International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Achieving Superior Predictive Performance through Deep Neural Networks for Enhanced Cardiovascular Disease Detection

Prof. (Dr.) Nikhil Srivastava[1], Pradipto Chatterjee[2]

[1]Professor and Dean, Major SD Singh University, Farrukhabad, UP, deanenggmsdsu@gmail.com
[2]*MTech Scholar, Major S D Singh University, Farrukhabad, UP, pradipto.chatterjee1@gmail.com*

ABSTRACT

Cardiovascular disease (CVD) is a heterogeneous group of heart disorders, primarily resulting from the deposition of plaque in arteries, high blood pressure, and poor lifestyle, thus being a major cause of mortality worldwide. Early detection of cardiovascular disease is very important since it enables timely intervention, brings about lifestyle modifications, and permits proper medical attention, thereby curbing the progression of the disease, avoiding severe complications such as myocardial infarction or stroke, and enhancing long-term outcomes and overall well-being. In this paper, we present a deep learning technique for the prediction of heart disease or cardiovascular disease (CVD) with increased accuracy, from a large and robust dataset. Traditional machine learning algorithms, such as Random Forest and Logistic Regression, have been able to attain moderate accuracy (70.2%) in the identification of heart disease in this dataset; however, our expert neural network model is far superior to these methods, with an accuracy rate of 73.67%. Our proposed neural network is a multi-layered architecture with L2 regularization, batch normalization, and dropout techniques to prevent overfitting while capturing complex patterns in the data. We have also used systematic hyperparameter tuning and early stopping by reducing the learning rate of our model. This study shows how deep learning can provide both superior predictive performance and interpretability through permutation importance analysis. Our study presents the potential of neural networks in healthcare and disease detection, particularly for CVD risk assessment, where even small enhancement in accuracy can have substantial impact in the clinical field.

Keywords: Cardiovascular Disease Detection, Heart Disease Detection, Deep Learning, Neural Network.

1. Introduction

CVDs happen to be one of the most crucial global health issues, accounting for around 30% of total deaths all over the world. Early detection of persons who are at risk is estimated to be a vital part of the preventive approach in order to lower mortality rates. The traditional risk assessment tools, although helpful, often are based on simplified statistical models that might not be able to reveal the complex interactions among risk factors. Recent developments in ML and AI have contributed to the improvement in the accuracy of prognosis of CVD[1][2]. Supervised learning algorithms such as Random Forests managed to get moderate precision of something about 70% in identifying at-risk patients[3]. However, these models might also lack the ability to entirely simulate the intricate relationships in the classic clotting parameters, lifestyle factors, and genetic predispositions. Deep Learning, a part of artificial intelligence, is based on the human brain's structure and it has a strong alternative way. Neural networks can automatically identify complicated patterns in the huge data sets and thus are able to make predictions which are more precise than those made by traditional methods of statistics. This study looks into the implementation of deep learning in the area of CVD prediction through a dataset that consists of a huge number of patient records.

The dataset used in this study includes a huge range of clinical and lifestyle factors, making it suitable for predictive models. The features we have selected for this study were chosen based on their strong relation with cardiovascular risk. Demographic Factors like Age and gender are some fundamental risk factors for CVD. The risk of cardiovascular disease generally increases with age due to natural and physiological changes in the cardiovascular system. Gender differences in cardiovascular disease are also sometimes interconnected and with biological differences between genders risk factors for CVDs increase. Hemodynamic Factors like Systolic and diastolic blood pressure are among the most critical factors for the prediction of cardiovascular disease. Increased systolic blood pressure makes the heart and blood vessels work harder, leading to the progression of atherosclerosis and an increase in the chance of heart failure and stroke. Cholesterol and glucose levels are metabolic factors that very closely and directly impact the probability of getting cardiovascular diseases. Dyslipidemia, an abnormal level of cholesterol made up of a lot of low-density lipoprotein (LDL) cholesterol and little high-density lipoprotein (HDL) cholesterol, is a potent factor in artery wall plaque formation. The same goes for high blood glucose levels which are a sign of prediabetes or diabetes, both of which contribute to the increase of cardiovascular risk. Height and weight, which are the main anthropometric measures used to calculate body mass index (BMI) and give us information about body composition and adiposity. Happily enough, obesity, mainly central obesity, tends to be a high-risk marker of cardiovascular disease due to poor insulin sensitivity, inflammation, and endothelial dysfunction. Smoking, consuming alcohol, physical activity etc. and many more lifestyle factors play a critical role in cardiovascular health. Smoking, for example, is causative in endothelial dysfunction and atherosclerosis, and on the other hand, the excessive intake of alcohol can provoke hypertension and cardiomyopathy. Physical inactivity is an independent risk factor for the development of obesity, diabetes, and cardiovascular diseases. The all-out evaluation for the cardiovascular risk profile is achievable by all these pieces put together.

Apart from the majority of the smaller datasets which thus are more specialized in clinical parameters, our dataset is wide enough as it holds not only the objective measurements (blood pressure, cholesterol) but also rather subjective lifestyle factors like (smoking, alcohol use). In this way, we get an exhaustive light on cardiovascular risk. The inclusion of both modifiable and non-modifiable risk factors is what gives the possibility of coming up with some models that can identify people at high risk despite seemingly normal clinical parameters, as well as, those who are most likely to be benefited from the lifestyle interventions. The dataset used in this study is fantastic in a few ways if we compare it with the previously used datasets. One of the major advantages of the dataset is that it offers a large sample size, which gives sufficient statistical power to detect meaningful relationships between features and outcomes. In the dataset, you will find nearly equal numbers of patients with and without cardiovascular diseases, and for this reason, it is preventing the bias in model development. The features which are mostly used for clinical data collection are the components making the dataset highly relevant for the establishment of practical prediction models to be integrated into healthcare practice. The wide variety of diagnosed patients with all age groups and genders, and the use of in-depth statistical analysis gives the assurance that the findings are transferable to similar demographic groups.

We have designed a customized neural network that handles feature data from the csv file to predict whether a person is having any cardiovascular disease or not. After training the model we have used hyperparameter tuning and regularization that helped us to achieve the test accuracy of 73.67%. This accuracy represents a significant improvement over the past ML based algorithms. Our neural network based approach not only outperforms traditional models but also provides insights about the importance of different risk factors.Hence by discussing how deep neural networks can enhance heart disease prediction while maintaining interpretability, our work offers a practical framework for developing support tools for clinical decision making that can potentially save lives through earlier detection.

2. Related Work

Machine learning and deep learning techniques are applied in the prediction of cardiovascular disease (CVD) and it has gained significant traction in recent time due to the increasing availability of large-scale health datasets and advancements in computational power. Traditional statistical methods, such as logistic regression, have long been employed for CVD risk assessment. For instance, Framingham Risk Score models, which rely on logistic regression, have been widely used to estimate the 10-year risk of CVD based on factors like age, cholesterol levels, blood pressure, and smoking status[1]. However, these models often oversimplify the complex interplay of risk factors, leading to moderate predictive performance, typically around 70% accuracy, as noted in various studies[2]. Supervised ML models, such as Random Forests and SVM, have demonstrated improvements over traditional statistical approaches by capturing non-linear relationships in data. A study by Weng et al. (2017) compared the performance of Random Forests, logistic regression, and gradient boosting machines using electronic health record data, achieving accuracies ranging from 67% to 72% for CVD prediction[3]. While these models outperform simpler statistical tools, their limitations in handling high-dimensional data and intricate feature interactions have paved the way for deep learning approaches. Neuron Networks are based on human neurons. Deep learning is a further advanced part of Machine Learning and it is inspired by the structure of the human brain and nervous system. Deep learning algorithms are designed in such a way that it can extract and learn complex patterns from raw input data without the explicit training[4]. Neural networks have shown promise in medical diagnostics, particularly for CVD prediction. Krittanawong et al. (2019) conducted a systematic review of deep learning applications in cardiology, highlighting that convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) consistently outperformed traditional ML models when applied to structured clinical datasets[5]. For example, a study by Dutta et al. (2020) utilized a deep neural network to predict heart disease using the UCI Heart Disease dataset, achieving an accuracy of 71.8%, surpassing Random Forest’s 68.5% on the same dataset[6]. The success of these models is often attributed to techniques such as regularization, dropout, and batch normalization, which mitigate overfitting and enhance generalization to unseen data[7]. Despite these advancements, challenges remain in applying deep learning to CVD prediction. Many prior studies have relied on smaller datasets with limited feature diversity, such as the Cleveland Heart Disease dataset, which includes only 14 attributes and 303 patient records[8]. This restricts the ability of models to generalize across diverse populations. In contrast, larger and more comprehensive datasets, such as those derived from national health registries, have enabled more robust predictions by incorporating a broader range of clinical and lifestyle factors[9]. Additionally, interpretability remains a critical concern in deep learning models, as their “black box” nature can hinder clinical adoption. Techniques like permutation importance and SHAP (SHapley Additive exPlanations) values have been proposed to address this, offering insights into feature contributions while maintaining predictive power[10]. Our study builds on this foundation by leveraging a large, balanced dataset with both modifiable and non-modifiable risk factors, addressing some of the limitations of prior work. Unlike earlier efforts that focused solely on clinical parameters, our inclusion of lifestyle factors aligns with emerging evidence that behavioral variables significantly enhance CVD risk stratification[11]. By employing a customized neural network with advanced regularization and hyperparameter tuning, we aim to push the boundaries of predictive accuracy beyond the benchmarks set by traditional ML and earlier DL models.

3. Research Methodology

In this research we have used a large dataset with nearly about 70,000 patient records. In each row we have 12 features like age, gender, blood pressure etc.

3.1 Data Preprocessing:

We have prepared the data by converting the original value into some new value that can be processed well by the neural network.

a. **Age Conversion**: Initially the age was in days for every patient. We have converted the age of patients into years for better understanding.

Age(P)(age in years) $\frac{age\in days}{365}$

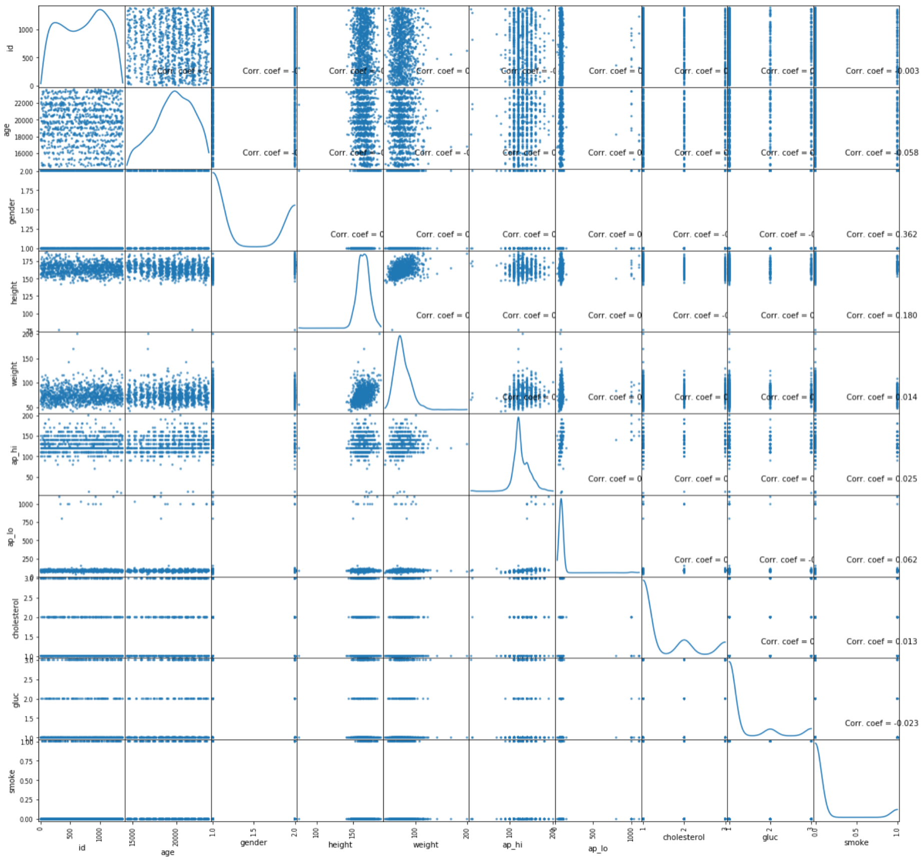

Fig.1. Scatter and Density plot of different features

b. **BMI Calculation:** BMI is a very important feature that can predict whether a person is having any cardiovascular disease or not to a certain extent. This feature was missing from the original dataset that we have downloaded. So, we calculated the feature by the following formula,

$$\text{BMI} = \frac{weight}{}$$

c. **Hypertension Flag:** Just like BMI, Hypertension is also another important variable to check the presence of heart disease. This feature was also not present in the original dataset. We have calculated this using the simple logic : *if the systolic blood pressure was* ≥ 140 *or diastolic blood pressure was* ≥ *90.*

d. **Normalization:** The features in the dataset had different scales. To get similar scales, we had used *StandardScaler*. The StandardScaler normalizes all the numerical features by first subtracting the mean and then dividing by the standard deviation. If, X = Original feature value, μ = Mean of the feature, σ = Standard Deviation of the feature, X_{scaled}= Normalized feature, then we can have the formula as follows

$$X_{scaled} = \frac{X-\mu}{\sigma}$$

e. **One-Hot Encoding:** When we have categorical variables like gender, for better data handling we convert them into numerical(binary) values using One-Hot Encoding. For every value in the feature vector we convert it into binary numbers,

```
for every k in K:
    $if category = k$:
        $k_{encoded} = 1$
    else:
        $k_{encoded} = 0$
```

3.2 Model Architecture:

In this study, we have designed a feedforward neural network(where we can observe unidirectional data flow from input to output) with a sequential architecture to classify CVD. The reason behind choosing a sequential model is it provides a stack of different layers of the network in a linear and straightforward fashion. In our model we have used an input layer that has 14 neurons, two hidden layers with batch normalization and dropout and an output layer with sigmoid activation function. Complex patterns in structured data can be captured in an efficient way using this model. We have also used regularization techniques to prevent overfitting.

We have used some Dense layers as they are also well-suited for learning complex relationships in structured data. Dense layers are nothing but fully connected layers where each neuron is connected to every other neuron in the previous layer. Every dense layer calculates a weighted sum and an activation function by which it transforms the input data. It can be represented by the below formula,

$$\Upsilon_n = \omega_n . \alpha_{n-1} + \beta_n$$

$$\alpha_n = ReLU(\Upsilon_n)$$

where, Υ_n is the linear output of the layer *n*, ω_n is the weight matrix of layer *n*, α_{n-1} is the activation output from the previous layer *n-1*, β_n is the bias vector of layer *n*. ReLU(Rectified Linear Unit) activation function is computationally efficient and it has the ability to introduce non-linearity into our deep learning model. Hence, we have used ReLU as the activation function in the Hidden Layers of our application.

3.2.1 Layer-by-Layer Architecture

3.2.1.1 Input Layer: The input layer is used here as it accepts the patient data using the 14 neurons(one for each feature). $\alpha^{[0]} = \chi$. where, $\alpha^{[0]}$ = input activations, χ = input feature matrix.

3.2.1.2 First Hidden Layer: Here we have 128 neurons with ReLU activation. It can be expressed by the formula of Υ_n above.

3.2.1.3 Batch Normalization Layer: This layer stabilizes training by normalizing inputs of layers.

$\mu^{[1]}{}_{\square} = \frac{1}{m}\sum_{i=1}^{m} \quad z_i{}^{[1]}$;

$$\Rightarrow \sigma^{2[1]} = \frac{1}{m}\sum_{i=1}^{m}$$

3.2.1.4. Dropout Layer: This layer deactivates neurons randomly and prevents overfitting.

3.2.1.5. Second Hidden Layer: This layer consists of 64 neurons with ReLU activation function.

3.2.1.6. Second Batch Normalization Layer: It is also having the same structure as the first normalization layer and the core functionality is also the same.

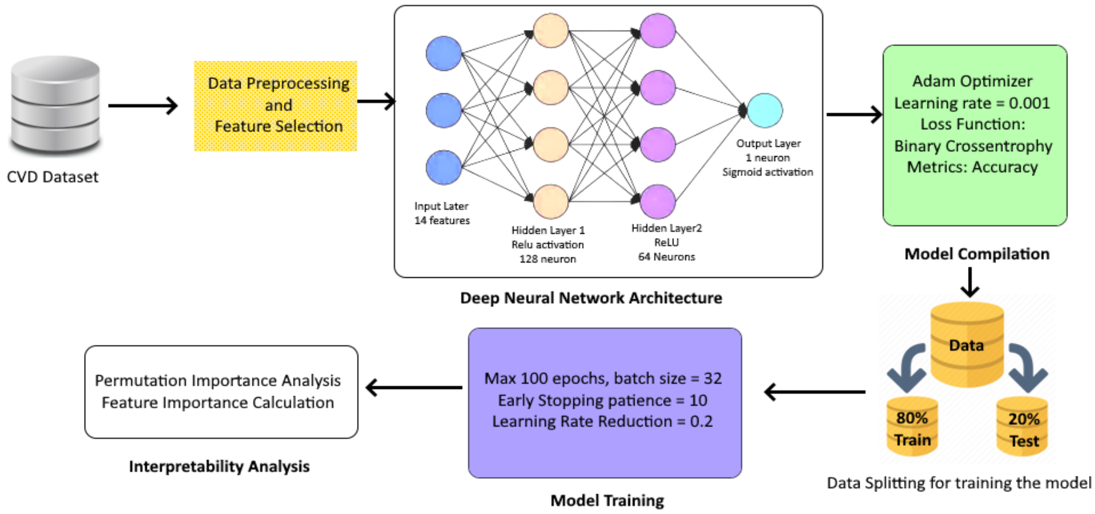
CVD Dataset
Data Preprocessing and Feature Selection
Input Later 14 features
Hidden Layer 1 Relu activation 128 neuron
Hidden Layer2 ReLU 64 Neurons
Output Layer 1 neuron Sigmoid activation
Deep Neural Network Architecture
Adam Optimizer
Learning rate = 0.001
Loss Function:
Binary Crossentrophy
Metrics: Accuracy
Model Compilation
Data
80% Train
20% Test
Data Splitting for training the model
Max 100 epochs, batch size = 32
Early Stopping patience = 10
Learning Rate Reduction = 0.2
Model Training
Permutation Importance Analysis
Feature Importance Calculation
Interpretability Analysis

Fig.2 Model Architecture of our Deep Learning Model

3.2.1.7. Second Dropout Layer: This layer is also having the same functionality and structure as the first dropout layer.

3.2.1.8. Output Layer: This layer is the main classification layer with one neuron. We need the sigmoid activation function so that we can classify the input into two classes - whether the person is having cardiovascular disease or not. It can be represented as follows

$z^{[3]} = w^{[3]}a^{[2]} + b^{[3]}$

$\Rightarrow \qquad \hat{Y} = \sigma(z^{[3]}) = \frac{1}{1+e^{-z[3]}}$

Ŷis the predicted probability of CVD, b = Bias scalar

4. Online license transfer

All authors are required to complete the Procedia exclusive license transfer agreement before the article can be published, which they can do online. This transfer agreement enables Elsevier to protect the copyrighted material for the authors, but does not relinquish the authors' proprietary rights. The copyright transfer covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microfilm or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder, the permission to reproduce any figures for which copyright exists.

Acknowledgements

Acknowledgements and Reference heading should be left justified, bold, with the first letter capitalized but have no numbers. Text below continues as normal.

An example appendix

Authors including an appendix section should do so before References section. Multiple appendices should all have headings in the style used above. They will automatically be ordered A, B, C etc.

Example of a sub-heading within an appendix

There is also the option to include a subheading within the Appendix if you wish.

References

Van der Geer, J., Hanraads, J. A. J., & Lupton, R. A. (2000). The art of writing a scientific article.*Journal of Science Communication, 163*, 51–59.

Strunk, W., Jr., & White, E. B. (1979).*The elements of style* (3rd ed.). New York: MacMillan.

Mettam, G. R., & Adams, L. B. (1999).How to prepare an electronic version of your article. In B. S. Jones & R. Z. Smith (Eds.), *Introduction to the electronic age* (pp. 281–304). New York: E-Publishing Inc.

Fachinger, J., den Exter, M., Grambow, B., Holgerson, S., Landesmann, C., Titov, M., et al. (2004).Behavior of spent HTR fuel elements in aquatic phases of repository host rock formations, 2nd International Topical Meeting on High Temperature Reactor Technology. Beijing, China, paper #B08.

Fachinger, J. (2006). Behavior of HTR fuel elements in aquatic phases of repository host rock formations.*Nuclear Engineering & Design,236*, 54.