JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Prediction of Coronary Artery Disease (CAD) using Deep Learning Models

[1]Mohammed Khaja Rehanuddin, [1]Mohammed Raihaan, [1]Mohammed Arhamuddin,

[1]Mohammad Aliya Firdous,[2] S.V.S. Hanumantha Rao, [3]Dr. Thayyaba Khatoon

[1]UG Student, [2]Assistant Professor, [3]Professor & H.O.D
[123]Artificial Intelligence and Machine Learning, School of Engineering
[123]Malla Reddy University, Maisammaguda, Hyderabad, India

Abstract : Coronary Artery Disease (CAD) poses significant global health risks and timely prediction of this cardiovascular disease is crucial for improved patient outcomes. This research aims to contribute to the field of medical sciences by implementing Deep Learning Algorithms for early-stage prediction of CAD. The study utilizes two deep learning algorithms, namely Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN), trained on the Framingham Dataset [6] . This dataset has been collecting comprehensive data on cardiovascular diseases and associated risk factors since 1948. Our findings indicate that the LSTM model achieved an accuracy of 83%, while the RNN model achieved an accuracy of 84%. These results demonstrate the effectiveness of deep learning algorithms in predicting CAD at an early stage. The application of such predictive models holds great promise for timely interventions and ultimately improving patient outcomes..

***Key Words* - Coronary Artery Disease, CAD, cardiovascular disease, deep learning algorithms, Long Short-Term Memory, LSTM, Recurrent Neural Networks, RNN, early-stage prediction, Framingham Dataset [6].**

I. INTRODUCTION

Coronary Artery Disease is a widely spread cardiovascular disease that is caused when the coronary arteries shrink. This leads to a reduction in the flow of blood to the heart. It is one of the major health problems in the world which poses a huge threat to individuals and the healthcare systems internationally. Predicting Coronary Artery Diseases early will be of huge help to society as it will increase the chances of saving a patient before the disease becomes chronic or more dangerous to an individual. Early detection of CAD will play a significant role in timely intervention and preventing any possible adverse cardiovascular events.

Since the advancements in deep learning technologies and techniques, the field of medical research and predictive analysis has improved significantly. Many advanced deep learning models such as Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN) have shown very high potential and demonstrated unrealistic capabilities and capacities to analyze, visualize and improvise the entire medical industry and the healthcare data. These deep learning models flourish remarkably in capturing and identifying patterns that are not visible to human observations, and identifying relationships and dependencies with huge amounts of data which make them suitable for predictive analysis tasks such as predicting Coronary Artery Diseases. In this process of building a deep learning model to predict coronary artery disease, we have extensively trained this model on the infamous Framingham Heart Study Dataset [6] .

By harnessing the vast amounts of information and patterns in this dataset we intend to develop a usable predictive model that can identify individuals who are prone to Coronary Artery Disease and are prone to develop this disease. The sole aim of this research is to ease the domain of healthcare provision to intervene in the early stages, implement accurate and early measures to prevent the advancement of this disease in any individual, and reduce the mortality and morbidity caused by this chronic disease.
Throughout this research, we aspire to contribute to CAD prediction to help in improving the evaluation metrics and prediction capabilities of the existing models. This is to contribute and propose an effective model for predicting CAD. The findings and deductions of this study will have the potential to add to the existing clinical practices and risk stratification so as to guide personalized treatment for individual and exclusive cases of CAD.

II. LITERATURE SURVEY

The prediction of Coronary Artery Disease (CAD) using deep learning models has gained significant attention in recent years. Several studies have explored different approaches to enhance the accuracy and efficiency of CAD detection and risk prediction. Zhou et al. (2021) proposed [1] a 3D deep learning approach for the automatic detection of CAD using coronary CT angiography. Their study demonstrated the effectiveness of deep learning models in accurately identifying CAD cases from medical imaging data.

Wu et al. (2020) conducted [2] a study focusing on predicting CAD risk based on echocardiographic data. They developed a deep-learning model that leverages the power of machine learning algorithms to analyze echocardiographic features and provide reliable risk predictions.

In a different approach, Wang et al. (2021) introduced [3] a transfer learning-based deep learning framework for the classification of CAD. They extracted features from electrocardiogram (ECG) signals and utilized transfer learning to improve the accuracy of CAD classification.

These studies highlight the potential of deep learning models in the field of CAD prediction and detection. By utilizing advanced techniques and leveraging different types of medical data, such as coronary CT angiography, echocardiographic data, and ECG signals, researchers aim to enhance the accuracy and efficiency of CAD diagnosis, leading to improved medical interventions and patient outcomes.

III. DATASET DESCRIPTION

We used the Framingham dataset (FHS / Framingham Heart Study) [6] to train and develop our model over time. The Framingham dataset [6] is a dataset which is a landmark study. This study started back in the year 1948 in Framingham, Massachusetts, United States. It is the most influential and widely acknowledged research study in cardiovascular studies internationally. The study has been made available to the research community enabling researchers and scientists to utilize and explore factors and new possibilities.

This study continues to date, and the dataset we used while training the model had a record of 4241 entries with 15 variables.
The variables are –

1) Male – This variable is for the individual's gender. 1 represents that the individual is a male, and 0 indicates that the individual is a Female.

2) Age – This variable represents the age of the individual when the feature was recorded.

3) Education - This variable indicates the highest level of Education or the level of qualification of the individual who was examined.

4) Current Smoker - This variable indicates whether an individual is a smoker or not. 1 represents active/current smoker, and 0 represents non-smoker.

5) Cigarettes Per Day – This variable indicates the number of cigarettes an individual smokes every day. This is for individuals who recognize as current smokers.

6) BP meds – This variable indicates whether an individual is on Blood Pressure regulation medication or not. 1 represents active medication, and 0 represents null.

7) Prevalent Stroke – This variable indicates if an individual has had a stroke in the past or not. 1 represents yes, and 0 indicates none.

8) Prevalent Hypertension (Hyp): This variable indicates whether the individual has prevalent hypertension or high blood pressure. A value of 1 typically represents the presence of hypertension, while a value of 0 indicates no hypertension.

9) Diabetes – This variable indicates whether the individual has diabetes or not. 1 represents a diabetic patient and 0 represents a non-diabetic patient.

10) Total Cholesterol (totChol) – This variable represents the total cholesterol level of an individual in milligrams per deciliter (mg/dL).

11)Systolic Blood Pressure (sysBP): This variable represents the systolic blood pressure of an individual in millimeters of mercury (mmHg).

12)Diastolic Blood Pressure (diaBP): This variable represents the diastolic blood pressure of the individual in millimeters of mercury (mmHg). It indicates the pressure exerted on the artery walls when the heart is at rest between contractions.

13)Body Mass Index (BMI): This variable represents the body mass index of the individual, calculated as the weight in kilograms divided by the square of the height in meters. It provides information about the individual's body composition and weight status.

14)Heart Rate: This variable represents the heart rate of the individual in beats per minute (bpm). It indicates the number of times the heart beats in one minute and provides information about the individual's cardiovascular health.

15)Glucose: This variable represents the glucose level of the individual in milligrams per deciliter (mg/dL). It indicates the blood sugar level and provides information about the individual's glucose metabolism.

16)Ten-Year CHD (Coronary Heart Disease) Risk: This variable represents the risk of developing coronary heart disease within ten years. It is typically calculated using various risk factors and predictive models.

IV. Our Approach

In this experiment, we used deep learning models to find coronary artery disease. Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) are the two models we used.

The first thing we did was import the required libraries, which included pandas, NumPy, sci-kit-learn, and Keras. We also loaded the Framingham dataset [6], which has several cardiovascular health-related variables. We then carried out some preprocessing operations. To ensure a clean dataset, we removed rows with missing values. The target variable was then isolated from the characteristics. The goal variable ("TenYearCHD") was recorded in the variable "y," whereas the features were kept in the variable "X." We used the MinMaxScaler to normalize the features in order to get the data ready for the LSTM model. This process is essential for ensuring that all features have a comparable scale and range. Deep learning models perform better when data is normalized. Using the train_test_split function from scikit-learn, we divided the data into training and testing sets after normalizing the features. 20% of the data was set aside for testing, and the remaining 80% was used to train the models. To meet the LSTM input form requirements, we rearranged the training and testing data for the LSTM model. Samples, timesteps, and features are the types of input that LSTM models require. In this instance, the data was reshaped to include one timestep and the quantity of characteristics.

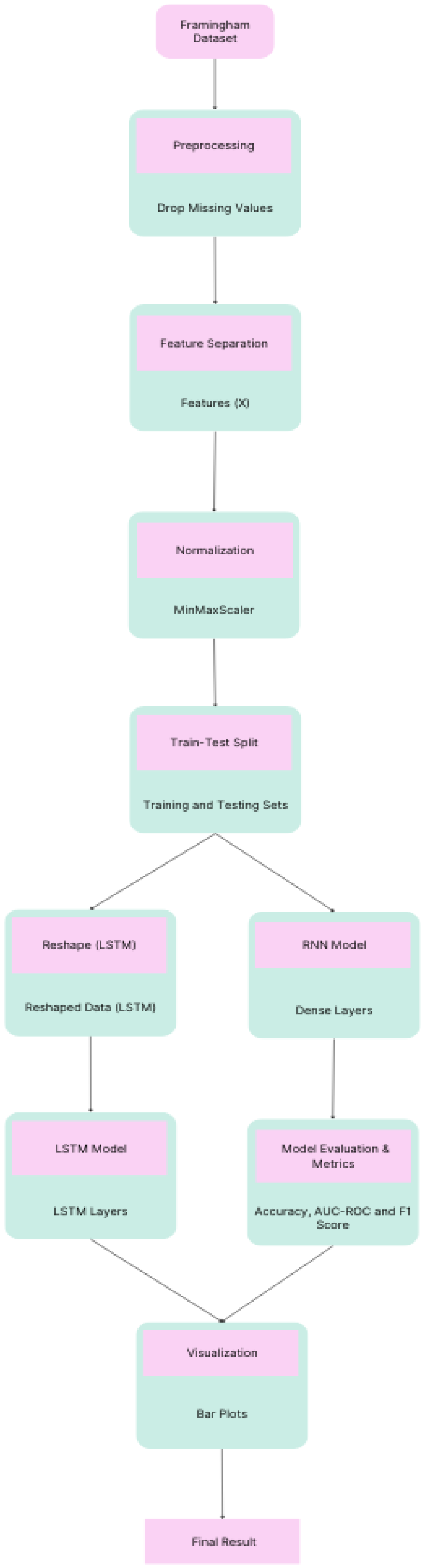
Two LSTM layers with 64 units each made up the LSTM model architecture, which was followed by dropout layers to avoid overfitting. To generate the binary classification output, a final Dense layer with a sigmoid activation function was added. Binary cross-entropy was used as the loss function and the Adam optimizer was used to build the model. The training data was then used to train the LSTM model. With verbose set to 1, we ran 50 epochs with a batch size of 32 to show the training progress. We made predictions on the test set after training the LSTM model. We rounded the model's projected probability to provide binary predictions. For the LSTM model, we computed a number of evaluation metrics, including accuracy, area under the ROC curve (AUC-ROC), and F1 score. These metrics give insight into the model's effectiveness.

Similarly, we used Keras' Sequential API to create an RNN model. Two Dense layers with 64 units each made up the RNN model, which was followed by a final Dense layer with a sigmoid activation function. Using the same inputs as the LSTM model, we created and trained the RNN model.

We determined the RNN model's evaluation metrics, such as accuracy, AUC-ROC, and F1 score. We printed the accuracy, AUC-ROC, and F1 scores for each model to evaluate the effectiveness of the LSTM and RNN models. To compare the metrics graphically, we also made bar graphs. The accuracy, AUC-ROC, and F1 score for the two models were displayed side by side in the bar graphs.

In conclusion, this effort addressed the detection of coronary artery disease using LSTM and RNN models. While the RNN model has its own set of measures, the LSTM model was more accurate, had a higher AUC-ROC, and scored higher on the F1 test. The performance comparison between the two models was represented visually by the bar plots.

V. ARCHITECTURE



VI. SIMULATION AND RESULTS

In this section, we present the simulation and results of our research paper. This section will focus on the results we acquired while predicting Coronary Artery Disease (CAD) using deep learning models.

We used the Framingham Dataset [6] which comprises clinical and demographical features of patients including various factors of high importance for us to train our model.

6.1 MODEL TRAINING AND EVALUATION

In this model, we divided the dataset into training and test sets with a 70:30 ratio, ensuring the model's ability to generalize and understand the dataset. We implemented two significantly advanced algorithms Long Short-Term Memory and Recurrent Neural Networks in this model.

We chose to use RNN and LSTM because of their undisputable ability to handle sequential data and capture temporal dependencies.

6.2 PERFORMANCE METRICS

We evaluated the performance of LSTM and RNN models on the testing set using performance metrics including AUC-ROC, F1 Score, and accuracy. Accuracy measures the proportion of correctly predicted CAD. AUC-ROC provides an overall assessment of the model's ability to discriminate between CAD-positive and CAD-negative instances. The F1 score is a measure of the model's precision and recall, accounting for both false positives and false negatives.

6.3 RESULTS

The results of our simulations are as follows:

RNN Model:

- Accuracy: 0.8456 (84%)
- AUC-ROC: 0.7212
- F1 Score: 0.1594

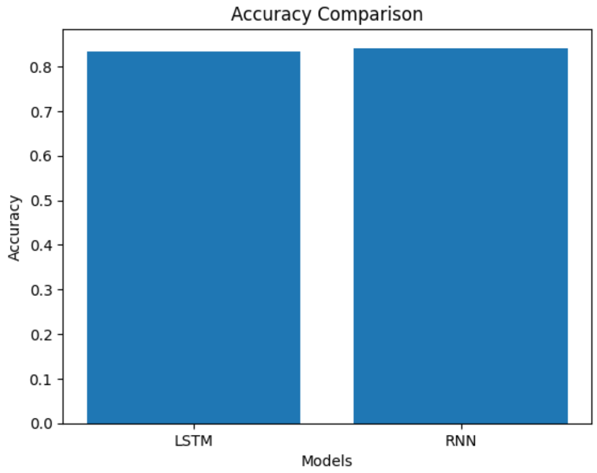
LSTM Model:

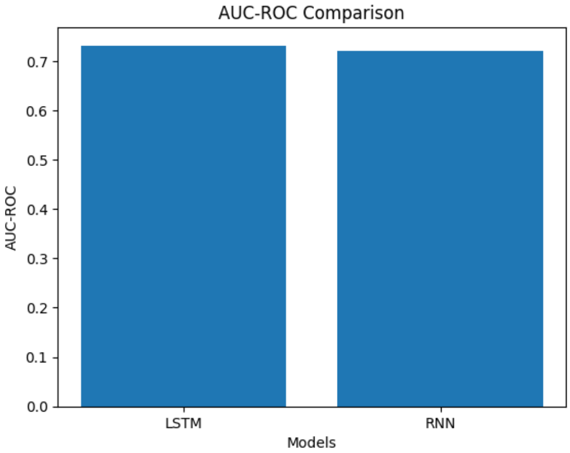
- Accuracy: 0.8360 (83%)
- AUC-ROC: 0.7315
- F1 Score: 0.0163

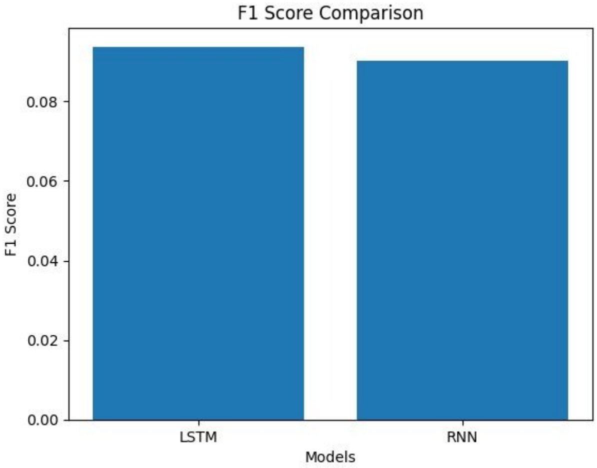
The RNN model achieved an accuracy of 0.8456, indicating that it correctly classified 84% of the CAD cases in the testing set. The AUC-ROC value of 0.7212 suggests a moderate discriminatory power of the RNN model in distinguishing between CAD-positive and CAD-negative instances. The F1 score of 0.1594 indicates a trade-off between precision and recall for the RNN model.

The LSTM model achieved an accuracy of 0.8360, correctly identifying 83% of the CAD cases. The AUC-ROC value of 0.7315 suggests a slightly improved discriminatory power compared to the RNN model. However, the F1 score of 0.0163 indicates challenges in achieving a balance between precision and recall for the LSTM model.

Both LSTM and RNN showed promising performance in predicting CAD. The LSTM model showed a slightly improved AUC-ROC score whereas deferred in accuracy by a slight margin.







VII. CONCLUSIONS

In conclusion the usage of LSTM and RNN models in predicting CAD offers several advantages. These deep learning models are suitable for the process of predicting CAD. The ability of these deep learning models especially to process variable length inputs and retain long-term memory makes them tailor fit for analyzing and understanding the huge dataset and getting valuable insights from this invaluable dataset.

Our study's findings demonstrated that both the RNN and LSTM models performed effectively in predicting CAD.

The RNN model had an AUC-ROC score of 0.7212 and an accuracy of 84%. The LSTM model achieved an AUC-ROC value of 0.7315 and an accuracy of 83%. These findings show that the models can distinguish between CAD-positive and CAD-negative situations. It is significant to remember that both models encountered difficulties in obtaining a balanced F1 score. The LSTM model had an F1 score of 0.0163, compared to the RNN model's 0.1594. This shows that additional optimization and threshold adjustments for categorization are required to balance precision and recall.

Our study demonstrates the potential of deep learning models, particularly RNN and LSTM, in CAD prediction. The models offered useful insights for identifying those at risk of getting CAD and showed strong accuracy and discriminatory power. These results can help medical practitioners with CAD burden reduction through early detection, risk assessment, and targeted therapies.

To increase the overall performance and reliability of CAD prediction using deep learning methodologies, more research is required to improve the models, add more features, and investigate different evaluation measures. We can improve the therapeutic utility of these models and help with CAD management and prevention by addressing these issues.

VIII. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Professor S.V.S. Hanumantha Rao for his invaluable guidance, expertise, and continuous support throughout the course of this research. His profound knowledge and insightful inputs were instrumental in shaping the direction and quality of our study.

We would also like to extend our heartfelt appreciation to Dr. Thayyaba Khatoon, the Head of the Department of Artificial Intelligence and Machine Learning at Malla Reddy University. Her encouragement, mentorship, and valuable feedback significantly contributed to the success of this research endeavor.

Furthermore, we would like to thank the faculty members of the Department of Artificial Intelligence and Machine Learning for their constant encouragement, assistance, and constructive discussions that enriched our understanding and strengthened our research. We are indebted to the participants who generously shared their data for this study, as their contribution was crucial in conducting meaningful analysis and generating reliable results. Lastly, we extend our appreciation to all our family members and friends who provided unwavering support, understanding, and encouragement throughout this journey.

We acknowledge and appreciate the contributions of all individuals who played a part, directly or indirectly, in the completion of this research paper.

REFERENCES

[1] Zhou, Y., Cai, H., & Huang, S. (2021). Automatic detection of coronary artery disease using 3D deep learning on coronary CT angiography. Computers in Biology and Medicine, 135, 104589.

[2] Wu, G., Zhou, Y., Zhang, S., Zhao, Y., Chen, W., & Wang, C. (2020). Deep learning-based echocardiographic risk prediction for coronary artery disease. Frontiers in Physiology, 11, 1057.

[3] Wang, Z., Li, H., Zhang, Y., Xu, Z., & Zhu, X. (2021). Classification of coronary artery disease based on transfer learning and deep learning. Biomedical Signal Processing and Control, 68, 102617.

[4] "Long Short-Term Memory" by Sepp Hochreiter and Jürgen Schmidhuber - https://www.bioinf.jku.at/publications/older/2604.pdf

[5] "Recurrent Neural Networks" by Ilya Sutskever, Oriol Vinyals, and Quoc V. Le - https://arxiv.org/pdf/1409.3215.pdf

[6] "Framingham Heart Study" - Original Study: https://www.framinghamheartstudy.org/