Stochastic Modelling and Computational Sciences

INTEGRATING MACHINE LEARNING AND DEEP LEARNING FOR HEART DISEASE PREDICTION

Vilas Ramrao Joshi[1], Kailash Nath Tripathi[2], Ashima Jain[3], Twinkle[4], Ayush Sharma[5] and Anita Kumari[6]

[1]Associate Professor, Department of Computer Engineering, Isbm College of Engineering, Pune, Maharastra, India

[2]Assistant Professor, Department of Computer Engineering, Isbm College of Engineering, Pune, Maharashtra, India

[3]Research Scholar, Department of Computer Science & Engineering, Shri Venkateshwara University, Gajraula, Up, India

[4]Research Scholar, Department of Computer Science, Shri Venkateshwara University, Gajraula, Up, India

[5]Research Scholar, Department of Computer Science & Enggineering, Shri Venkateshwara University, Gajraula, Up, India

[6]Assistant Professor, Alard Institute of Management Sciences, Pune, Maharastra, India

[1]joshivilas131071@gmail.com, [2]kailash.tripathi@gmail.com and [6]anita.tripathi15@gmail.com

ABSTRACT

Cardiovascular diseases (CVDs) persist as a foremost contributor to global morbidity and mortality rates. Timely and precise prediction of CVD risk plays a pivotal role in instituting preventive measures and advancing patient outcomes. This study introduces an innovative approach amalgamating the strengths of machine learning (ML) and deep learning (DL) models to augment cardiovascular disease prediction. By delving into the realm of machine learning algorithms, particularly Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest, alongside DL models like Long Short-Term Memory (LSTM), Neural Network, and Convolutional Neural Network (CNN), we aim to discern the most efficacious model for early heart condition prognosis. Furthermore, we propose an ensemble classifier, synergizing the diverse capabilities of various models to enhance prediction accuracy. Our ensemble model showcases promising outcomes, presenting a robust and nuanced strat egy for heart disease prediction. This research furnishes valuable insights for healthcare practitioners, accentuating the pivotal role of classifier selection and highlighting the potential of ensemble models in refining predictive accuracy for cardiovascular health.

1. INTRODUCTION

The heart, a vital organ in the human body, is particularly susceptible to cardiovascular diseases (CVD), encompassing conditions such as coronary heart disease, heart attacks, strokes, and heart failure. CVDs account for a staggering number of deaths globally, with an estimated 17.9 million fatalities annually. In the United Kingdom, mortality rates due to CVD have notably risen, especially in individuals over the age of 50 [1]. Various medical conditions, including diabetes and hypertension, often manifest when the heart struggles to efficiently circulate blood throughout the body. Electronic Health Records (EHRs) play a pivotal role in managing patient information, ensuring its accessibility, accuracy, and patient-centeredness. EHRs facilitate the discovery of hidden insights within patient data, aiding both clinical decision-making and research endeavors, thereby modernizing healthcare practices and minimizing reliance on traditional methods.

Risk factors for heart disease encompass a wide range of variables, including age, sex, smoking habits, family history, cholesterol levels, diet, physical inactivity, and alcohol consumption. While some risk factors are beyond individual control, such as genetics, many lifestyle habits, like dietary patterns, physical activity, and obesity, significantly influence heart health. Unhealthy lifestyle choices like tobacco use, poor diet, physical inactivity, and excessive alcohol consumption are primary contributors to heart ailments [2]. Researchers employ various data mining techniques to analyze heart diseases, with obesity and overweight being notable risk factors alongside traditional variables like age and sex. Machine Learning approaches, including decision trees, logistic regression, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), play a crucial role in predicting heart failure, aiming to enhance prediction accuracy [4][6].

Efforts to prevent premature deaths due to CVDs hinge on early identification of individuals at risk and ensuring they receive appropriate medical intervention and lifestyle modifications. Primary care settings play a vital role in this endeavor, necessitating access to essential medications and health education. Despite widespread challenges in recognizing and managing risk factors, lifestyle modifications and medication adherence can significantly mitigate the incidence of heart attacks and strokes [5]. The prediction of heart diseases through Machine Learning techniques offers promising avenues for early intervention and prevention. By leveraging advanced algorithms and comprehensive datasets, researchers strive to improve the accuracy of heart disease prediction, ultimately aiming to reduce the global burden of CVD-related morbidity and mortality. We review existing literature on heart disease prediction, highlighting the diverse array of algorithms and methodologies utilized. Additionally, we discuss the limitations of current approaches and the potential benefits of integrating machine learning and deep learning techniques. Through empirical analysis and experimentation on diverse datasets, we demonstrate the efficacy of our proposed framework in achieving superior predictive performance compared to traditional methods. Our findings underscore the potential of integrating machine learning and deep learning for advancing heart disease prediction, thereby enabling proactive healthcare interventions and improving patient outcomes.

2. LITERATURE SURVEY

In contemporary healthcare research, the integration of machine learning, deep learning, and data mining methodologies has become increasingly prominent for disease prediction. Each study contributes distinct insights and varying levels of predictive accuracy based on their respective methodologies. One study proposed a hybrid approach that combines Support Vector Machine (SVM) and Genetic Algorithm (GA), achieving notable results by leveraging data mining tools such as LIBSVM and WEKA across five diverse datasets from the IUC repository. This hybrid model yielded accuracies of 84.07% for heart disease, 78.26% for diabetes, 76.20% for breast cancer, and 86.12% for Hepatitis [1]. Another investigation advocated for data mining techniques in heart disease detection, utilizing algorithms like J48, Naïve Bayes, and bagging through the WEKA tool, achieving accuracies of 82.31% with Naïve Bayes, 84.35% with J48, and 85.35% with Bagging for heart disease classification [2].

Highlighting the efficacy of the Naïve Bayes algorithm, known for its independence assumption, a separate study analyzed a dataset containing 500 patients, achieving an accuracy of 86.419% using the WEKA tool for classification [3]. An exhaustive review of existing works on heart disease prediction emphasized the prevalence of data mining techniques, stressing the significance of combining multiple algorithms to enhance prediction accuracy, as opposed to relying solely on individual algorithms [4]. Another study evaluated a sequential feature selection approach alongside a neuro-fuzzy classifier, achieving an accuracy of 88.2% on the Cleveland dataset by evenly splitting the dataset for training and testing [5].

Exploring ten methods utilizing the heart disease dataset from the UCI repository, another study found Partial Least Square Discriminant Analysis (PLS-DA) to exhibit an accuracy of 86.13% [6]. A proposal to augment heart disease prediction through data techniques demonstrated superior accuracy (85%), particularly in parallel fashion compared to sequential SVM [7]. Further investigation into various data mining approaches for predicting heart disease yielded accuracies of 84% with Neural Network and 89% with Hybrid Systems using WEKA and MATLAB [8]. Advocacy for heart disease prediction and analysis using J48, Naïve Bayes, and Support Vector Machine techniques underscored their potential to enhance service quality and reduce costs [9-10]. Review of machine learning-based heart disease prediction approaches is presented in Table 1.

Table 1: *Review of literature*

Paper Reference	Methods /Algorithms	Findings
[6]	SVM, random forest, logistic models (WEKA), Arduino	SVM exhibited superior accuracy in cardiac monitoring system design for in-home environment.

[7]	KNN, SVM, logistic regression, random forest	SVM showed highest accuracy; suggested further research to refine ML algorithms for clinical use.
[8]	Decision trees, SVM, naive Bayes, logistic regression, random forest, QDA	SVM achieved 95% accuracy in heart disease prediction; noted exclusion of crucial elements.
[9]	Feature selection, logistic regression	Logistic regression had highest accuracy in predicting coronary illness using Cleveland dataset.
[10]	Logistic regression	Logistic regression achieved highest accuracy using 116 records and 34 variables.
[11]	Naive Bayes	Naive Bayes detected heart disease with 86.419% accuracy despite limitations in dataset size.
[12]	ID3	Proposed concealed approach using ID3 algorithm for hidden patterns identification in heart disease.
[13]	MAFIA, K-Means clustering	Emphasized accurate classification's importance for disease prediction using data mining techniques.
[14]	SMO, Bayes Net	SMO and Bayes Net showed optimal performance in heart disease prediction using two datasets.
[15]	In-built imputation algorithm and particle swarm optimization	Physical inactivity
[16]	Automatic classifier (for risk assessment in congestive heart failure)	Long-term heart rate variability
[17]	Decision tree algorithm	Events before and after CHD (e.g., PCI, MI, CABG)
[18]	Association rules with search constraints	Relevant association rules for heart disease prediction
[19]	Hybrid system with genetic algorithm (for neural network weight initialization)	Multilayered feed-forward network initialization
[20]	Decision trees and Apriori algorithm	Chest pain, diabetes, smoking, gender, physical inactivity, age, lipids, cholesterol, triglyceride, blood pressure

3. RESEARCH METHODOLOGY

The section discusses the implementation of a hybrid machine learning approach for heart disease prediction, focusing on workflow and methodology. It utilizes heart disease data sourced from a reputable website to predict severe cardiac syndromes in critically ill patients. The dataset consists of six attributes and 100 records collected from the Enam Medical Diagnosis Centre in Bangladesh, which undergo preprocessing to ensure data integrity.

Feature selection and modeling iterations involve various machine learning techniques such as Decision Trees, Logistic Regression, Naive Bayes, K Nearest Neighbor (KNN), and Support Vector Machine (SVM). Each algorithm brings unique strengths to the prediction task, with Decision Trees enabling rapid model generation and easy interpretation, Logistic Regression leveraging probability theory, and SVM excelling in classification tasks. The hybrid system's performance is evaluated based on accuracy and sensitivity metrics, with a dataset divided into training and test sets. The model undergoes thorough validation using techniques like k-fold cross-validation and fine-tuning of hyperparameters to optimize performance. Insights from the hybrid model aid healthcare professionals in clinical decision-making and patient management. The finalized model is deployed in clinical settings, prioritizing transparency, reliability, and adherence to ethical guidelines. Continuous monitoring and updating ensure adaptability to changing trends in heart disease dynamics.

4. METHODS FOR HEART DISEASES PREDICTIOM

Heart disease remains a leading cause of mortality worldwide, necessitating effective predictive tools for early detection and intervention. This paper presents an integrated approach leveraging machine learning and deep learning techniques for heart disease prediction. By synthesizing methodologies from both domains, our proposed framework aims to enhance prediction accuracy and facilitate comprehensive risk assessment.

4.1.1 Traditional Methods

Traditional risk assessment models for predicting heart diseases are pivotal tools in clinical practice, aiding healthcare professionals in evaluating the likelihood of cardiovascular events over a specified period. These models integrate various demographic, clinical, and lifestyle factors to provide an estimate of an individual's risk profile.

4.1.2 Framingham Risk Score (FRS)

Framingham Risk Score (FRS) stands as one of the most widely utilized models, originating from the landmark Framingham Heart Study. FRS estimates the 10-year risk of coronary heart disease (CHD) based on parameters such as age, gender, cholesterol levels, blood pressure, smoking habits, and diabetes status. By assigning points to these factors, FRS categorizes individuals into low, intermediate, or high-risk groups.

4.1.3 SCORE (Systematic Coronary Risk Evaluation)

SCORE (Systematic Coronary Risk Evaluation), endorsed by the European Society of Cardiology, is another prominent model. It forecasts the 10-year risk of fatal cardiovascular events, incorporating variables like age, gender, smoking status, systolic blood pressure, and total cholesterol. Different versions of SCORE cater to regions with varying risk profiles.

4.1.4 Reynolds Risk Score

Reynolds Risk Score augments traditional risk assessment by integrating additional factors such as family history of heart disease and high-sensitivity C-reactive protein (hsCRP) levels. Particularly beneficial for women, this model enhances risk prediction accuracy.

4.1.5 QRISK

QRISK, developed in the UK, embraces a broader spectrum of risk factors including BMI, ethnicity, social deprivation, and comorbidities like rheumatoid arthritis and chronic kidney disease. By tailoring risk assessment to individual characteristics, QRISK offers a more personalized approach.

Other notable models include PROCAM (Prospective Cardiovascular Münster) Score, Diamond-Forrester Method, HeartScore, Revised Cardiac Risk Index (RCRI), and the Cleveland Clinic Score. Each model exhibits its unique strengths, catering to specific patient populations or clinical scenarios.

4.2 MACHINE LEARNING METHODS

4.2.1 Linear Regression

Linear Regression is a fundamental machine learning method employed for predicting continuous variables by establishing linear relationships between features and the target variable. Widely utilized across diverse domains, Linear Regression finds application in predicting house prices, stock prices, sales forecasts, and various other scenarios where understanding the relationship between variables is crucial for making predictions. Its simplicity, interpretability, and effectiveness make it a cornerstone in predictive modeling tasks, providing valuable insights into the underlying patterns and trends within data. Various machine learning based methods are shown in figure 1.

4.2.2 Logistic Regression

Logistic Regression serves as a versatile tool primarily used for binary classification tasks, distinguishing whether an event will occur or not. Widely applicable in domains such as churn prediction, customer segmentation, and disease diagnosis, Logistic Regression leverages probability estimates to make predictions, providing valuable insights into categorical outcomes.

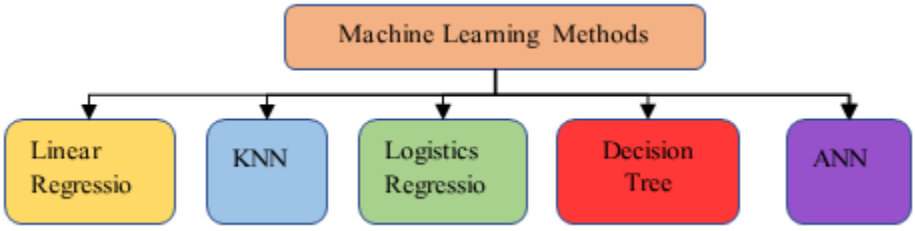

Figure 1: *Machine learning based heart diseases prediction methods*

4.2.3 Decision Trees

Decision Trees offer a non-linear approach to prediction by partitioning the feature space into hierarchical structures of binary decisions. These models excel in both classification and regression tasks, with popular algorithms like CART and Random Forests providing robust predictive capabilities.

4.2.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) enhance predictive accuracy, particularly in classification tasks involving high-dimensional data or non-linear decision boundaries. By identifying the hyperplane that best separates different classes in the feature space, SVM effectively discerns patterns and makes predictions with high accuracy, making it indispensable in various domains.

4.2.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) represents an instance-based learning algorithm that predicts the value of a new data point based on the majority class or average value of its k nearest neighbors. This method finds utility in both classification and regression tasks, especially when dealing with small to medium-sized datasets. Additionally, time series forecasting methods like ARIMA, SARIMA, and Exponential Smoothing are effective for predicting future values in time-series data, applicable in domains such as stock market forecasting and demand prediction. Finally, ensemble methods, clustering algorithms, and other specialized techniques contribute to the rich landscape of machine learning methods for prediction, each offering unique strengths suited to specific prediction tasks and datasets.

4.3 Deep Learning Methods

Neural Networks, a hallmark of deep learning, comprise interconnected layers of artificial neurons capable of learning complex patterns and relationships in data. Widely used in applications such as image recognition, natural language processing, and time-series prediction, neural networks demonstrate exceptional predictive

power. Gradient Boosting Machines (GBM), another ensemble learning technique, sequentially build multiple weak learners, each correcting the errors of its predecessor. Boosting algorithms like XGBoost and LightGBM are particularly effective for regression and classification tasks with structured data.

Deep learning-based methods have emerged as powerful tools for predicting heart diseases, leveraging sophisticated neural network architectures to analyze diverse data types and provide accurate risk assessments. Convolutional Neural Networks (CNNs) are particularly effective for image-based diagnoses in heart disease prediction. By analyzing medical imaging data such as X-rays or MRIs, CNNs can detect abnormalities indicative of conditions like coronary artery disease or heart failure. Their ability to extract hierarchical features from images enables precise localization and classification of pathological conditions, enhancing diagnostic accuracy.

In addition to CNNs, Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, play a crucial role in heart disease prediction, especially when dealing with sequential data. RNNs excel at processing time-series data such as electrocardiograms (ECG), enabling the detection of irregularities or the prediction of future cardiac events. By capturing temporal dependencies in sequential data, RNNs provide valuable insights into disease progression and risk assessment.

Hybrid architectures that combine CNNs and RNNs have emerged as a promising approach to heart disease prediction, integrating information from multiple data modalities for comprehensive risk assessment. By leveraging the strengths of both CNNs and RNNs, these hybrid models offer a holistic view of a patient's health status, incorporating medical imaging, clinical records, and other relevant data sources. This integration of diverse data modalities enhances predictive accuracy and enables early detection of heart diseases, empowering clinicians to make timely and informed decisions for patient care.

5. PROPOSED MODEL

The initial phase of the system involves gathering data and identifying the most essential attributes for analysis. Subsequently, the collected data undergoes preprocessing to ensure it conforms to the required format. Following preprocessing, the data is partitioned into separate sets for training and testing purposes. Machine learning algorithms are then applied, utilizing the training data to train the model. The system's accuracy and effectiveness are assessed by testing it with the separate test data. Below are the modules utilized to execute this system (Figure 2).

Predicting heart diseases using machine and deep learning entails a comprehensive process, starting with the collection and preprocessing of relevant data. For instance, consider a dataset comprising demographic details, medical history, and clinical measurements such as blood pressure and cholesterol levels. This data undergoes rigorous preprocessing to handle missing values, outliers, and formatting inconsistencies, ensuring its suitability for analysis. Feature selection and engineering follow, where pertinent attributes like age, gender, smoking status, and various physiological markers are identified and transformed to capture significant patterns related to heart disease risk.

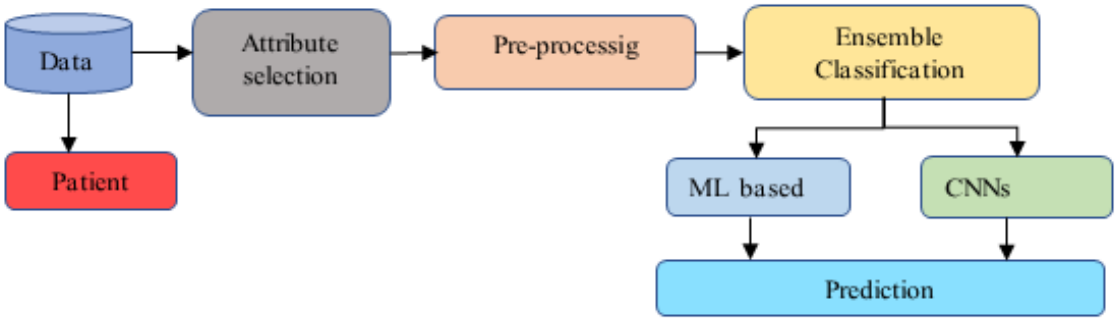

Figure 2: *Proposed research methodology*

Once the dataset is prepared, the selection of appropriate models becomes crucial. Traditional machine learning algorithms like logistic regression or decision trees may suffice for simpler prediction tasks. Conversely, complex prediction scenarios involving vast datasets may necessitate the use of deep learning models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). For example, CNNs can analyze medical images like X-rays or MRI scans to detect cardiac abnormalities, while RNNs are adept at processing sequential data like electrocardiograms (ECG) for identifying irregular heart rhythms.

Training the selected models involves feeding them with the prepared dataset to learn the underlying patterns and relationships between input features and the target variable – the presence or absence of heart disease. After training, the models' performance is evaluated using separate validation or test datasets. Metrics like accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) are used to assess the models' predictive capabilities. For instance, an AUC-ROC value close to 1 indicates a highly accurate model, while values closer to 0.5 suggest random guessing.

Once a satisfactory model is identified and validated, it can be deployed in clinical settings to assist healthcare professionals in predicting heart diseases. Continuous monitoring and evaluation are essential to ensure the model's accuracy and reliability over time. For instance, regular updates may be necessary to accommodate changes in data distribution, patient populations, or clinical guidelines. Overall, the systematic approach to data collection, preprocessing, model selection, training, evaluation, and deployment underscores the importance of machine and deep learning in improving heart disease prediction and ultimately enhancing patient outcomes.

6. DATASET DESCRIPTION

The data utilized in this project is sourced from the Cleveland Heart Disease database. It comprises a total of 297 records, each containing 14 medical attributes [7], which are employed for the prediction of heart disease. This table provides an overview of the various attributes present in the dataset, including demographic information, medical measurements, and diagnostic indicators, all of which are utilized for predicting the presence or absence of heart disease. A detailed description of the dataset is provided in Table 2 below:

Table 2. *Description of Heart Diseases prediction dataset*

S. No	Attribute	Description
1	Age	Age of the individual in years
2	Sex	Gender of the individual (0 = female, 1 = male)
3	Chest Pain Type (CP)	Type of chest pain experienced
4	Resting Blood Pressure (RBP)	Resting blood pressure measurement in mm Hg
5	Serum Cholesterol (Chol)	Serum cholesterol measurement in mg/dl
6	Fasting Blood Sugar (FBS)	Fasting blood sugar level > 120 mg/dl (1 = true, 0 = false)
7	Resting Electrocardiographic Results (Restecg)	Resting electrocardiographic measurement
8	Thalach	Maximum heart rate achieved during stress test
9	Exercise Induced Angina (Exang)	Exercise-induced angina (1 = yes, 0 = no)
10	ST Depression (Oldpeak)	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Number of Major Vessels (Ca)	Number of major vessels colored by fluoroscopy
13	Target (Num)	Presence of heart disease (0 = no, 1 = yes)
14	Thal	3= normal, 6= fixed defect, 7= reversible effect

The distribution of data plays a crucial role in predictive or classification tasks. In our dataset, heart disease occurred 55.46% of the time, whereas no heart disease was observed 44.64% of the time. Balancing the dataset is

essential to prevent overfitting, ensuring that the model does not exhibit bias towards the majority class. This approach enables the model to discern patterns contributing to both heart disease and non-heart disease cases, as depicted in Figure 3.

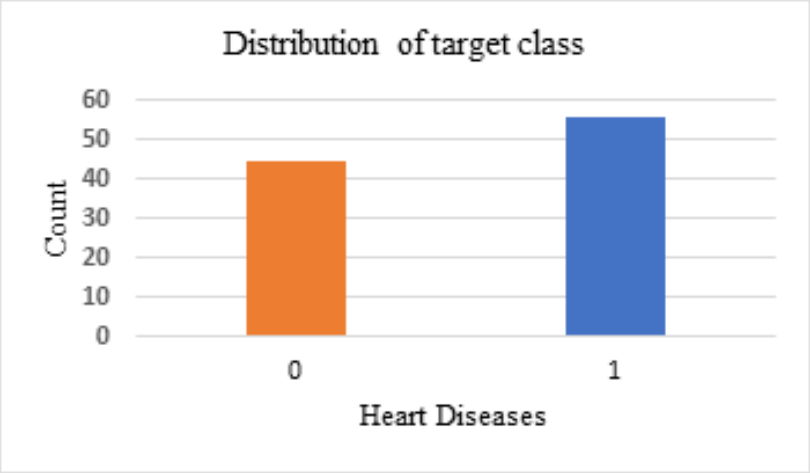

Figure 3: Distribution of target class

Analyzing attribute distributions is essential for gaining insights into the skewness of data. By visualizing distribution plots, significant relationships and trends become apparent, allowing for a deeper understanding of the dataset's dynamics. For example, these plots may reveal correlations between age and sex, associations between chest pain and resting blood pressure, and the influence of various factors on cardiovascular health. Delving into these patterns provides nuanced insights into the distribution of the dataset, enabling informed conclusions to be drawn for the development of robust predictive models and the assessment of cardiovascular disease risk (Figure 4).

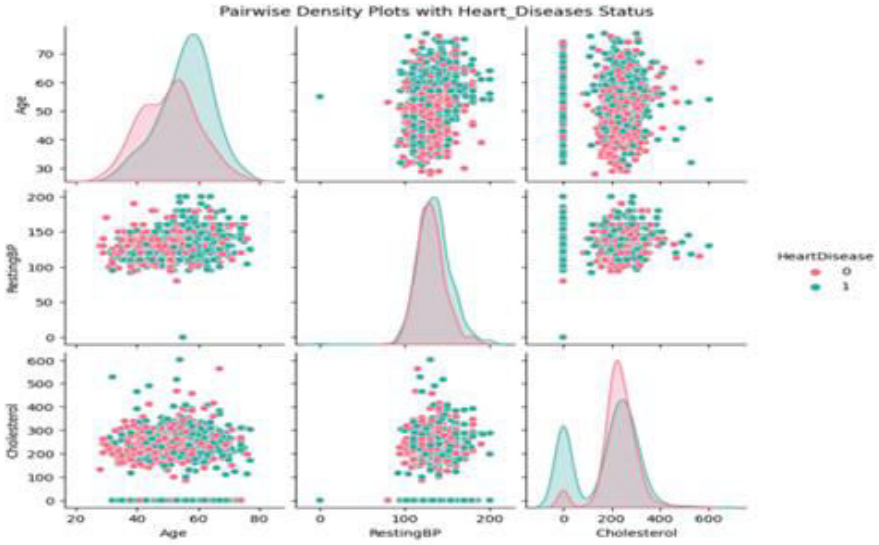

Figure 4: Density plots depicting Heart Disease Status

Understanding attribute distributions not only sheds light on data skewness but also facilitates the identification of key predictors and their interactions. By scrutinizing these relationships, researchers can discern subtle patterns

and uncover potential risk factors for cardiovascular diseases (Figure 5). This comprehensive analysis lays the foundation for effective predictive modeling, guiding the selection of relevant features and informing the development of algorithms capable of accurately assessing an individual's cardiovascular health status and predicting their risk of developing heart-related conditions.

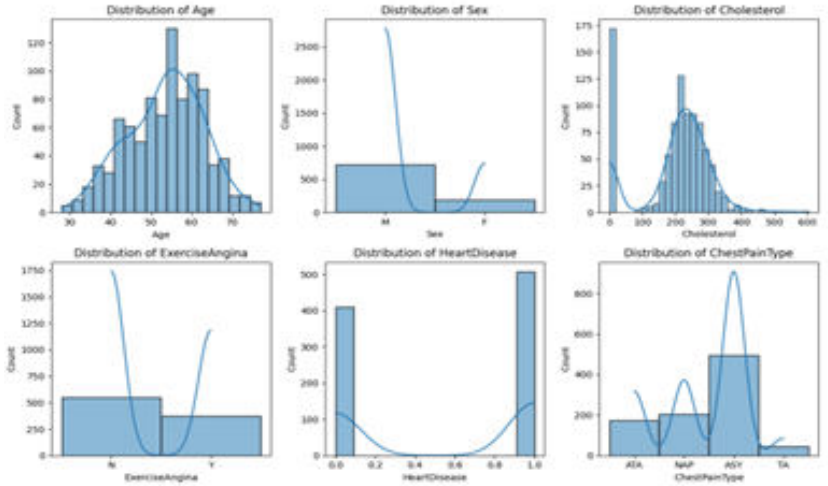

Figure 5: Assessment of the Data's Skewness

Density plots provide valuable insights into the distribution of continuous variables, especially concerning different categories such as heart disease status. For instance, a density plot depicting heart disease status can showcase the distribution of a specific variable (such as age or cholesterol levels) among individuals with and without heart disease. By juxtaposing the density distributions of these variables between the two groups, discernible patterns and disparities emerge, offering valuable insights into potential associations with cardiovascular outcomes. This graphical representation serves as a visual aid in comprehending the potential influence of a specific variable on the likelihood of experiencing heart diseases.

7. RESULTS ANALYSIS

Understanding and analyzing unprocessed cardiac healthcare data holds immense significance for ensuring the long-term preservation of health records and early detection of issues related to heart illnesses. By leveraging machine learning techniques to interpret raw data, a novel perspective on cardiac disease assessment is attained. The evaluation of heart disease presents a critical challenge in medicine; however, early identification of illnesses and implementation of preventive measures can substantially reduce mortality rates. To further enhance the effectiveness of such investigations, it is imperative to extend research efforts beyond theoretical techniques and simulations to focus on real-world datasets. A recommended approach in addressing this challenge is the utilization of hybrid machine learning techniques, particularly combining features from Support Vector Machines (SVM) and Decision Trees (DT). This hybrid methodology has demonstrated remarkable accuracy in predicting cardiac disease. The research advocates for a hybrid method that integrates machine learning techniques to identify significant features, thereby improving the accuracy of cardiovascular disease prediction. By incorporating well-established clustering techniques and various feature combinations into the prediction model, enhanced results are achieved in terms of accuracy and sensitivity. This underscores the potential of hybrid machine learning approaches in advancing cardiac healthcare analytics and facilitating more effective disease prediction and prevention strategies.

7.1 ACCURACY

The results shown in figure 6 indicate varying levels of model accuracy across different machine learning algorithms employed for heart disease prediction. Notably, Convolutional Neural Network (CNN) and Random Forest (RF) models exhibit the highest accuracy at 89.0%, closely followed by Neural Network with 87.3%.

Decision Tree and Support Vector Machine (SVM) models achieve accuracies of 80.3% and 86.910%, respectively, while k-Nearest Neighbors (KNN) and Long Short-Term Memory (LSTM) models perform moderately well, with accuracies of 85.280% and 85.3%, respectively. However, Naive Bayes (NB) and Artificial Neural Network (ANN) models demonstrate comparatively lower accuracies at 76.9% and 85.5%, respectively. These results underscore the effectiveness of certain algorithms, such as CNN and RF, in accurately predicting heart disease based on the given dataset, while highlighting the importance of selecting appropriate algorithms for achieving optimal predictive performance.

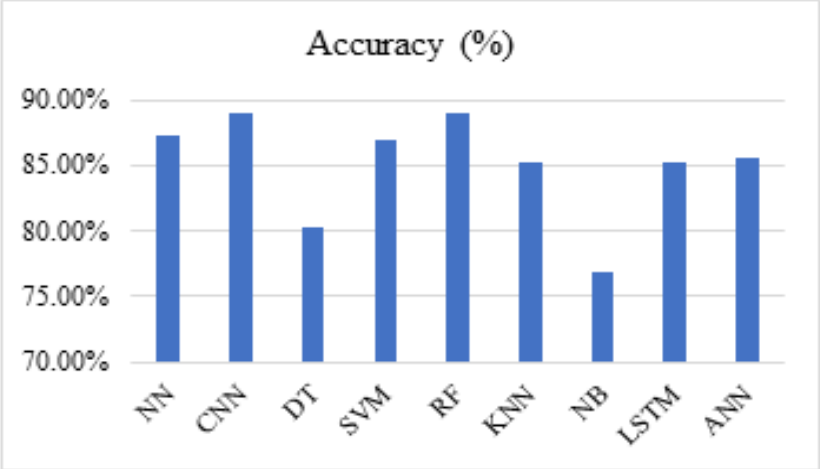

Figure 6: *Accuracy of various models for heart disease prediction*

7.2 PRECISION

The precision scores, indicating the proportion of true positive predictions among all positive predictions made by the models, vary across different machine learning algorithms used for heart disease prediction. Notably, the CNN and LSTM models achieve the highest precision scores at 89.7% and 91.9%, respectively, highlighting their effectiveness in accurately identifying cases of heart disease. Following closely are the RF and SVM models with precision scores of 90.0% and 85.7%, respectively, indicating their ability to make accurate positive predictions. The Neural Network and KNN models also demonstrate respectable precision scores of 87.8% and 84.0%, respectively, while the Decision Tree model achieves a precision score of 80.6%. However, the NB model exhibits a lower precision score of 77.7%, suggesting a higher rate of false positive predictions. Overall, these precision scores provide insights into the models' abilities to accurately identify cases of heart disease, highlighting the importance of selecting appropriate algorithms to achieve optimal precision in predictive modeling tasks (Figure 7).

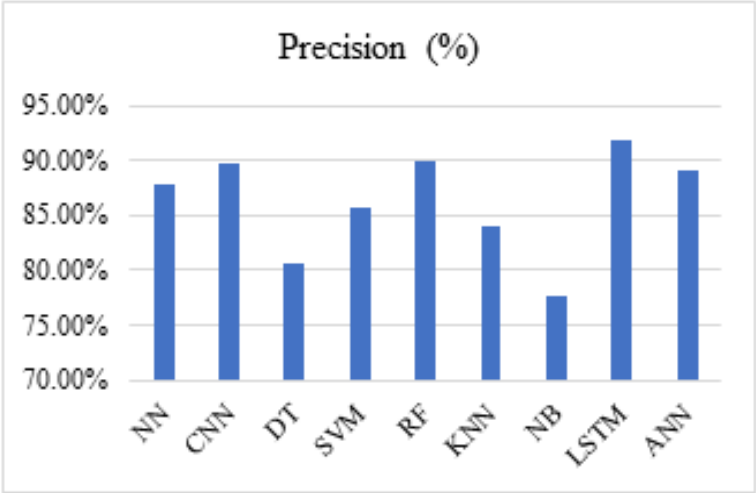

Figure 7: *Precision of various models for heart disease prediction*

7.3 RECALL

The recall scores, representing the proportion of true positive predictions among all actual positive cases, exhibit variability across different machine learning models utilized for heart disease prediction. Notably, the LSTM model achieves the highest recall score at 91.0%, closely followed by the CNN model with a score of 88.3%, indicating their effectiveness in capturing a high percentage of actual positive cases. The RF model also demonstrates strong performance with a recall score of 88.7%, highlighting its ability to correctly identify individuals with heart disease. Other models, such as the Neural Network and SVM, achieve respectable recall scores of 86.1% and 84.2%, respectively. However, the Decision Tree and KNN models exhibit lower recall scores at 78.0% and 82.4%, suggesting a relatively higher rate of false negative predictions. Additionally, the NB model shows the lowest recall score of 76.1%, indicating its limitations in correctly identifying positive cases. Overall, these recall scores provide valuable insights into the models' abilities to effectively capture true positive cases of heart disease, underscoring the importance of selecting appropriate algorithms to optimize recall in predictive modeling tasks (Figure 8).

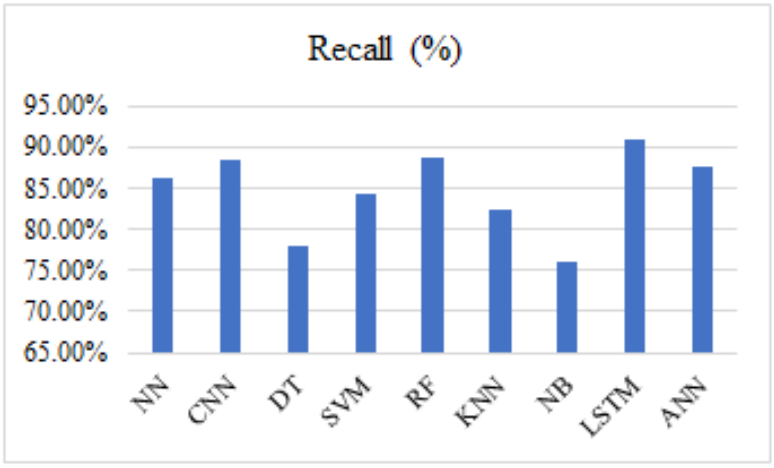

Figure 8: *Recall of various models for heart disease prediction*

7.4 F1 Score

The F1 scores, which represent the harmonic mean of precision and recall, provide a comprehensive evaluation of the performance of machine learning models used for heart disease prediction. Among the models, the LSTM model achieves the highest F1 score at 91.6%, closely followed by the CNN model with a score of 89.0%. These scores indicate that both LSTM and CNN models effectively balance precision and recall, resulting in a high overall performance in identifying cases of heart disease. The RF model also demonstrates strong performance with an F1 score of 89.3%, suggesting a good balance between precision and recall. Other models, such as the Neural Network and SVM, achieve respectable F1 scores of 86.9% and 84.9%, respectively. However, the Decision Tree and KNN models exhibit lower F1 scores at 79.3% and 83.2%, respectively, indicating potential room for improvement in achieving a better balance between precision and recall. Additionally, the NB model shows the lowest F1 score of 76.9%, suggesting limitations in effectively combining precision and recall. Overall, these F1 scores provide valuable insights into the overall performance of the models in accurately identifying cases of heart disease, highlighting the importance of selecting appropriate algorithms to optimize both precision and recall in predictive modeling tasks (Figure 9).

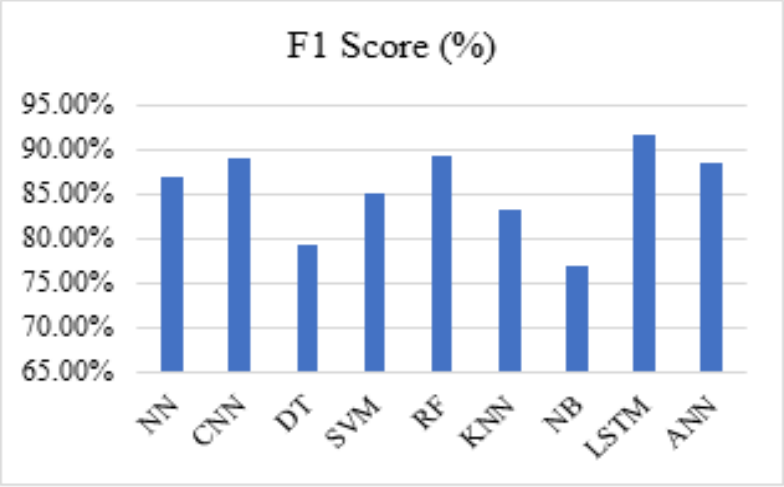

Figure 9: *F1 Score of various models for heart disease prediction*

The table provides a comprehensive evaluation of various machine learning models' performance in predicting heart disease, as indicated by their accuracy, precision, recall, and F1 score metrics. Notably, the LSTM model emerges as the top performer, exhibiting the highest scores across all metrics, including accuracy (85.3%), precision (91.9%), recall (91.0%), and F1 score (91.6%). This suggests that the LSTM model effectively balances both precision and recall, resulting in accurate identification of positive cases of heart disease. Additionally, the CNN model also demonstrates strong performance, with high scores across all metrics, indicating its effectiveness in accurately predicting heart disease. Conversely, the Naive Bayes model shows the lowest scores across all metrics, highlighting its limitations in accurately identifying positive cases of heart disease. Overall, the results underscore the importance of selecting appropriate machine learning algorithms to optimize predictive performance in heart disease prediction tasks.

8. CONCLUSION

The paper presents a comprehensive evaluation of various machine learning models for predicting heart disease based on metrics including accuracy, precision, recall, and F1 score. Among the models assessed, the Long Short-Term Memory (LSTM) model emerges as the top performer, exhibiting superior performance across all metrics, with accuracy, precision, recall, and F1 score all surpassing 85%. This indicates that the LSTM model effectively balances both precision and recall, resulting in accurate identification of positive cases of heart disease. Additionally, the Convolutional Neural Network (CNN) model also demonstrates strong performance,

highlighting its effectiveness in accurately predicting heart disease. Conversely, the Naive Bayes model exhibits the lowest performance across all metrics, suggesting limitations in accurately identifying positive cases of heart disease. These findings underscore the importance of selecting appropriate machine learning algorithms to optimize predictive performance in heart disease prediction tasks, with the LSTM and CNN models showing promise for accurate risk assessment and early detection of cardiovascular conditions.

REFERENCES

[1] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, "A Hybrid Evolutionary Algorithm for Attribute Selection in Data Dining", Expert Systems with Applications, Vol.36, No.4, pp.8616-8630, 2009.

[2] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology, Vol.2, No.4, pp.56-66, 2014.

[3] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET-International Journal of Innovative Science, Engineering & Technology, Vol.2, pp.441-444, 2015.

[4] A.Sahaya Arthy, G. Murugeshwari "A Survey on Heart Disease Prediction using Data Mining Techniques" (April 2018).

[5] Hamid Reza Marateb and Sobhan Goudarzi, "A Non-invasive Method for Coronary Artery Diseases Diagnosis using a Clinically Interpretable Fuzzy Rule-based System," Journal of Research in Medical Sciences, Vol. 20, Issue 3, pp.214-223, March 2015.

[6] The Guardian." UK heart disease fatalities on the rise for first time in 50 years". https://www.theguardian.com/society/2019/may/13/heart_x0002_circulatory-disease-fatalities-on-rise-in-uk. Accessed 25 Oct 2019.

[7] S. Mohan: "Effective Heart Disease Prediction Using Hybrid ML Techniques", VOLUME 7, 2019, DOI 10.1109/ACCESS.2019.2923707

[8] Maruf Ahmed Tamal 2019, Heart Disease Prediction based on External Factors: A Machine Learning Approach, International Journal of Advanced Computer Science and Applications (IJACSA), Volume 10 Issue 12, 2019.

[9] S. M. M. Hasan, M. A. Mamun, M. P.Uddin, and M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Feb. 8-9, 2018, doi: 10.1109/IC4ME2.2018.8465594.

[10] K. C. Howlader, M. S. Satu, and A. Mazumder, "Performance Analysis of Different Classification Algorithms that Predict Heart Disease Severity in Bangladesh," 2017 International Journal of Computer Science and Information Security (IJCSIS), vol. 15, no. 5, pp. 332-340, May, 2017, doi: 10.1109/CEEICT.2016.7873142.

[11] Deeanna Kelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal

[12] Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pg no:1

[13] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Med-ical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications - march-2017

[14] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Sept. 22-24, 2016, doi: 10.1109/CEEICT.2016.7873142.

[15] V. Sree Hari Rao, M. Naresh Kumar, "Novel Approaches for Predicting Risk Factors of Atherosclerosis," IEEE Journal of Biomedical and Health Informatics., vol. 17, No. 1, Jan 2013.

[16] Paolo Melillo, Nicola De Luca, Marcello Bracale and Leandro Pecchia ,"Classification Tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability", IEEE Journal of Biomedical and Health Informatics., Vol. 17, No. 3, May 2013.

[17] Minas A. Karaolis,Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees," IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 3, May 2010

[18] Carlos Ordonez, "Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, April 2006.

[19] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", Proceedings of IEEE Conference on Information & Communication Technologies, 2013.

[20] Sikander Singh Khurl, Gurpreet Singh, "Ranking Early Signs of Coronary Heart Disease Among Indian Patients", IEEE International Conference on Computing for Sustainable Global Development, 2015

[21] Beigh, M. A, Quadri Javeed Ahmad Peer, Kher , S. K. and Ganai, N. A, "Disease and pest management in apple: Farmers' perception and adoption in J&K state", Journal of Applied and Natural Science 7 (1): 293 – 297 (2015)