A Genetic Algorithm based Modified Machine Learning Approach for the Prediction of Heart Disease

Sukanya Roy
Dept. Computer Science Engineering
University of Engineering and Management
Kolkata, India
piu.sep.1993@gmail.com

Bijoy Kumar Mandal
Dept. Computer Science Engineering
NSHM Knowledge Campus
Durgapur, India
writetobijoy@gmail.com

Koustov Khamaru
Dept. Computer Science Engineering
University of Engineering and Management
Kolkata, India
koustov.khamaru@gmail.com

Arghya Bhattacharya
Dept. Computer Science Engineering
University of Engineering and Management
Kolkata, India
arghyabhattacharya4444@gmail.com

Debarghya Basak
Dept. Computer Science Engineering
University of Engineering and Management
Kolkata, India
basakdebarghya01@gmail.com

***Abstract*— This considers employments a novel strategy of information investigation, highlight choice and hyper-parameter tuning to get it and select the highlights predominant in heart illnesses from a therapeutically practical dataset. Customarily, heart infections are measured at afterward stages and they may pass the point of no return for certain patients. The need of early location of heart illnesses is one the prime reason for the outlandish misfortune of blameless lives. The current technique includes the utilization of an altered machine learning calculation alongside Hereditary algorithm-based highlight choice and advancement to detect the highlights that influence the heart and cause maladies the foremost. Components like resting blood pressure, gender, fasting blood pressure, resting electrocardiographic comes about and a few other variables are taken under consideration to attain the finest conceivable comes about. The altered approach accomplishes a part superior precision as compared to conventional machine learning approaches. In this paper, Direct Straightforward Vector Classifier, Ada-Boost Classifier and Calculated Relapse Classifier have been utilized. All the hyper-parameters are finely tuned to progress the ultimate precision.**

Keywords—Genetic Algorithms, Machine Learning, Hyper-parameter tuning, Feature Selection, Optimization, Heart Disease

I. Introduction

Specialized inquire about in heart maladies incorporates a wealthy history, dating back to the 17th century with William Harvey's disclosure of blood circulation. Key breakthrough incorporates the development of the stethoscope in 1819, the disclosure of X-rays in 1895, and the advancement of electrocardiography within the early 20th century. These progressions, together with the presentation of drugs like digitalis and diuretics, [1] altogether made strides the conclusion and treatment of heart conditions. Within the mid-20th century, the development of cardiac catheterization and coronary supply route bypass surgery revolutionized heart care. More as of late, progresses in restorative imaging, negligibly obtrusive strategies, and sedate treatments have kept on make strides results for patients with heart maladies [2].

In cutting edge times, conventional machine learning calculations have been utilized to foresee the chances of a quiet causing a heart illness with direct victory rates. Conventional machine learning calculations like Ada Boost Classification, [3] Calculated Relapse and Straightforward Vector Classification have been utilized as benchmarks to compare the comes about of the more seasoned calculations and the proposed altered calculation [4].

Within the current paper hereditary calculations have been utilized to guarantee maximal include extraction and optimization and miniature hyper-parameter tuning has too been done. The Machine Learning calculations have been altered to work in conjunction with the instrument of hereditary calculation in an exertion to progress upon the already utilized procedures. Correctness's higher than 5-6% as compared to conventional calculations have been accomplished [5].

II. Literature Survey

Latest research studies have shown promising findings with regards to the application of technology in heart disease detection and prediction. Machine learning algorithms can potentially predict heart disease with accuracy by using many parameters from patients.[6][7] These studies use techniques such as random forests, support vector machines, and artificial neural networks to analyse information obtained from patients and identify heart disease-related patterns [8].

Medical imaging, in the form of CT scans and MRI, has further improved the visual and evaluation of the structure and functioning of the heart, hence helping in early detection. Wearable technology, like smartwatches, increasingly monitors heart rate and other vital signs that can be invaluable data for the personal risk assessment and early intervention [9].

Machine Learning has played a pivotal role in advancing the field of heart disease research and heart disease prevention owing to its quick and accurate nature. On the basis of this, we concluded that early detection of heart diseases is pivotal if the long-term risks are to be mitigated[11]. In this current paper, we have made use of Logistic Regression, Ada-Boost Classification and Linear SVC to achieve our goals [12].

Genetic algorithms fall under the category of evolutionary and optimization search techniques that make use of the principle of Darwinian natural selection. These search algorithms work by creating candidate solutions, evaluating them regarding their fitness, and generating a new generation by cross-over and mutation to form candidate solutions from the selected ones within the population. [19][20]][21] The cycle progresses until a satisfactory solution emerges or a termination criterion is satisfied. Genetic algorithms are particularly useful for solving complex optimization problems where traditional methods may struggle [17][18]. Genetic algorithm mechanisms do very well in optimal feature selection and code optimization.[13][14] Therefore, the current research methodology has well combined the mechanisms of both machine learning and genetic algorithms to achieve high rates of accuracy in the detection of heart disease.[15][16]

III. PROPOSED METHODOLOGY

The dataset used is of medical data regarding heart conditions of patients. Every column denotes a specific-criteria that impacts the heart, and every one of them represents the medical data of an individual. The dataset is in ASCII CSV format wherein every row denotes a heart-related medical criteria. Thus, there are 1025 rows and 14 columns in the dataset (1025 x 14).

A. *Traditional Methodology:*

Feature selection and extraction are critical steps in machine learning that enhance the performance of a model and its interpretability. Feature selection is concerned with identifying only the most important information in the data and discarding irrelevant or redundant details. This process streamlines the model, preventing overfitting and enhancing efficiency. Techniques such as correlation analysis, forward selection, and L1 regularization are commonly used. Feature extraction, on the other hand, changes the original data into a new representation that often captures better the underlying patterns. Among them are PCA, LDA, and t-SNE that are used to reduce the dimensionality and form more informative features.

Classification algorithms are one of the key tools in supervised machine learning. They help computers classify data into pre-specified categories, like spam or not spam, or diagnose medical conditions. In the case of the current study, Logistic Regression, Linear SVC, and Ada-Boost are considered.

1) *Logistic Regression*

Logistic regression is used to model a linear combination of one or more independent variables in terms of the probability of an event occurring.

In logistic regression, we model the probability of an event occurring as a linear combination of one or more independent variables. The relationship is expressed through the following equation:

$$log\left(\frac{p}{1-p}\right) = \beta^0 + \beta^1 X^1 + \beta^2 X^2 + \ldots + \beta_k X_k$$

where

- $\boldsymbol{p}$: The probability of the event occurring
- $\boldsymbol{X^1 \ldots X_k}$: The independent variables
- $\boldsymbol{\beta_0 \ldots \beta_k}$:The coefficients (weights) associated with each variable

The left-hand side of the equation, $log\left(\frac{p}{1-p}\right)$, is known as the logit or log-odds. It transforms the probability, which is bounded between 0 and 1, into a linear function that can take on any value.[22]

To obtain the probability directly, we can apply the inverse of the logit function, which is the sigmoid function:

$$p = 1\left(1 + \exp\left(-(\beta^0 + \beta^1 X^1 + \beta^2 X^2 + \ldots + \beta_k X_k)\right)\right)$$

This equation provides a probabilistic prediction for the occurrence of the event based on the values of the independent variables and their corresponding coefficients.

2) *Linear SVC*

Linear SVC is a machine learning algorithm for classification. It tries to find an optimal hyperplane that separates the data points of different classes with the maximum margin. In case of linear SVC, this hyperplane is a line.[23][24]

The decision boundary for a linear SVC can be represented by the following equation:

$$w^T x + b = 0$$

where:

- **w**: Weight vector (normal to the hyperplane)
- **x**: Input feature vector
- **b**: Bias term (determines the offset of the hyperplane)

a) Interpretation

- If $w^T x + b > 0$, the data point is classified as belonging to one class.
- If $w^T x + b < 0$, the data point is classified as belonging to the other class.

3) *AdaBoost (Adaptive Boosting)*

AdaBoost is a popular ensemble learning method that combines multiple weak learners (classifiers that performed slightly better than random guessing) to create a strong learner with significantly improved accuracy.[25][26]. The core equation in AdaBoost represents the final prediction of the ensemble:

$$H(x) = sign\left(\sum\left(\alpha_t * h_{t(x)}\right)\right)$$

where:

- $H(x)$: The final prediction of the ensemble for a given input$x$.
- α_t: The weight assigned to the $\mathrm{t_{th}}$ weak learner.
- $h_{t(x)}$: The prediction of the $\mathrm{t_{th}}$ weak learner for input x.

a) Initialization

Assigning equal weights to all training samples.

b) Iterative Training

For each iteration t: , Training a weak learner h_t using the current weights of the Training samples.

Calculating the error rate (ε_t) of h_t.

Calculating the weight α_t for h_t based on its error rate:

$$\alpha_t = 0.5 * \log\left(\frac{(1 - \varepsilon_t)}{\varepsilon_t}\right)$$

Updating the weights of the training samples based on the performance of h_t:

Increasing the weights of misclassified samples.

Decreasing the weights of correctly classified samples.

c) Final Prediction

Combining the predictions of all weak learners using their respective weights α_t .[27]. The sign of the weighted sum determines the final prediction.

B. Proposed Methodology:

In the proposed methodology, we used finely hyper-parameter tuned machine learning algorithms in conjunction with genetic algorithm mechanisms in an attempt to achieve higher accuracies.[28]

1) Logistic Regression

Genetic algorithms can be applied to optimize the coefficients $\beta_0.\beta_k$ of the logistic regression model. The idea is that every set of coefficients should be represented as a "chromosome" or individual in a population. Then, the fitness of every chromosome is calculated by considering the performance of the logistic regression model corresponding to the given chromosome.[29]

The optimized coefficients obtained from the genetic algorithm are then used to create the final logistic regression model:

$$p = 1\left(1 + exp\left(-(\beta^0 + \beta^1 X^1 + \beta^2 X^2 + \ldots + \beta_k X_k)\right)\right)$$

2) Linear SVC

Genetic algorithms can be used to optimize the weight vector **w** and bias term **b** in the Linear SVC model:

- Chromosome Representation: Each chromosome in the genetic algorithm population represents a set of weights (w) and a bias term (b).
- Fitness Function: The fitness function evaluates the performance of the Linear SVC model with the given weights and bias. Common metrics include accuracy, precision, recall, F1-score, or area under the ROC curve (AUC).[30]
- The optimized weights and bias obtained from the genetic algorithm are then used to create the final Linear SVC model:

$$w^T x + b = 0$$

C. Ada-Boost

$$H(x) = sign\left(\sum\left(\alpha_t * h_{t(x)}\right)\right)$$

Instead of directly calculating weights in the AdaBoost algorithm, genetic algorithms can find an optimal set of weights for a given weak learner in an ensemble. The parameters for every weak learner can be optimized with genetic algorithms.

1) Algorithm

- Step 1: The heart disease dataset is imported.
- Step 2: Data cleaning and pre-processing are performed.
- Step 3: Linear SVC, Logistic Regression and Ada-Boost are applied on the dataset to achieve the base accuracies and the ROC curves are displayed.
- Step 4: The genetic algorithms are applied for feature selection, feature extraction and optimization.
- Step 5: Linear SVC, Logistic Regression and Ada-Boost are applied on each and every generation of optimized features for five different generations.
- Step 6: The improved accuracies are checked and the changes in the generations are displayed graphically.

IV. RESULT AND ANALYSIS

To check the results of the newer methodology, we used mathematical data pertaining to the heart conditions of several patients. We found that the modified machine learning methodology improves the accuracies of machine learning algorithms by a large margin and aids in the early detection of heart diseases. Genetic algorithms ensure that the unwanted features are eliminated and thus, higher accuracies with better results are obtained.

The Genetic Algorithm (GA) optimization of the Ada-Boost Classifier has significantly improved its performance, achieving a remarkable 89.35% accuracy rate which is a major jump from the traditional accuracy of 83.20%.

This substantial enhancement surpasses the capabilities of existing algorithms, demonstrating the effectiveness of the GA in fine-tuning the Ada-Boost classifier parameters for optimal classification outcomes.

A. Procedural Analysis

Fig 1. Shows us the dataset that has been used. Fig 2. Shows us the corelation heatmap of the features present in the dataset.

On comparing Fig 3. with Fig 7,8,9 we can see that the modified machine learning algorithm gives us a higher accuracy as compared to the traditional technique.

On comparing Fig 4,5,6 with Fig 10,11,12 (respectively), we see the graphical differences in accuracies between the traditional technique with the modified machine learning algorithm.

```
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
```

Fig. 1. Dataset Used

Fig. 2. Corelation of Heatmap and features

	Classifier	Accuracy
0	Ada Boost	0.832685
1	Linear SVC	0.789883
2	Logistic Regression	0.782101

Fig. 3. Accuracies of the classification algorithms used without genetic algorithms

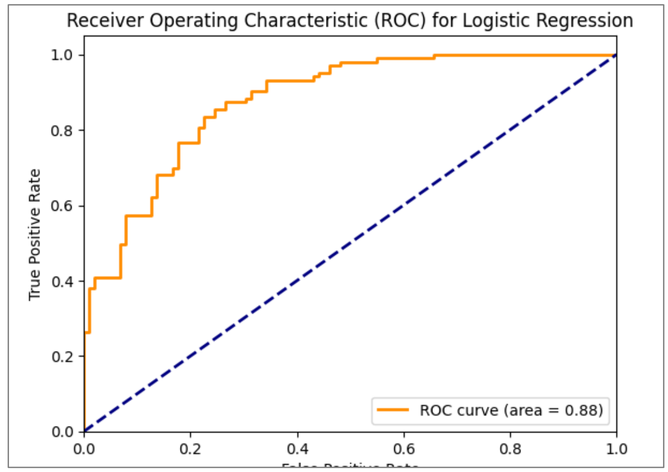

Fig. 4. ROC curve for Logistic Regression

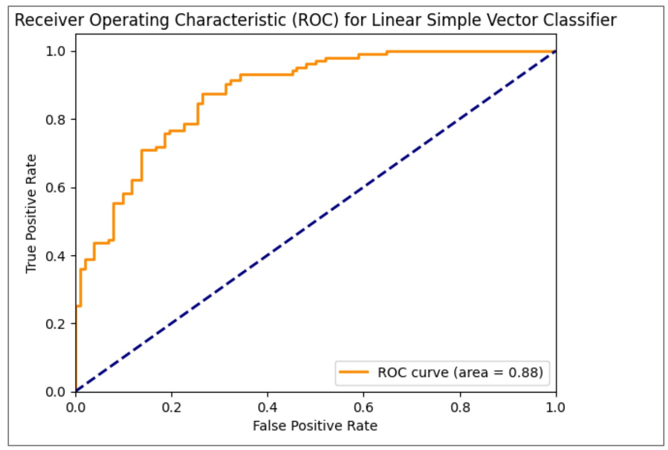

Fig. 5. ROC curve for Linear Simple Vector Classifier

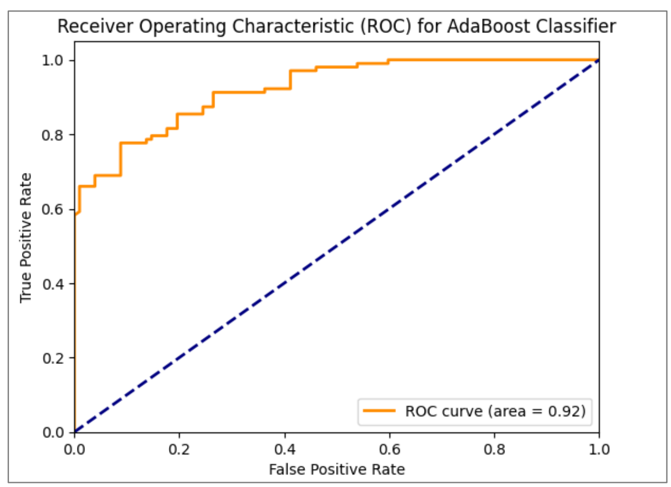

Fig. 6. ROC curve for AdaBoost Classifier

```
Best score in generation 1 : [0.844106463878327]
Best score in generation 2 : [0.8479087452471483]
Best score in generation 3 : [0.844106463878327]
Best score in generation 4 : [0.844106463878327]
Best score in generation 5 : [0.8555133079847909]
```

Fig. 7. Accuracies of Logistic Regression using genetic algorithm for 5 generations

```
Best score in generation 1 : [0.8555133079847909]
Best score in generation 2 : [0.8517110266159695]
Best score in generation 3 : [0.8517110266159695]
Best score in generation 4 : [0.8479087452471483]
Best score in generation 5 : [0.8517110266159695]
```

Fig. 8. Accuracies of Linear SVC using genetic algorithm for 5 generations

```
Best score in generation 1 : [0.8935361216730038]
Best score in generation 2 : [0.8859315589353612]
Best score in generation 3 : [0.8859315589353612]
Best score in generation 4 : [0.8821292775665399]
Best score in generation 5 : [0.8935361216730038]
```

Fig. 9. Accuracies of AdaBoost Classifier using genetic algorithm for 5 generations

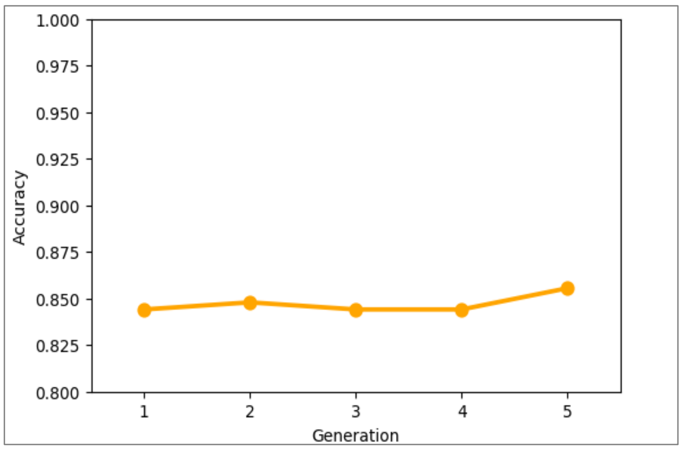

Fig. 10. Accuracy graph of Logistic Regression using genetic algorithm for 5 generations

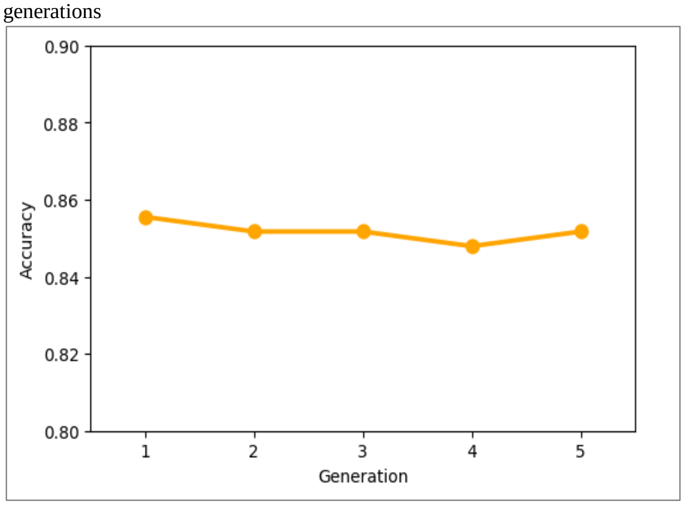

Fig. 11. Accuracy graph of Linear SVC using Genetic algorithm for 5 generations

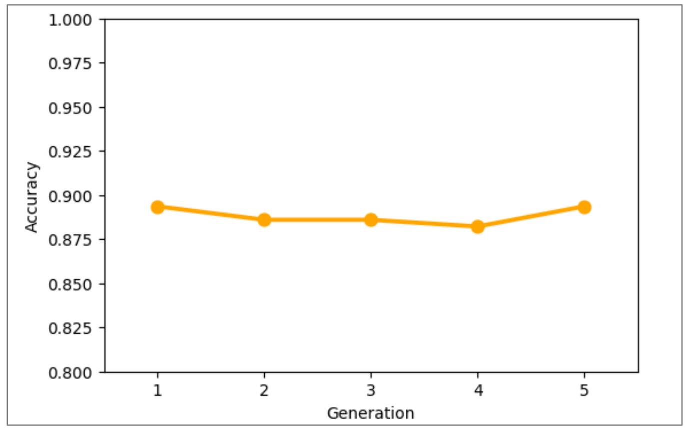

Fig. 12. Accuracy graph of AdaBoost using Genetic algorithm for 5 generations

V. CONCLUSION AND FUTURE SCOPE

We would like to conclude the proceedings of this research by stating our firm belief in the fact that the enhanced accuracy in the early detection and prediction of heart diseases in potential patients will surely help doctors and medical professionals avoid doomsday scenarios and save a lot of lives. There are countless methods of heart disease detection and feature selection but they are not completely optimized and their accuracies are not maximized, as seen via this research.

Thus, a modified methodology where modified and highly hyper-parameterized classification algorithms work in pair with genetic algorithms offer far higher efficiency and accuracy in the early detection of heart diseases in potential patients. Manual intervention by doctors and medical professionals becomes redundant as the algorithm can directly detect the presence of heart diseases from their medical files themselves. The research shall be used a basis for more advanced research into heart disease detection and its prevention with higher accuracy and efficiency. In accordance with the needs of the current times, we also intend to work on ways to predict the exact specific regions of the heart that may have issues and create a framework that will allow potential patients to have personalized reports on possible heart conditions they may have or develop.

REFERENCES

[1] Alshraideh, Mohammad, Najwan Alshraideh, Abedalrahman Alshraideh, Yara Alkayed, Yasmin Al Trabsheh, and Bahaaldeen Alshraideh. "Enhancing heart attack prediction with machine learning: A study at jordan university hospital." Applied Computational Intelligence and Soft Computing 2024, no. 1 (2024): 5080332.

[2] Anjum, Nishat, Cynthia Ummay Siddiqua, Mahfuz Haider, Zannatun Ferdus, Md Azad Hossain Raju, Touhid Imam, and Md Rezwanur Rahman. "Improving Cardiovascular Disease Prediction through Comparative Analysis of Machine Learning Models." Journal of Computer Science and Technology Studies 6, no. 2 (2024): 62-70.

[3] Mandava, Manjula. "MDensNet201-IDRSRNet: Efficient cardiovascular disease prediction system using hybrid deep learning." Biomedical Signal Processing and Control 93 (2024): 106147.

[4] Mishra, Jyoti, and Mahendra Tiwari. "IoT-enabled ECG-based heart disease prediction using three-layer deep learning and meta-heuristic approach." Signal, Image and Video Processing 18, no. 1 (2024): 361-367.

[5] Singh, M. Sheetal, Khelchandra Thongam, Prakash Choudhary, and P. K. Bhagat. "An Integrated Machine Learning Approach for Congestive Heart Failure Prediction." Diagnostics 14, no. 7 (2024): 736.

[6] Naser, Marwah Abdulrazzaq, Aso Ahmed Majeed, Muntadher Alsabah, Taha Raad Al-Shaikhli, and Kawa M. Kaky. "A Review of Machine Learning’s Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges." Algorithms 17, no. 2 (2024): 78.

[7] Ogunpola, Adedayo, Faisal Saeed, Shadi Basurra, Abdullah M. Albarrak, and Sultan Noman Qasem. "Machine learning-based predictive models for detection of cardiovascular diseases." Diagnostics 14, no. 2 (2024): 144

[8] Yashudas, A., Dinesh Gupta, G. C. Prashant, Amit Dua, Dokhyl AlQahtani, and A. Siva Krishna Reddy. "DEEP-CARDIO: Recommendation System for Cardiovascular Disease Prediction using IOT Network." IEEE Sensors Journal (2024).

[9] Feng, Mingyang, Xiaosong Wang, Zhiming Zhao, Chufeng Jiang, Jize Xiong, and Ning Zhang. "Enhanced Heart Attack Prediction Using eXtreme Gradient Boosting." Journal of Theory and Practice of Engineering Science 4, no. 04 (2024): 9-16.

[10] Baseer, K. K., K. Sivakumar, Duggineni Veeraiah, Gunjan Chhabra, Prasanna Kumar Lakineni, M. Jahir Pasha, Ramu Gandikota, and Gopakumar Harikrishnan. "Healthcare diagnostics with an adaptive deep learning model integrated with the Internet of medical Things (IoMT) for predicting heart disease." Biomedical Signal Processing and Control 92 (2024): 105988.

[11] Omankwu, Obinnaya CB, M. C. Okoronkwo, and Chigbundu Kanu. "Cardiovascular (Heart) Diseases Prediction using Deep Learning Neural Network Model." NIPES-Journal of Science and Technology Research 6, no. 1 (2024).

[12] Krishna, Anala Hari, AravapalliBhavya Sri, KurakulaYuvaVenkata Sai Priyanka, SachinTaran, and Varun Bajaj. "Emotion classification using EEG signals based on tunable-Q wavelet transform." IET Science, Measurement & Technology 13, no. 3 (2019): 375-380.

[13] Ma, Cai-Yi, Ya-Mei Luo, Tian-Yu Zhang, Yu-Duo Hao, Xue-Qin Xie, Xiao-Wei Liu, Xiao-Lei Ren et al. "Predicting coronary heart disease in Chinese diabetics using machine learning." Computers in Biology and Medicine 169 (2024): 107952.

[14]]. Jasmine, Jinitha A., and S. Mangayarkarasi. "Heart Disease Prediction and Analysis Using Ensemble Classifier in Machine Learning Techniques." In Human Machine Interaction in the Digital Era, pp. 99-105. CRC Press, 2024.

[15] Roy, Sukanya, et al. "Predictive Analysis & Classification of Diseases in Organs using Modified-C Clustering Technique." 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom).IEEE, 2020.

[16] Roy, Sukanya, et al. "Identification of medical disorders in eye and biometric authentication analysis with iris retina scan using machine learning." Biotechnology and Biological Sciences.CRC Press, 2019.29-33.

[17] Alarcao, Soraia M., and Manuel J. Fonseca. "Emotions recognition using EEG signals: A survey." IEEE Transactions on Affective Computing 10.3 (2017): 374-393.

[18] Siuly, Siuly, Yan Li, and Yanchun Zhang. "EEG signal analysis and classification." IEEE Trans Neural SystRehabilitEng 11 (2016): 141-4.

[19] Zhao, Guozhen, et al. "Emotion analysis for personality inference from EEG signals." IEEE transactions on affective computing 9.3 (2017): 362-371.

[20] Roy, Sukanya, SoumyaBhattacharjee, Bijoy Kumar Mandal, and MrigankaSaha. "A Comparative Study on Different Techniques for Classification of Brain Waves From EEG Signals." American Journal of Science & Engineering 2, no. 4 (2022): 1-7.

[21] Roy, Sukanya, et al. "Brain Anomaly Prediction Using Modified Clustering (MC)." 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2019.

[22] Roy, Sukanya, et al. "Tumor detection using spatial hybrid clustering (SHCM) from brain MRI scans." Contemporary Medical Biotechnology Research for Human Health. Academic Press, 2022. 149-159.

[23] Roy, Sukanya, Bijoy Kumar Mandal, Koustov Khamaru, and Arghya Bhattacharya. "Prediction and Classification of Brain Tumor using Modified Transfer Learning." In 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 850-854. IEEE, 2024.

[24] Roy, Sukanya, Nilesh Dutta, Siddhartha Ghosh, Samapti Kundu, and Tama Rani Saha. "Brain tissue segmentation using modified fuzzy C-means clustering." In Contemporary Medical Biotechnology Research for Human Health, pp. 141-149. Academic Press, 2022.

[25] Rani, Pooja, Rajneesh Kumar, Anurag Jain, Rohit Lamba, Ravi Kumar Sachdeva, Karan Kumar, and Manoj Kumar. "An Extensive Review of Machine Learning and Deep Learning Techniques on Heart Disease Classification and Prediction." Archives of Computational Methods in Engineering (2024): 1-19.

[26] Tigga, Onima, Jaya Pal, and Debjani Mustafi. "Bi-SMOTE: a novel framework for handling imbalanced datasets using machine learning techniques." International Journal of Information Technology (2024): 1-15.

[27] Pingle, Yogesh Prabhakar, and Lakshmappa K. Ragha. "Statistical assessment of Indian classical music’s role in diabetes, thyroid, hypertension management." International Journal of Information Technology 17, no. 1 (2025): 423-429.

[28] Maheswari, G., and S. Gopalakrishnan. "A smart multimodal framework based on squeeze excitation capsule network (SECNet) model for disease diagnosis using dissimilar medical images." International Journal of Information Technology (2024): 1-19.

[29] KN, Sunil Kumar, Pavan P. Kashyap, Darshan A. Bhyratae, Suhas A. Bhyratae, and A. Kalaivani. "CNN-BO-LSTM: an ensemble framework for prognosis of liver cancer." International Journal of Information Technology (2024): 1-7.

[30] Bagheri Tofighi, Abolfazl, Abbas Ahmadi, and Hadi Mosadegh. "Improving lung cancer detection via MobileNetV2 and stacked-GRU with explainable AI." International Journal of Information Technology 17, no. 2 (2025): 1189-1196.