

# OHE2LM: A Hybrid Approach Towards Heart Attack Prediction using One-Hot Encoding based Extreme Learning Machine Model

Pawan Kumar Mall<sup>1</sup>, Swapnita Srivastava<sup>2</sup>, Mitul M. Patel<sup>3</sup>, Aniruddh Kumar<sup>4</sup>, Vipul Narayan<sup>5</sup>, Sanjay Kumar<sup>6</sup>, P. K. Singh<sup>7</sup>, D. S. Singh<sup>8</sup>

<sup>1</sup>Assistant Professor, GL Bajaj Institute of Technology and Management  
pawankumar.mall@gmail.com

<sup>2</sup>Assistant Professor, GL Bajaj Institute of Technology and Management  
swapnitasrivastava@gmail.com

<sup>3</sup>Assistant Professor, Department of Electronics and Communication Engineering, Parul Institute of Engineering and Technology,  
Parul University, Vadodara, India;  
patelmitul4388@gmail.com

<sup>4</sup>Department of Computer Science and Engineering, Galgotias College of Engineering and Technology  
Knowledge Park-2 Greater Noida  
aniruddh.knit@gmail.com

<sup>5</sup>Assistant Professor, Galgotias University  
Gautam Buddha Nagar, Uttar Pradesh  
vipulpsainian2470@gmail.com

<sup>6</sup>Assistant Professor, Rajkiya Engineering College Azamgarh  
sanjay@gecazamgarh.ac.in

<sup>7</sup>Professor, Computer Science and Engineering Department, Madan Mohan Malaviya University Of Technology, Gorakhpur, 273010, Uttar Pradesh, India.  
topksingh@gmail.com

<sup>8</sup>Associate Professor, Computer Science and Engineering Department, Madan Mohan Malaviya University Of Technology, Gorakhpur, 273010, Uttar Pradesh, India.  
dss\_mec@yahoo.co

**Abstract:** Predicting heart attacks stands as a significant concern contributing to global morbidity. Within clinical data analysis, cardiovascular disease emerges as a pivotal focus for forecasting, wherein Data Science and machine learning (ML) offer invaluable tools. These methodologies aid in predicting heart attacks by considering various risk factors Just like high blood pressure, increased cholesterol levels, irregular pulse rates, and diabetes, this research aims to enhance the accuracy of predicting heart disease through machine learning techniques. This study introduces a MLdriven approach, termed ML-ELM, dedicated to forecasting heart attacks by analysing diverse risk factors. The proposed ML-ELM model is compared with alternative Utilizing machine learning techniques like Support Vector Machines, Logistic Regression, Naïve Bayes, and XGBoost is a key aspect of this exploration into different approaches for predictive modeling., is part of the research strategy. The dataset utilized for heart disease symptoms is sourced from the UCI ML Repository. The outcomes reveal that our proposed ML-ELM model has demonstrated superior predictive performance among the ML techniques tested. ML models show notable efficiency in identifying heart attack symptoms, particularly with boosting algorithms. Accuracy assessments were employed to gauge the predictive ability, Our suggested model demonstrated an outstanding accuracy rate of 96.77%.

## I. Introduction

Cardiovascular disease (CVD) is currently the primary concern in the medical field, representing a significant and persistent health threat with high global fatality rates. According to recent World Health Organization statistics, about 20.5 million of people succumb to cardiovascular disease annually, constituting 31.5% of worldwide deaths. Projections indicate a potential increase in this annual death toll to 24.2 million by 2030. The majority (85%) of

cardiovascular disease-related fatalities are attributed to heart attacks and strokes. A heart attack occurs when arterial plaque buildup hinders blood flow to the heart, while a stroke results from a blood clot in a brain artery, disrupting blood circulation to this crucial organ. Primarily, heart disease emerges when the heart struggles to adequately supply blood to various parts of the body [1]. This leads to initial indications like irregular heart rhythms, breathing difficulties, chest pain, abrupt dizziness, queasiness, swollen feet, and perspiration. Timely and precise forecasting along with correct diagnosis of heart

disease show an essential role in enhancing. The survival rates of using patients are influenced by various factors related to cardiovascular disease. The symptoms include high pressure of blood, cholesterol levels, alcohol consumption, smoking, obesity, insufficient physical activity, and genetic variations. Early identification of warning signs and lifestyle modifications, including increased physical activity, abstaining from smoking, and undergoing suitable medical evaluations conducted by healthcare professionals, can significantly contribute to decreasing mortality rates. [2].

Presently, the methods utilized for forecasting and identifying heart disease rely largely on scrutinizing a patient's medical background, symptoms, and physical assessment findings conducted by physicians. Frequently, medical professionals encounter challenges in accurately forecasting a patient's heart condition, achieving an accuracy rate of approximately 67%. This difficulty arises due to the current diagnostic approach, which relies on associating the symptoms observed in the patient under consideration with those identified in previously diagnosed cases [3] [4]. Consequently, there is a pressing need in the medical field for an automated, intelligent system to precisely forecast heart disease. Achieving this objective involves tapping into the extensive pool using patient data within the medical domain and utilizing other machine algorithms. [5]. Currently, research teams in the field of data science have increasingly focused on disease prognosis. This heightened interest is a direct result of the swift advancement of sophisticated computing technologies in healthcare and the accessibility of extensive health-related databases. The fusion of cutting-edge deep-learning technologies and intelligent decision-making systems holds significant promise in enhancing healthcare support within our society. Data stands as the most invaluable resource for acquiring fresh insights, gathering crucial information, and augmenting existing knowledge.

Various sectors, including science, technology, agriculture, business, education, and health, harbor an immense volume of data commonly referred to as "big data." This data exists in a raw state, either structured or unstructured, awaiting processing and analysis. [6]. Deriving valuable insights from big data requires the storage, processing, analysis, management, and visualization of this data through comprehensive data analysis [7].

Presently, within the healthcare domain, databases containing patient-related medical reports are abundantly available and expanding continuously. However, this raw data is notably redundant and imbalanced. Effectively leveraging this data requires preprocessing steps. Advances in computing capabilities and the flexible nature of machine learning play a crucial role in enhancing these procedures. This opens up new

avenues for research using healthcare sector, especially in early disease prediction, such as cardiovascular disease as well as cancer, with the ultimate goal of improving survival rates. [8].

Machine learning presents superior predictive modelling tools that aim to tackle the existing constraints [9]. It holds significant promise in leveraging big data for the development of prediction algorithms. This approach depends on computational systems to comprehend intricate and non-linear relationships among characteristics by minimizing discrepancies between observed results and actual one [10]. The system assimilates patterns using the features present within the current dataset and employs these patterns to predict outcomes within an unfamiliar dataset. Among the potent machine learning techniques for prediction, classification stands out. It's a supervised machine learning method known for its efficacy in disease identification when trained with fitting data [11].

## II. Related work:

In [12] author have aim to create a model for accurately predicting cardiovascular diseases to mitigate the fatalities associated with such conditions. The study presents a method that incorporates k-modes clustering initialization to enhance classification accuracy. Various models like random forest, decision tree, multilayer perceptron, and XGBoost, were employed. The GridSearch CV method was utilized to fine-tune the models' parameters for optimal results. In [13] authors have introduced systematic bias, impacting the performance of models within specific demographic subgroups. This research aims to explore this issue using electronic health records data in conjunction with diverse machine learning models. Additionally, the study assesses the impact of three bias mitigation strategies: eliminating protected attributes, adjusting sample sizes through resampling, and balancing by proportion.

In [14] author have aim to create a model is to anticipate the probability of individuals developing heart disease to address this concern. In particular, the study compared different models to effectively classify and predict instances of cardiac arrest using concise features.

In [15] author have introduced a prediction model based on machine learning (MLbPM), which utilizes a fusion technique based on data scaling, split ratios, optimal parameters, utilizing a range of prediction algorithm based on machine learning.

In reference [16], the author presents machine learning approach focused on quantum machine technique based on classifiers. The proposed method enhances the performance compare to traditional approaches.

In [17] author have use, both the UCI dataset for heart disease and dataset based on real time are utilized to evaluate techniques for deep learning, comparing them against old methods. To enhance traditional methods accuracy, a cluster-based bi-directional long-short term memory technique is introduced.

In reference [18], the author presents a hybrid algorithm that combines the DT with Ada Boosting algorithms for predicting coronary heart disease (CHD). The evaluation of this approach's performance relies on metrics i.e accuracy, True Positive Rate and Specificity.

In [19], the writer introduces framework for smart healthcare intelligent designed to predict cardiovascular disease of heart using the Swarm-Artificial Neural Network approach. The procedure starts by generating a set number of NN randomly to train and test the framework, scrutinizing their solution consistency. Furthermore, the NN populations go through a two-stage training process that includes adjusting weights and integrating a newly formulated heuristic method. Ultimately, the precision of cardiovascular disease prediction relies on altering neuron weights through the distribution of globally optimal weights among other neurons.

III. Proposed Model

The dataset used sourced from the Cleveland database as well as which is publicly accessible at the UCI Learning Repository for machine under the title "Heart Disease Data Set." It comprises 11 features, namely age, sex, exang, ca, trtbps, chol, fbs, rest\_ecg, thalach, and target. Refer to table 1 for a detailed explanation of these features.

Table 1: Dataset details

| S.N. | Features | Details   |
|------|----------|---|
| 1.   | Age      | Patient's age   |
| 2.   | Sex      | Assigns patients a categorical label of either 1 or 0 based on gender. However, it does not clarify whether 1 corresponds to male or female, and the same ambiguity applies to 0. |
| 3.   | Exang    | Exercise-induced angina is denoted as 1 for "yes" and 0 for "no."   |
| 4.   | Ca       | Count of major vessels (0-3)  |
| 5.   | Cp       | Categorization of chest pain type: 1 for typical angina, 2 for atypical angina, 3 for non-anginal pain, and 4 for asymptomatic.   |
| 6.   | Trtbps   | If the fasting blood sugar level goes beyond 120 mg/dl, we represent it as 1 to convey true and as 0 to express false   |

|     |          |   |
|-----|----------|---|
| 7.  | Chol     | Cholesterol measured in mg/dl obtained through a BMI sensor   |
| 8.  | Fbs      | When the fasting blood sugar level exceeds 120 mg/dl, it's denoted as 1 to signify true, and 0 to indicate false. |
| 9.  | Rest_ecg | Electrocardiographic outcome  |
| 10. | Thalach  | Maximum achieved heart rate   |
| 11. | Target   | 0 indicates a lower likelihood of a heart attack, while 1 signifies a higher probability of a heart attack.       |

In this section, we explore the proposed methodology, encompassing pre-processing of data, classifiers machine learning, and metrics for performance.

Data pre-processing:

The data pre-processing alters raw data format understandable by machines. Real-world patient data is frequently inconsistent, incomplete, and prone to errors. Data pre-processing techniques are employed to address these issues. [20][21].

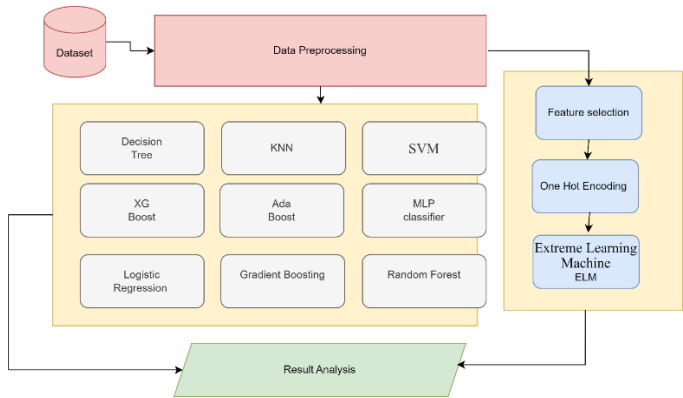


Figure 1: Block layout of proposed model

Decision Tree:

The Decision Tree, a widely used machine learning algorithm, functions in supervised learning for classification tasks. It adeptly classifies both categorical and continuous dependent variables by partitioning the population into more homogeneous sets using essential attributes or independent variables [22].

KNN (K-Nearest Neighbours) Algorithm:

Adaptable for classification as well as regression challenges, KNN is a prevalent choice in the field of Data Science, especially for classification tasks. This uncomplicated algorithm stores existing cases and classifies new ones by majority vote from its k neighbors, assigning the case's class

based on the most similar neighbors, determined through a distance function. [23].

### SVM Algorithm:

SVM algorithm classifies that represents raw data as points in an n-dimensional space associating each feature with a specific coordinate to aid in data classification. Using classifiers, it divides and visualizes the data [24] [25][26].

### Logistic Regression:

Logistic Regression estimates discrete values, typically binary (0/1), from independent variables by predicting event probabilities through a logit function. It's also known as logit regression.

### Gradient Boosting:

Particularly useful for complex datasets, Gradient Boosting excels in both speed and accuracy for prediction. This method reduces model bias error by iteratively building models and correcting the flaws of the preceding ones.

### Random Forest:

The "random forest" consists decision trees working collectively. It selects the most voted class from individual tree predictions to enhance accuracy, employing bagging and feature randomization techniques [27][28].

### MLP Classifier:

An MLP (multilayer perceptron) is an artificial neural network that produces outputs from inputs through multiple layers, trained using backpropagation—a technique associated with deep learning [29].

### AdaBoost:

AdaBoost, an ensemble learning technique, commonly utilizes decision trees with just one split (decision stumps) to improve overall prediction accuracy.

### XGBoost:

Employing decision trees sequentially and assigning significant weights to independent variables, XGBoost utilizes a combination of classifiers/predictors to build a more powerful model capable of addressing various problems like regression, classification, ranking, and custom predictions.

### One-Hot Encoding:

This method transforms categorical data into a numerical format compatible with machine learning algorithms, generating a binary variable for each unique integer value.

### Feature Selection:

Feature selection entails identifying the most pertinent attributes for a given dataset enhancing model accuracy by reducing noise and optimizing computational efficiency.

### Extreme Learning Machine (ELM):

ELM is an algorithm offering high performance at an exceptionally rapid learning pace. Unlike traditional neural network learning methods, ELM doesn't rely on gradient-based techniques, tuning all parameters at once without iterative training.

ELM algorithm has, by far, the easiest implementation of all. The implementation is easy, the algorithm yields excellent results with minimal computational time.

$$X = \{X_1, X_2, \dots, X_N\}, \quad (1)$$

We possess the following training set: where N denotes the training set, and X<sub>i</sub> is an array of values to database. X<sub>i</sub> shape length is equal to the database columns number without the class column. We have the following set of classes:

$$T = \{T_1, T_2, \dots, T_N\}, \quad (2)$$

where T<sub>i</sub> array of binary tree that includes the X<sub>i</sub> entry class in the training set.

Steps for implementing ELM:

Create the input layer weights matrix;

$$W = \begin{bmatrix} rand & \dots & rand \\ \vdots & \ddots & \vdots \\ rand & \dots & rand \end{bmatrix}, \quad (3)$$

Compute output matrix for hidden layer. Subsequently, activate the output matrix using any desired activation function.;

$$H = W * X, \quad (4)$$

Compute the Moore-Penrose pseudoinverse.

$$G^+ = (G^T * G)^{-1} * G^T, \quad (5)$$

Compute weight matrix, beta.;

$$\beta = H^+ * T, \quad (6)$$

We iterate through step 2 dataset was testing, generating a new matrix "H". Following that, we create result matrix called O, utilizing the aknownd beta matrix.



$$O = H * \beta , \tag{7}$$

Apply the Soft Max algorithm to transform the O matrix. Subsequently, compare the O matrix with the T matrix using the Winner Takes All algorithm.

IV. Result:

In this section we are going to analysis the purposed model and other algorithms accuracy parameter from confusion matrix. First of all we will examine the dataset figure 2 provide an overview age count of the patient. The figure 3 provides an overview age count of the patient with respect to cholesterol. The figure 4 depicts

es an overview age count of the patient with respect to trtbps. The figure 5 depicts an overview age count of patient with respect to thalachh. The figure 6 provides an overview age count of the patient with respect to heart attack rate. The figure 7 provides an overview age count of the patient with respect to thall.

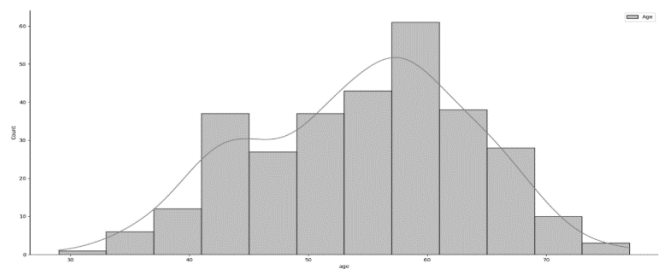


Figure 2: Age vs Count

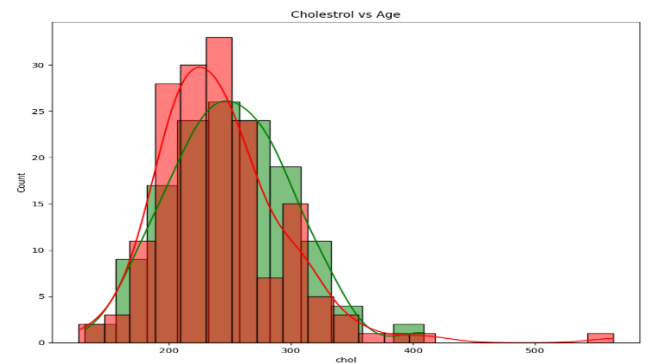


Figure 3: Chol vs Count

The figure 4 provides an overview age count of the patient to age groups.

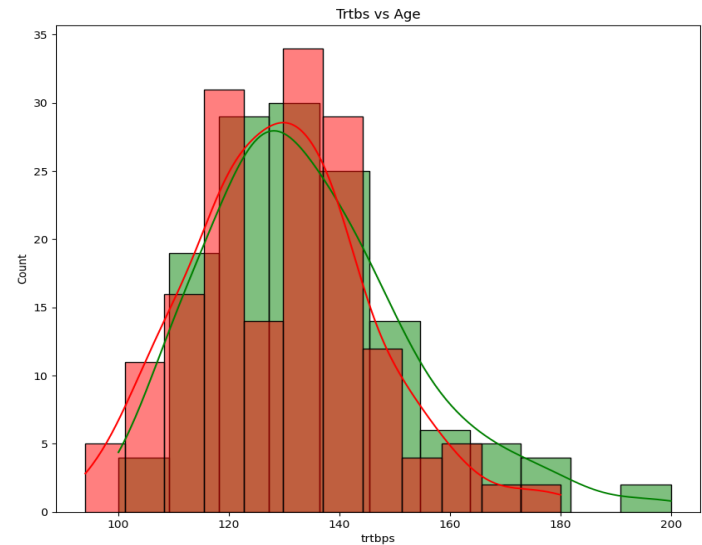


Figure 4: Age vs Trtbps

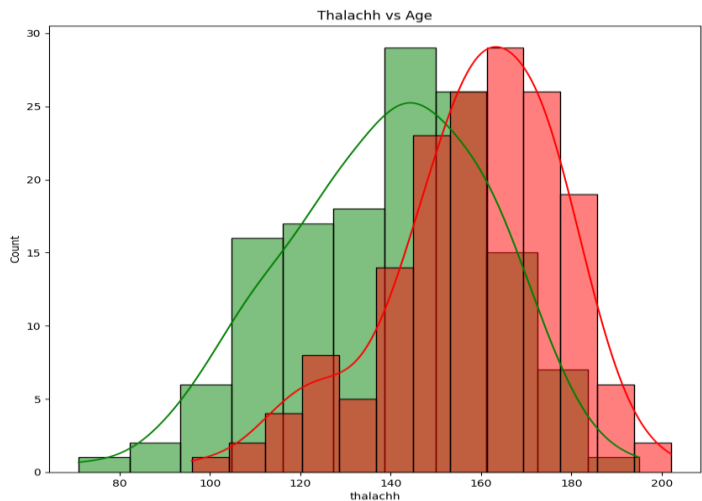


Figure 5: Age vs Thalachh

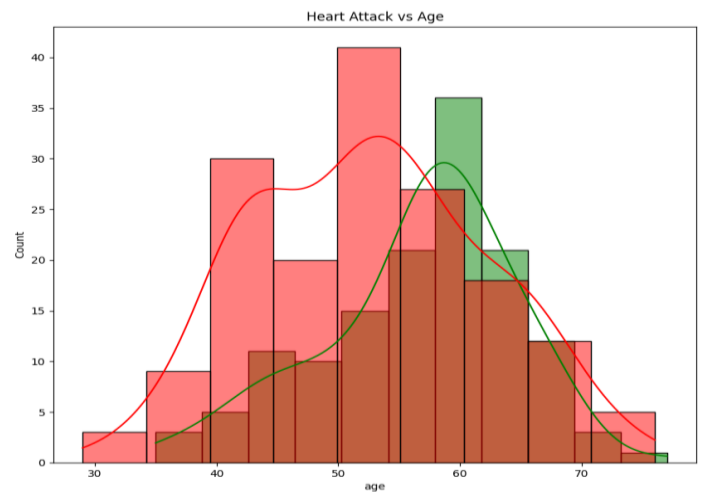


Figure 6: Age vs heart attack rate

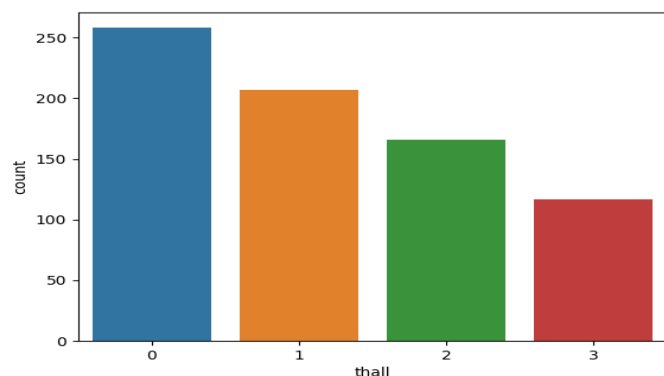


Figure 7: Age vs thall

In second phase of study we will examine correlation and important features from dataset to be considered provide an overview age count of the patient. The figure 8 provides an overview correlation between features. The figure 9 provides rating for important features from dataset.



Figure 8: Correlation matrix

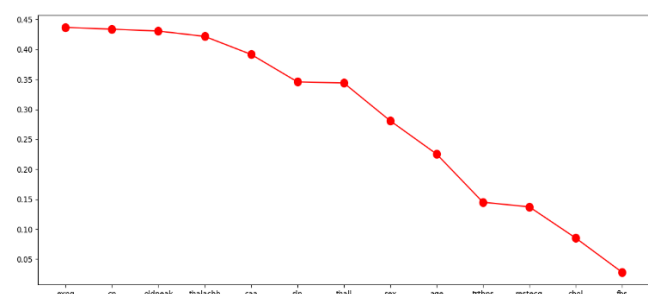


Figure 9 rating of features

Below in figure 10 we have plot a box-cum-swarm plot for the outlier's detection. It is a great way to observe the data and it's stats like **median**, **max**, **min**, and the **quartiles** by just hovering on the plots.

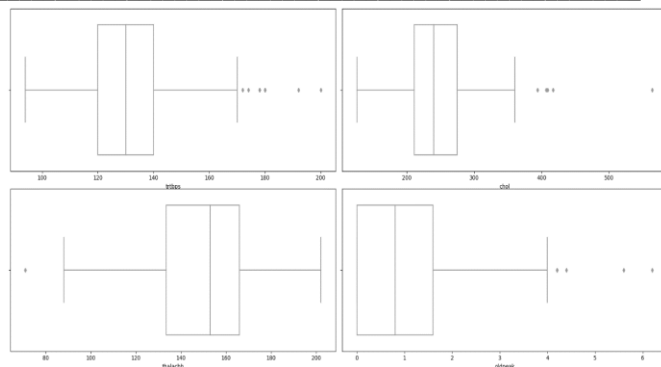


Figure 10: Box plot for outlier's detection.

Figure 11 provides the prediction accuracy to predict the heart attack. The proposed model archived accuracy of 96.77%. In comparison with other machine learning models. XG-boost, Ada Boost, MLP classifier, random forest, gradient boosting, logistic regression, SVM, KNN, decision tree achieved 95.08%, 93.44%, 93.44%, 91.8%, 91.8%, 90.16, 90.16%, 88.52%, 81.97%.

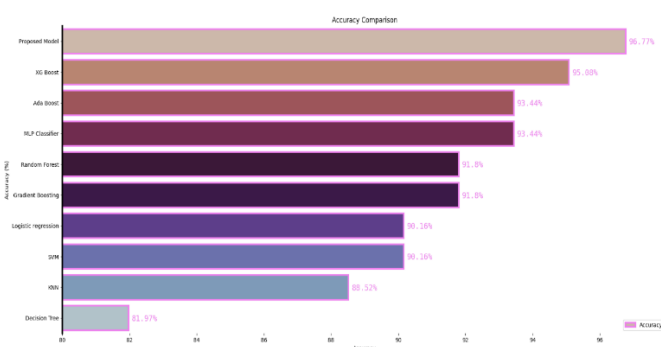


Figure 11: Comparison of existing techniques with proposed model

The forecast has significant clinical implications for analyzing disease risk factors and interpreting a patient's situation. The boosting algorithm demonstrated promising outcomes in predicting heart disease symptoms, indicating potential for further optimization by delving into the associated risk factors for this condition.

## V. Conclusion:

In the present era, a critical focus is essential on cardiac arrest due to statistics indicating a high number of fatalities, recording nearly 617,000 deaths in 2017. Early prediction and preventive measures for heart attacks are crucial in averting premature fatalities. The Extreme Learning Machine algorithm stands out as a highly efficient machine learning technique within the realm of neural networks, particularly adept at handling extensive datasets. Its non-iterative training approach involves tuning all parameters at once, resulting in rapid training. Notably, this algorithm boasts user-friendly implementation and the capability to address intricate

problems. The proposed model archived accuracy of 96.77%. In comparison with other machine learning models. XG-boost, Ada Boost, MLP classifier, random forest, gradient boosting, logistic regression, SVM, KNN, decision tree achieved 95.08%, 93.44%, 93.44%, 91.8%, 91.8%, 90.16, 90.16%, 88.52%, 81.97%. In future same technique can be implemented different medical domain.

## References

- [1] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–16, 2020.
- [2] T. Obasi and M. O. Shafiq, "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases," in *2019 IEEE international conference on big data (big data)*, 2019, pp. 2393–2402.
- [3] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, 2018.
- [4] V. V Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.
- [5] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. Rudd, and M. van der Schaar, "Cardiovascular Disease Risk Prediction Using Machine Learning: A Prospective Cohort Study of 423,604 Participants," *Circulation*, vol. 138, no. Suppl\_1, pp. A15234–A15234, 2018.
- [6] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.
- [7] M. Aljanabi, M. H. Qutqut, and M. Hijjawi, "Machine learning classification techniques for heart disease prediction: a review," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 5373–5379, 2018.
- [8] D. Swain, S. K. Pani, and D. Swain, "A metaphoric investigation on prediction of heart disease using machine learning," in *2018 International conference on advanced computation and telecommunication (ICACAT)*, 2018, pp. 1–6.
- [9] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PloS one*, vol. 12, no. 4, p. e0174944, 2017.
- [10] Y. Khan, U. Qamar, N. Yousaf, and A. Khan, "Machine learning techniques for heart disease datasets: a survey," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 2019, pp. 27–35.
- [11] J. I. Z. Chen and P. Hengjinda, "Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study," *Journal of Artificial Intelligence*, vol. 3, no. 01, pp. 17–33, 2021.
- [12] Bhatt CM, Patel P, Ghetia T, Mazzeo PL. *Effective Heart Disease Prediction Using Machine Learning Techniques*. Algorithms. 2023; 16(2):88. <https://doi.org/10.3390/a16020088>
- [13] Li, F., Wu, P., Ong, H. H., Peterson, J. F., Wei, W. Q., & Zhao, J. (2023). Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *Journal of Biomedical Informatics*, 138, 104294.
- [14] Kresoja, K. P., Unterhuber, M., Wachter, R., Thiele, H., & Lurz, P. (2023). A cardiologist's guide to machine learning in cardiovascular disease prognosis prediction. *Basic research in cardiology*, 118(1), 10.
- [15] Bizimana, P. C., Zhang, Z., Asim, M., El-Latif, A., & Ahmed, A. (2023). An Effective Machine Learning-Based Model for an Early Heart Disease Prediction. *BioMed Research International*, 2023.
- [16] Abdulsalam, G., Meshoul, S., & Shaiba, H. (2023). Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach. *Intell. Autom. Soft Comput*, 36, 761-779.
- [17] Dileep, P., Rao, K. N., Bodapati, P., Gokuruboyina, S., Peddi, R., Grover, A., & Sheetal, A. (2023). An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm. *Neural Computing and Applications*, 35(10), 7253-7266.
- [18] Sk, K. B., Roja, D., Priya, S. S., Dalavi, L., Vellela, S. S., & Reddy, V. (2023, March). Coronary Heart Disease Prediction and Classification using Hybrid Machine Learning Algorithms. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 1-7). IEEE.
- [19] Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2023). An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications*, 35(20), 14723-14737.
- [20] Narayan, Vipul, and A. K. Daniel. "Novel protocol for detection and optimization of overlapping coverage in wireless sensor networks." *Int. J. Eng. Adv. Technol* 8 (2019).
- [21] Narayan, Vipul, A. K. Daniel, and Ashok Kumar Rai. "Energy efficient two tier cluster based protocol for wireless sensor network." *2020 international conference on electrical and electronics engineering (ICE3)*. IEEE, 2020.
- [22] Narayan, Vipul, and A. K. Daniel. "Multi-tier cluster based smart farming using wireless sensor network." *2020 5th international conference on computing, communication and security (ICCCS)*. IEEE, 2020.
- [23] Narayan, Vipul, and A. K. Daniel. "RBCHS: Region-based cluster head selection protocol in wireless sensor network." *Proceedings of Integrated Intelligence Enable Networks and Computing: IINCC 2020*. Springer Singapore, 2020.
- [24] Narayan, Vipul, and A. K. Daniel. "A novel approach for cluster head selection using trust function in WSN." *Scalable Computing: Practice and Experience* 22.1 (2021): 1-13.
- [25] Narayan, Vipul, and A. K. Daniel. "IOT based sensor monitoring system for smart complex and shopping malls." *International conference on mobile networks and management*. Cham: Springer International Publishing, 2021.

- [26] Narayan, Vipul, and A. K. Daniel. "Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model." (2022).
- [27] Mall, Pawan Kumar, et al. "Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique." *Journal of Pharmaceutical Negative Results* (2022): 4784-4792.
- [28] Narayan, Vipul, et al. "FuzzyNet: Medical Image Classification based on GLCM Texture Feature." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.
- [29] Sawhney, Rahul, et al. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." *Decision Analytics Journal* 6 (2023): 100169.

