REAL TIME ANOMALY DETECTION IN ELECTRONIC HEALTH RECORDS

Dhanasekar S[1], Sangeethaa S N[2]

[1] Student, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India

[2]Faculty, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India

ABSTRACT

Electronic Health Records (EHRs) are invaluable in healthcare, offering a wealth of patient data. Timely anomaly detection in EHRs is critical for identifying emerging health risks, enabling early intervention, and enhancing patient care. The project addresses the need for effective anomaly detection methods in the context of coronary heart diseases using machine learning algorithms. The aim of this project is to evaluate and compare the performance of several machine learning algorithms, including Random Forest, Decision Tree, Naive Bayes, Artificial Neural Networks (ANN). Analyzing the best algorithm based on their suitability, accuracy, and adaptability to evolving data patterns. The project meticulously preprocesses EHR data, identifies relevant features, and the algorithm is used to improve anomaly detection. Each algorithm is trained and optimized, considering performance metrics such as precision, recall, and F1-score. Real-time inference capabilities are assessed, emphasizing the need for models to adapt to changing data patterns. Continuous monitoring and model updating are emphasized to minimize false alarms. The integration of an alerting mechanism facilitates timely healthcare professional intervention upon anomaly detection. In the comparative analysis, Random Forest and ANN demonstrate superior performance, capturing intricate data relationships effectively. Decision Tree and Naive Bayes show moderate performance.

The project underscores the significance of real-time anomaly detection in EHRs for coronary heart diseases. It highlights the strengths and weaknesses of various machine learning algorithms in this context. Random Forest and ANN emerge as promising choices, balancing accuracy and interpretability. Ultimately, this research contributes to the advancement of healthcare systems by enhancing their ability to proactively detect anomalies, thereby improving patient outcomes. To address this, data augmentation techniques, such as random flip, random rotation, random translation, random zoom, random contrast, random hue, random brightness, and random saturation, were employed to reduce perspective variability. The study evaluated the performance of different design strategies to identify the approach that achieves the highest accuracy in monument recognition.

Keyword : *Artificial Neural Networks (ANN), Decision Tree, Naive Bayes, Random Forest.*

1.INTRODUCTION

1.1 DATA MINING

Data mining is a crucial analytical process that involves uncovering patterns and correlations within data to make predictions. This multi-stage process begins with exploration, where data is prepared and preliminary feature selection is performed. The next stage involves model construction and validation, where different models are evaluated to select the most accurate one. Finally, in the deployment stage, the chosen model is applied to new data to generate forecasts or estimates.

1.2 MACHINE LEARNING

Machine learning is a subset of artificial intelligence that focuses on algorithms that learn from data and improve their performance over time without explicit programming. It encompasses a wide range of applications, including email sorting and computer vision, where traditional algorithms may be insufficient. Machine learning involves the use of training data to build models that can make predictions or decisions based on new input data, making it a powerful tool in various fields.

1.3 HEART DISEASE

Heart disease encompasses a wide range of conditions affecting the heart and blood vessels. These conditions can include congenital heart defects, arrhythmias, and vascular diseases like coronary artery disease. Coronary artery disease, in particular, can lead to reduced blood flow to the heart muscle, potentially causing weakness or even heart muscle death. Risk factors for heart disease include smoking, high cholesterol, hypertension, diabetes, and genetic predisposition, which can damage the blood vessels and increase the risk of blockages.

1.4 FEATURE SELECTION

Feature selection is a critical step in creating predictive models. It involves choosing the most relevant data variables while reducing computational complexity. Statistical measures are used to assess the relationship between input variables and the target variable, helping select the most significant features. Feature selection can be categorized into supervised and unsupervised methods, with the former relying on statistical measures to determine feature importance based on the output variable and the data type.

1.5 PREDICTIVE MODEL

Predictive models use statistics and machine learning techniques to make predictions about future events or outcomes based on historical and current data. These models are widely used in various fields, from identifying suspects in criminal investigations to assessing risk in financial markets. Predictive modeling involves selecting appropriate algorithms to make predictions and evaluating their performance. It can be closely related to artificial intelligence and is essential for making informed decisions in many applications, such as marketing, fraud detection, and healthcare.

2.EXISTING MODEL

In existing, there are different AI calculations like LR, KNN, SVM, and GBC, along with the Grid Search CV, anticipate heart illness. The framework involves a 5-overlap cross-approval method for confirmation. A near report is given for these four strategies. The Datasets for both Cleveland, Hungary, Switzerland, and Long Ocean side V and UCI Kaggle are utilized to investigate the models' presentation. It is apparent that among the proposed approach, the Outrageous Angle Helping Classifier with Grid Search CV is delivering the best hyper boundary for testing precision. The essential point of this paper is to foster an interesting model-creation method for taking care of true issues. To categorize the presence and absence of the condition, the different types of machine and deep learning algorithms are available. Logistic regression (LR) techniques are used in this study. To categorize heart illness, UCI dataset is used. In order to enhance the performance of the model, the dataset was pre-processed by being analyzed, the missing values were corrected, and features were chosen based on correlation with the intended value for each feature. The traits with the highest favorable associations were picked. The dataset is then divided into training and test sets to perform classification: 90:10, 80:20, 70:30, 40:60, and 50:50 testing ratios. As shown below, the splitting ratio of 90:10 provides the best accuracy. The LR model’s accuracy was 87%.

.

2.1 DRAWBACKS

- The level of data quality affects accuracy.
- The prediction stage could be cumbersome with extensive data and Attentive to irrelevant aspects and the size of data.

- When the target classes overlap and the amount of noise in the data set is higher, it performs poorly.
- The performance of the support vector machine is poor when there are more attributes for each data point than there are training data samples.

3.OBJECTIVES

- To find out which model is accurate in predicting coronary artery disease.
- To demonstrate that deep learning methods are more accurate than machine learning techniques at predicting CAD.
- To show the features which contribute more to coronary artery disease.
- To do a comparative analysis of different models in predicting CAD.

4.PROPOSED SYSTEM

The patient record of coronary illness is utilized as the information. The fundamental target of our venture is to characterize the informational index utilizing the WARM calculation. The expectation is performed from mining the patient's verifiable informational collection. In Weighted Affiliated rule mining (WARM), DT, ANN various loads are appointed to various qualities as per their foreseeing ability. It has been demonstrated that the choice tree classifiers are performing great than customary classifiers approaches, for example, choice tree and ANN. Further from exploratory outcomes it has been found that WARM is giving superior exactness as contrast with other previously existing Affiliated Classifiers. Consequently, the framework is utilizing WARM, DT, ANN as a strategy to produce rule baseThe input is the electronic dataset of patients. The WARM method is employed in this project to classify the data set. Utilizing historical patient data mining, the forecast is made. Weighted Associative Rule Mining (WARM) assigns various attributes various weights based on their likelihood to foresee. The algorithms Decision Tree, Random Forest, Naive Bayes, and ANN have used. Here, a comparative analysis is done to predict which algorithm correctly predicts heart disease. The WARM technique is used by the system to produce rules. It has been proven that ANN classifiers, which offer 95% accuracy, beat traditional classifiers like decision trees, NB, and RF.

4.1 ADVANTAGES

- Can easily predict the heart Disease level and severity using range level of queries.
- As deep learning algorithm(ANN) is used, it gives more accuracy as compared to machine learning algorithms.

- Accuracy of these algorithm is very high.

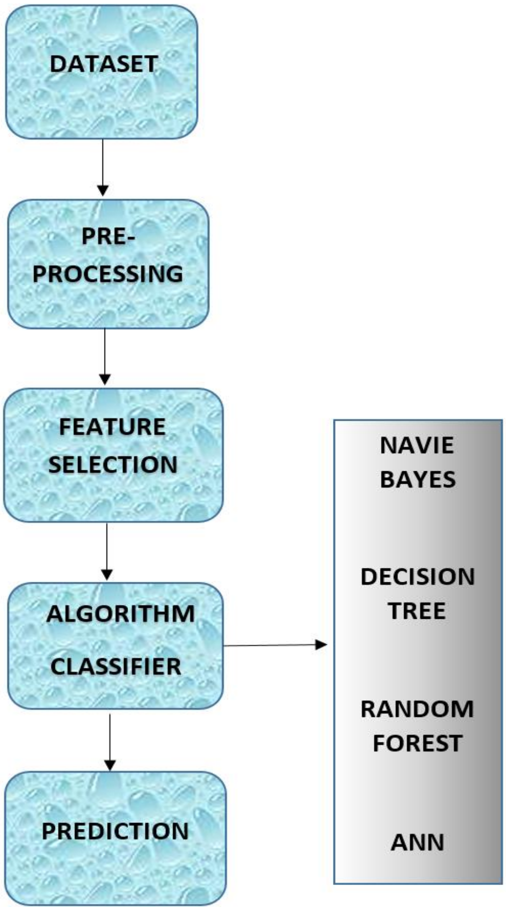

Figure 4.1 Proposed Methodology

4.2 PRE-PROCESSING

The missing qualities are supplanted with suitable qualities. The ID of the patient cases doesn't add to the classifier execution.

Consequently it is taken out and the result trait characterizes the objective or ward variable in this manner diminishing the list of capabilities size. The algorithmic procedures applied for highlight significance examination and order are extravagantly

introduced in the accompanying segments.

4.3 ALGORITHM IMPLEMENTATION

We carry out the new characterization approach that utilization affiliation rule mining and order has turned into a critical
instrument for information disclosure. A significant benefit of these characterization frameworks is that, utilizing weighted
Affiliation Rule Mining (WARM) they can look at a few elements all at once. While other condition of workmanship techniques like choice tree consider that component is free of one another. Numerous applications can profit from great arrangement model. Cooperative classifiers are particularly fit to applications where the model might help the space specialists in their choices. There are numerous spaces like clinical, where the most extreme exactness of the model is wanted and thus the precision of the cooperative classifiers.The forecast outcome is likewise recognized in the WAC, DTThe choice tree
classifier makes the grouping model by building a choice tree. Every hub in the treeto LateX determines a test on a quality, each branch plummeting from that hub relates to one of the potential qualities for that property.

4.4 PERFORMANCE EVALUATION

Grouping exactness is just the pace of right arrangements, either for an autonomous test set, or utilizing some variety of the cross- approval thought. the arranged precision result is determined and out put is displayed in the tree design with the quantity of significant vessels , parting into present or missing examples with the accurately and mistakenly characterized are recognized. Default peril is the open door that associations or individuals will be not ready to make the important portions on their commitment responsibilities. It gives reproducible and objective determination, and subsequently can be a significant
assistant device in clinical practices. Results are similarly, encouraging and in this manner the proposed technique will be useful in illness diagnostics. Later, the data has been divided into two types which is training data and testing data. For training it takes 80% and for testing it takes 20%. To train the model, we employed the Decision Tree, RF, NB, and ANN algorithms. The proposed approach evaluates the model using the testing set once it has been trained. A types of performance indicators, contains precision for exactness, recall for completeness, accuracy for precision and f1 scores for harmony.

5.PROPOSED WORD MODULES

5.1 SOFTWARE DESCRIPTION

5.1.1 IDLE

The name of a Python integrated development environment (IDE) is IDLE (Integrated Development and Learning Environment). The IDLE module is automatically included in the Python installation for Windows. IDLE has the following features:

- Code has 100% pure Python and Python shell window (interactive interpreter) with colourizing of code input, output, and error messages is cross-platform and performs largely the same on Windows, Unix, and macOS.
- a multi-window text editor with a wide variety of undo options, Python colorization, smart indent, call hints, auto completion, and other features.
- A debugger with continuous breakpoints, stepping, and the display of local and global namespaces; search in any window, replace in editing windows, and search through numerous files configuration, browsers, and other dialogues.

5.1.2 GOOGLE COLAB

The Colaboratory, sometimes known as "Colab," is a data analysis and machine learning tool that enables you to integrate the text, graphs, photos, executable code, HTML and LaTeX, and more into a single Google Drive document. Colaboratory, or "Colab" for short, is a creation of Google Research. Three fields where Colab excels include data analysis, education, and machine learning. Anyone can create and execute any Python code through the web. Technically speaking, Colab is a hosted Jupyter notebook service that offers free access to computer resources, including GPUs, and doesn't require any setup.

5.1.3 FEATURES OF COLAB

- Create and run Python code and describe the programming you used to support mathematical equations.
- Make, upload, and share notebooksand upload and save notebooks to Google Drive.
- Publish or import notebooks from GitHub and Import outside datasets, such as those from

Kaggle.

- Combine Keras, TensorFlow, PyTorch, and OpenCV and no-cost Cloud service with no-cost GPU.

5.2 HARDWARE DESCRIPTION

FEATURES	**MINIMUM REQUIREMENTS**	**MAXIMUM REQUIREMENTS**
PROCESSOR TYPE	Pentium 13	**Intel Corei7**
SPEED	3.40GHZ	3.90 GHZ
RAM	4GB DD2 RAM	16 GB NVIDIA
HARDDISK	500GB	1 TB
KEYBOARD	101/102 Standard Keys	104 keys
MOUSE	Optical Mouse	Track Pad

6.RESULT AND DISCUSSION

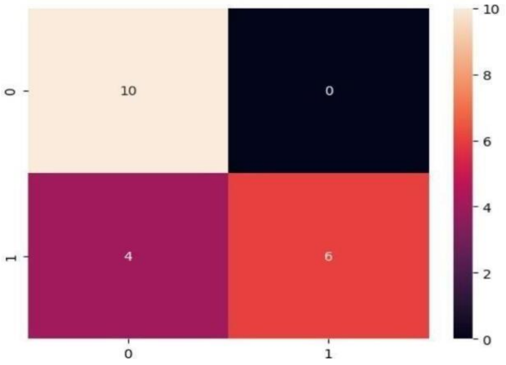

Figure 6.1 Confusion matrix of DT

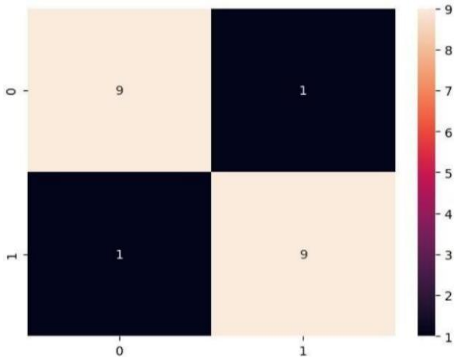

Figure 6.2 Confusion matrix of NB

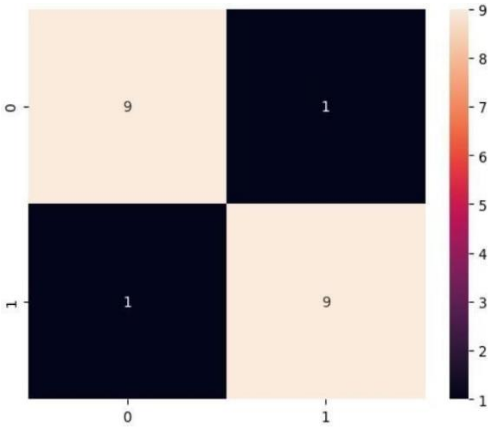

Figure 6.3 Confusion matrix of ANN

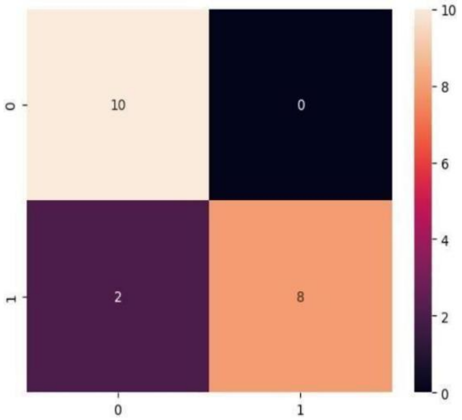

Figure 6.4 Confusion matrix of RF

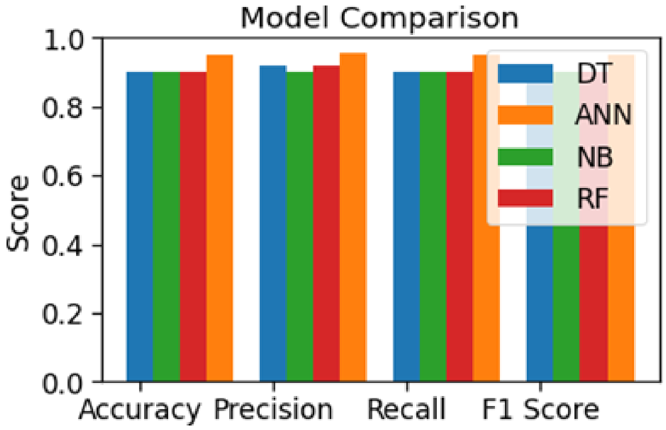

Figure 6.5 Model Comparison

6.1 DISCUSSION

The Dataset is taken from Cleveland and UCI repositories and is loaded using the pandas library. Pre-processing is carried out to enhance the performance of the dataset. The attributes in the dataset have gained weightage using the WARM algorithm and is classified into training and testing data. We use Decision Tree to classify problems and categorize objects depending on their learning features. We use Naïve Bayes for high accuracy feature selection and this algorithm helps in predicting results which are reliable and quick. We also use Random Forest algorithm as it gives both the features of Decision tree and Naïve Bayes thus giving a more reliable result. ANN aids in understanding the effects of changing the dataset's size either vertically or horizontally on computation time as well as the circumstances or conditions in which the model works best. It also aids in the explanation of why a particular model performs better in a particular setting or circumstance. All of these models precision for exactness, recall for completeness, accuracy for precision and f1 scores for harmony are evaluated to produce an effective model that effectively predicts heart disease. A comparison graph of the four models are plotted and the more possible cause for CAD is found using the feature comparison graph. The result obtained in ANN and Random Forest is higher and effective than the existing model.

7. CONCLUSION AND FUTURE WORK

In summary, this study is a significant advancement in the field of real-time anomaly detection in coronary heart disease electronic health records (EHRs). The usefulness of several machine learning algorithms, such as Decision Tree, Navie Bayes, Random Forest, Artificial Neural Networks (ANN), and the revolutionary "Warm" method, in the healthcare industry has been clarified by our thorough study of these algorithms. Our findings underscore the potential of Random Forest and ANN as promising choices for effective real-time anomaly detection due to their ability to capture intricate data patterns. However, the study also reveals that the "Warm" algorithm requires further refinement and investigation. Pre-processing of the corpus, such as cleaning and missing value detection, is done to increase performance. The crucial step is feature selection, which improves algorithm accuracy and especially focuses on the algorithm's behaviour. When compared to earlier research, the results outperformed it, and here ANN is providing an enhanced accuracy of 95% over other machine learning methods. Thanks to this improved precision, medical personnel will be able to anticipate cardiac sickness more quickly.

7.1 FUTURE WORK

Future work in real-time anomaly detection within electronic health records (EHRs) holds immense potential to revolutionize healthcare. One critical area for exploration is the seamless integration of these systems into the clinical workflow. By designing solutions that fit seamlessly within healthcare practices, we can provide timely alerts and decision support to healthcare professionals without causing disruptions, ultimately enhancing patient care and safety. Privacy-preserving approaches will continue to be a significant focus, allowing us to balance the need for anomaly detection with patient data privacy. Advancements in explainable AI (XAI) will play a pivotal role in making the outputs of anomaly detection models more interpretable and trustworthy for healthcare providers, further improving adoption rates and clinical utility. Additionally, the adaptability of these models is essential. Continual learning approaches can enable models to evolve alongside changing patient populations, emerging medical practices, and the incorporation of new data sources. Collaboration between healthcare institutions, scalability considerations, and comprehensive validation through real-world trials are also imperative for the successful implementation of real-time anomaly detection systems. Ultimately, future work in this field should be driven by a commitment to improving patient outcomes, enhancing healthcare quality, and ensuring the ethical and responsible use of EHR data.

8.REFERENCES

[1] Ambrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, Kiran Menisnkal, " Logistic Regression Technique for Prediction of Cardiovascular Disease".2022.

[2] Ghulab Nabi Ahmad, Hira Fatima, "Efficient Medical Diagnosis of human heart diseases using machine learning techniques with and without GridSearcCV," ,ACCESS.2022.3165792.

[3] Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, KarstenSternickel, Lijuan Zhu, "Using Efficient Supanova Kernel for Heart Disease Diagnosis", proc. ANNIE. 06,intelligent engineering systems through artificial neural networks, vol. 16, pp:305310, 2021.

[4] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021.

[5] S. I. Ansarullah and P. Kumar, ''A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method,'' Int. J. Recent Technol. Eng., vol. 7, no. 6S, pp. 1009–1015, 2021.

[6] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, ''Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection,'' in Proc. IEEE 5th Int. Conf. Converg. Technol. (ICT), Mar. 2020 pp. 1–4 .

[7] U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, ''A novel integrated diagnosis method for breast cancer detection,'' J. Intell. Fuzzy Syst., vol. 38, no. 2, pp. 2383–2398, 2020.

[8] S. Mohan, C. Thirumalai, and G. Srivastava, ''Effective heart disease prediction using hybrid machine learning techniques,'' IEEE Access, vol. 7, pp. 81542–81554, 2021

[9] G. G. N. Geweid and M. A. Abdallah, ''A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique,'' IEEE Access, vol. 7, pp. 149595–149611, 2020.

[10] Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, KarstenSternickel, Lijuan Zhu, "Using Efficient Supanova Kernel for Heart Disease Diagnosis", proc. ANNIE 06,intelligent engineering systems through artificial neural networks, vol. 16, pp:305310, 2021.

[11] SellappanPalaniappan, RafiahAwang, ―Intelligent Heart Disease Prediction System Using Data Mining Techniques‖; 9781424419685/08/$25.00©2020 IEEE.

[12] Chaitrali S. Dangareet. al., "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", (IJCA) (0975 – 8887), Vol. 47, No. 10, June 2020, page no. 4448.

[13] Jabbar MA, Deekshatulu BL, Priti C. Heart disease classification using nearest neighbour classifier with feature subset selection. Annals Computer Science 2021.

[14] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2020.

[15]R. Sivaranjani, V. S. Naresh, and N. V. Murthy, ‘‘4 coronary heart disease prediction using genetic algorithm based decision tree,’’ Intell. Decis. Support Syst., Appl. Signal Process., vol. 4, p. 71, Oct. 2020.