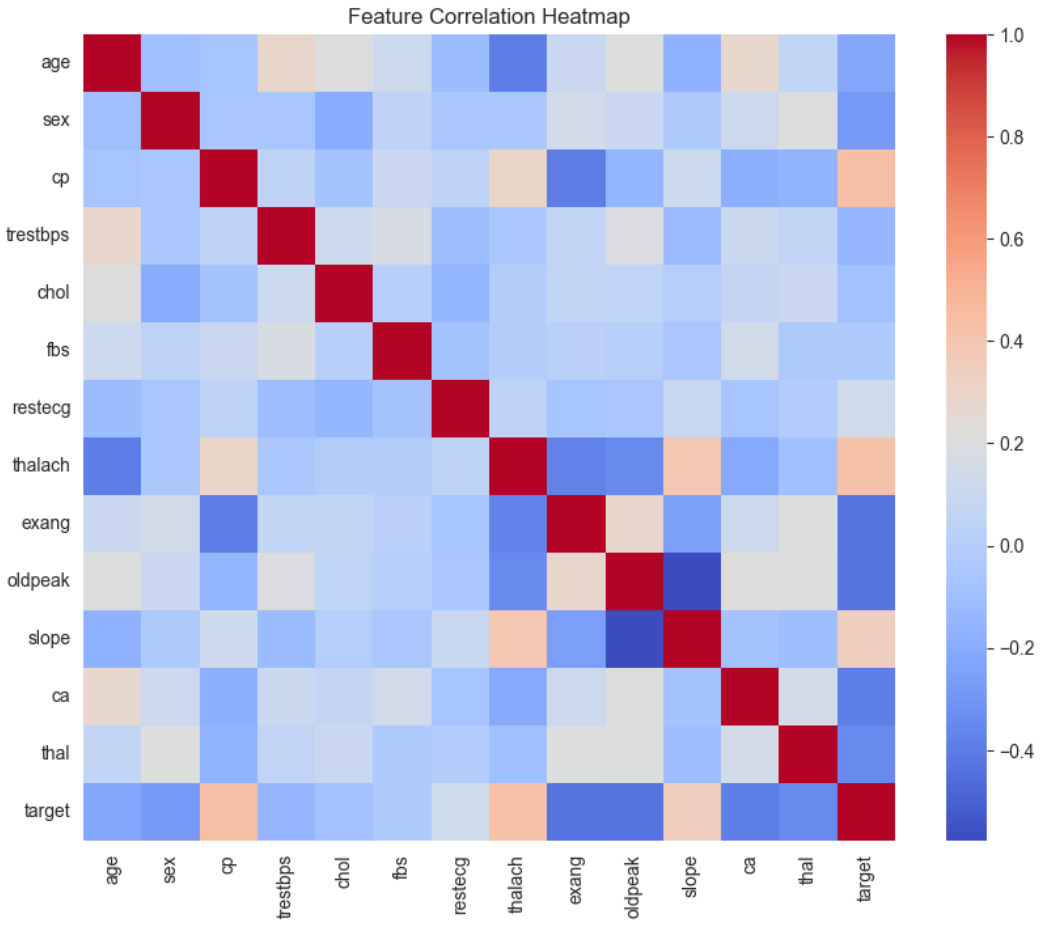

Figure 1: Feature Correlation Heatmap

trestbps, restecg, chol, and fbs have no correlation with the target. The results are shown at Table 2. slope-oldpeak pair, might have moderate collinearity, all other pairs have low or acceptable collinearity. Checking of the balance of the target predictor, shows that the

target is well balanced. Most of the predictors are categorial. At a later stage, when we preformed DL modeling we tried to remove the corelated predictors trestbps, restecg, chol, and fbs from the dataset.

Collinearity has been checked as well, and found slop-oldpeak is the one that maybe considered to have almost moderate collinearity. All the rest pairs have low or accesible collinearity.
Most of the predictors are categorial. The only numeric predictors are: age, trestbps, oldpeak, and ca. Scale the numeric predictors to the 0 to 1 interval. Most of the categorial predictors are binary. The non-binary categorial perdictors are: cp, restecg, slope, and thal.

6.1.4 Dataset split

The dataset is divided into a 60% training subset, 20% Validation subset, and test 20% test subset.

6.2 ML Modelling:

6.2.1 Decision Tree

We used the XGBoost classifier which implements gradient boosting for classification using decision trees. It's used to predict categorical predictors as well as numerical.

6.2.2 Randon Forest

An RF with a majority vote over 100 trees.

6.2.3 Logistic Regression

The model was calculated with a max iteration of 10,000,000

6.3 DL Modelling

Staring with basic NN with a single input layer, 64 neurons in hidden layer with a ReLU activation function and single output layer with a Sigmoid activation function and binary-cross-entropy as a loss function.
The weights are restored to the epoch with the best validation loss. It was implemented as a callback that finds the epoch with the best loss value and set the final weight of the model to the weights at this epoch. The best weights from set at epoch 43 with a loss value of 0.4472.

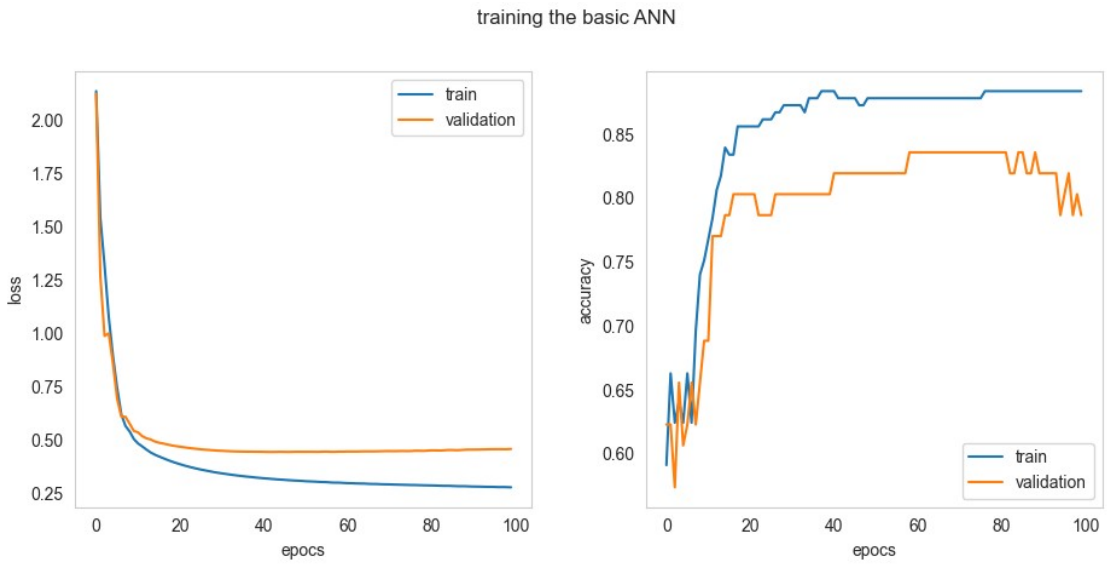

Figure 2: Basic NN training & accuracy & loss per Epoch

The hyperparameters to be checked are the depth, width, and activation functions. We try to find the optimal hidden layer width by changing the widths to 20,30,40,60,80,100 neurons. We selected 60 as the layer's width. We added a second hidden layer, we tried widths of 4,6,8,10 and 12 neurons. 2nd layer at a width of 10 achieved the best performance. The comparison between the performance of a single layer model and two hidden layers resulted similar performance. Adding a third hidden layer and testing at widths of 16, 20, 24, 28, 32 resulting a width of 28 with the best performance. The metrics of 3 hidden layers model are similar to single and two layers. In order to reduce complexity a simple model is preferred. We tried to improve by removing 55 (30%) of the potential outliers and 6 (3%) of the potential outliners from the training subset. It didn't lead to a more robust training. We continue "dropping" neurons during training, It didn't lead to an improvement. A Hinge loss function results are similar to Cross Entropy.

7. Results:

7.1 ML Results

Gradient boost tree-based classifier reached test accuracy of 0.77. RF reached test accuracy: 0.836. Logistic regression reached test accuracy of 0.803 which is reflected in Table 3.

Regression type	**Accuracy**
Logistic regression	0.803
Random forest	0.836
XGBC	0.77

Table 3: ML Test Accuracy Comparison

RF got the best test accuracy. RF metrics are:

Metric	**Value**
Accuracy	0.8361
Recall	0.9062
Precision	0.8056
F1-Score	0.8529

7.2 DL Results

7.2.1 Changing different parameters over the three setups of NN

A basic DL model slightly worse than those of the RF model

Metric	**RF**	**Bacis ANN**
Accuracy	0.8361	0.82
Recall	0.9062	0.91
Precision	0.8056	0.78
F1-Score	0.8529	0.84

Table 4: Basic NN Vs RF Metrics

The charts below show the loss per epoch for each tested width:

- Figure 3 – NN with single hidden layer
- Figure 4 – NN with two hidden layers
- Figure 5 – NN with three hidden layers

The metrics of a single and two hidden layers are the same.

Best weights at epoch 30 got a loss of 0.4184.

Table 5. reflects comparison of some metrics of three different NNs

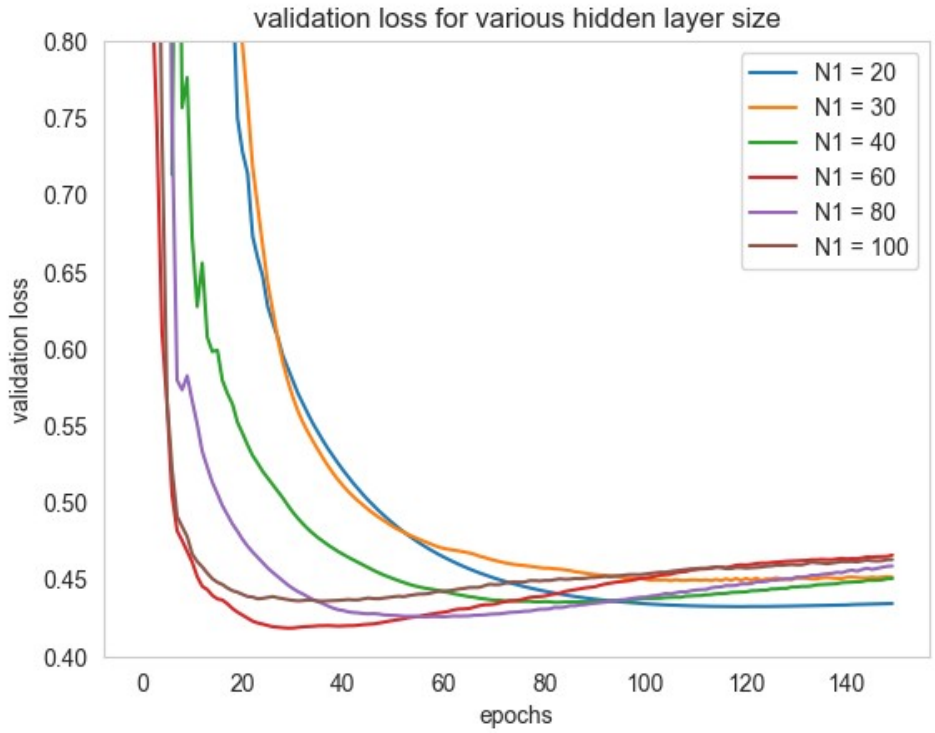

Figure 3: Single layer loss per epoch @ validation subset

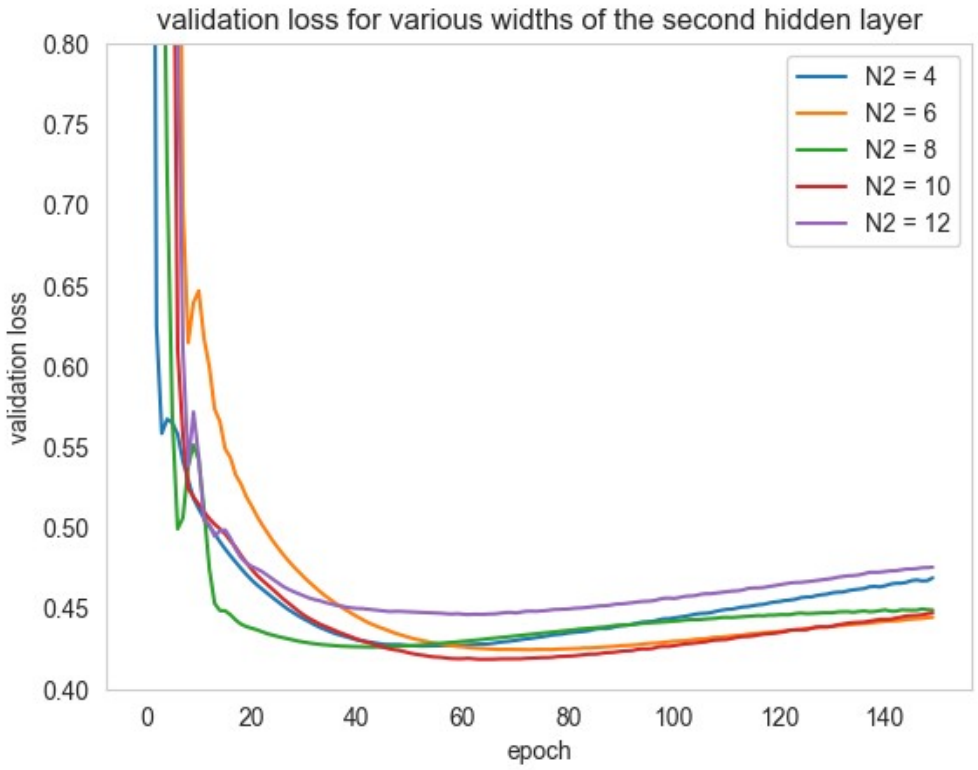

Figure 4: Two layers loss per epoch @ validation subset

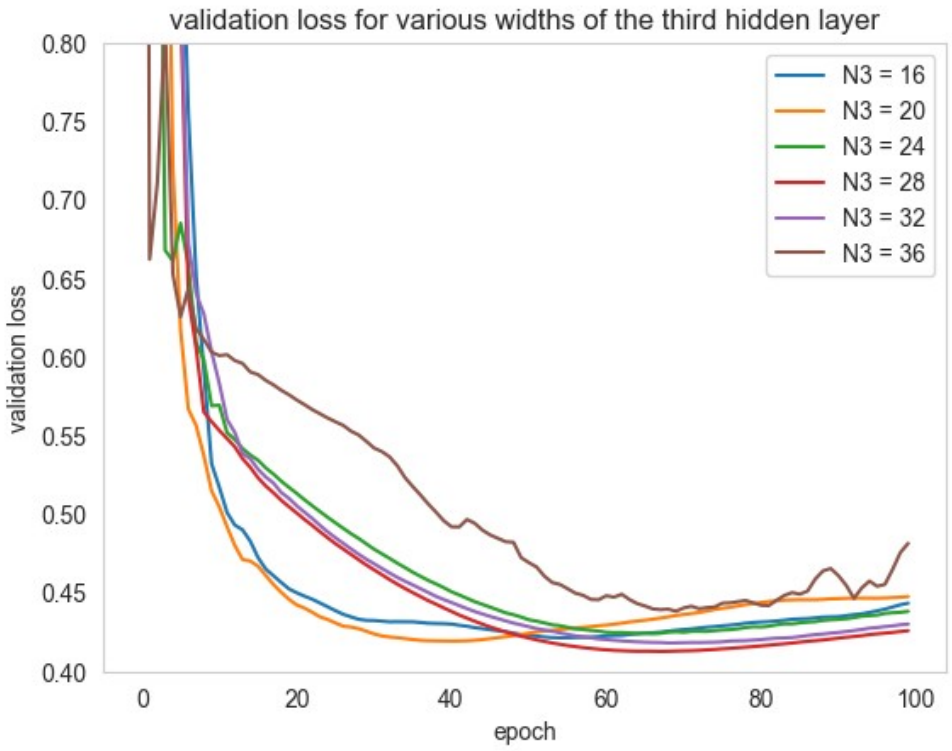

Figure 5: Three layers loss per epoch @ validation subset

Metric	**Single layer**	**Two layers**	**Three layers**
Accuracy	0.82	0.82	0.80
Recall	0.91	0.91	0.91
Precision	0.78	0.78	0.76
F1-Score	0.84	0.84	0.83

Table 5: Single, two and three layers metrics

We've tried to play with amount of outliers and dropout.

Below you can see Table 6. that reflects the performances before and after removing 30% of the potential outliers, 3% of the potential outliers and dropout accordingly:

Metric	**Origin**	**After removing 30% outliers**	**After removing 3% outliers**	**Dropout**
Accuracy	0.82	0.77	0.77	0.80
Recall	0.91	0.91	0.88	0.88
Precision	0.78	0.72	0.80	0.78

F1-Score	0.84	0.81	0.84	0.82

Table 6: metrics comparisons, outliners, dropout

7.2.2 Changing the balance of the dataset

The data set is well balanced. The negative/positive output ratio is 0.46/0.54. Therefore, changing the balance is expected to deterirate the performances. We have created modified dataset by removing 50% and 90% of the positive-target samples from the trainning and validation sets to check the metrics and got the next result.

For dataset modefied 50%:

The dataset is moderately imbalanced. The recall was highly degraded, while the precision is not affected. Consequently, the accuracy and the F1 score are also degraded.

For dataset modified 90%:

The dataset is highly imbalanced. The recall was completely degraded, while the precision is not affected. Consequently,the accuracy and the F1 score are also highly degraded.

See the Table 7. below:

Metric	**Original**	**Modified 50%**	**Modified 90%**
Accuracy	0.82	0.77	0.64
Recall	0.91	0.78	0.44
Precision	0.78	0.78	0.78
F1-Score	0.84	0.78	0.56

Table 7: Balanced train and evaluation subsets

7.2.3 Dimention reduction

Removing the predictors trestbps, restecg, chol, and fbs, which have the lowest correlation with the target, from the dataset

Given the dataset's limited number of predictors (13) and the relatively streamlined architecture of the neural network, there was minimal incentive to reduce dimensionality. Moreover, it was recognized that even weakly correlated predictors may offer valuable insights when considered collectively, making their removal potentially detrimental to model performance.

All merics of dimention reduction are reflected in Table 8.

Metric	**Original**	**Modified**
Accuracy	0.77	0.64
Recall	0.78	0.44

Precision	0.78	0.78
F1-Score	0.78	0.56

Table 8: Dimensions reduction results

8. Discussion and Concoctions:

The data set contains predictors that explain the target with a linear logistic regression. Techniques like, network architecture, balancing the data, dimensions reduction, Loss function modifications didn't improve the performance of the NN model. For this case NN is an "over keel".

9. Further Work:

Heart attacks prediction at this study and previous studies show that it can be done with good results. However, heart attacks prevention is more challenging and more important in terms of public health. Further work should include models that provide Inference. Once we can quantify how much each predictor contribute to the probability to suffers from heart attack, we might be able to prevent it. We should also look at data sets with other predictors which are not biological like environmental, social, and psychological.

10. References:

Bibliography

Aghamohammadi, M., Madan, M., Ki Hong, J., & Watson, I. (2019, December 4). Predicting Heart Attack through Explainable. *EasyChair Preprint*, *2093*. From https://link.springer.com/chapter/10.1007/978-3-030-22741-8_45

Kumar, S. G. (2021, August). A Machine Learning Approach for Heart Attack. *International Journal of Engineering and Advanced Technology (IJEAT)*, *10*(6). doi:10.35940/ijeat.F3043.0810621

Alshraideh, M., Alshraideh, N., Alshraideh, A., a Alkayed, Y., Al Trabsheh, Y., & Alshraideh, B. (2024, March 7). Enhancing Heart Attack Prediction with Machine Learning: A. *2024*. doi:https://onlinelibrary.wiley.com/doi/10.1155/2024/5080332

Rojek, I., Kotlarz, P., w Kozielski, M., Jagodzi ´nski, M., & Królikowski, Z. (2024, January 7). Development of AI-Based Prediction of Heart Attack Risk as an. *electronics*. doi:https://www.mdpi.com/2079-9292/13/2/272