IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

Heart Disease Prediction using Decision Tree

Ms. B. Ranjitha[1], Ms. K. Someshwari[2], Ms. N. Ishwarya[3], Ms. Md. Nazma[4]
Assistant Professor, Department of CSE[1]
Students, Department of CSE[2,3,4]
Guru Nanak Institute of Technology, Hyderabad, India

Abstract: *Heart disease is one of the most common causes of death around the world nowadays. Often, the enormous amount of information is gathered to detect diseases in medical science. All of the information is not useful but vital in taking the correct decision. Thus, it is not always easy to detect the heart disease because it requires skilled knowledge or experiences about heart failure symptoms for an early prediction. Most of the medical dataset are dispersed, widespread and assorted. However, data mining is a robust technique for extracting invisible, predictive and actionable information from the extensive databases. In this paper, by using info gain feature selection technique and removing unnecessary features, different classification techniques such that KNN, Decision Tree (ID3), Gaussian Naïve Bayes, Logistic Regression and Random Forest are used on heart disease dataset for better prediction. Different performance measurement factors such as accuracy, ROC curve, precision, recall, sensitivity, specificity, and F1-score are considered to determine the performance of the classification techniques. Among them, Logistic Regression performed better, and the classification accuracy is 92.76%.*

Keywords: Heart disease

I. INTRODUCTION

Heart disease is one of the most common causes of death around the world nowadays. Often, the enormous amount of information is gathered to detect diseases in medical science. All of the information is not useful but vital in taking the correct decision. Thus, it is not always easy to detect the heart disease because it requires skilled knowledge or experiences about heart failure symptoms for an early prediction. Most of the medical dataset are dispersed, widespread and assorted. However, data mining is a robust technique for extracting invisible, predictive and actionable information from the extensive databases. In this paper, by using info gain feature selection technique and removing unnecessary features, different classification techniques such that KNN, Decision Tree (ID3), Gaussian Naïve Bayes, Logistic Regression and Random Forest are used on heart disease dataset for better prediction.

The practice of examining large preexisting data bases in order to generate new information. It coverts raw data into useful information. It analyze the data for relationships that have not previously been discovered.The steps of data mining are: Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. Medical data mining is a domain of lot of imprecision and uncertainty. The clinical decisions are usually based on the doctors intuition. Therefore this may lead to disastrous consequences. Due to this there are many errors in the clinical decisions and it results in excessive medical costs.Serialization is also used in this system. It converts the data objects into streams of bytes and stores it into database.

Heart disease affects millions of people, and it remains the chief cause of death in the world. Medical diagnosis should be proficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Data mining is a software technology that helps computers to build and classify various attributes. This research paper uses classification techniques to predict heart disease. This section gives a portrayal of the related subjects like machine learning and its methods with brief descriptions, data pre-processing, evaluation measurements and description of the dataset used in this research.

We use Decision trees as it is

- Interpretability: Clinicians and healthcare professionals can easily follow the tree structure to understand how predictions are made.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

- Flexibility: Decision trees can handle both continuous (e.g., cholesterol levels) and categorical (e.g., family history of heart disease) variables.
- Easy to Understand: Decision trees mimic human decision-making by breaking down complex problems into smaller, more manageable decisions

II. REVIEW OF LITERATURE

1. Michael D. Seckeler, Tracey R. Hoke (2017)The Worldwide Epidemiology of Acute Rheumatic Fever(ARF) and Rheumatic Heart Disease (RHD) remain major global public health concerns, particularly in developing nations. This review explores the history, pathology, treatment of ARF, and current worldwide incidence of ARF alongside the prevalence of RHD. Despite a decreasing incidence, the disease burden persists significantly, necessitating continued focus on prevention and management strategies.
2. Thomas A. Gaziano, AsafBitton(2017)Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countriesis the leading cause of death in developed countries and a growing burden in developing nations, accounting for 7.3 million deaths globally in 2001. With three-fourths of these deaths occurring in low- and middle-income countries, the rapid rise in CHD is attributed to socioeconomic changes, aging populations, and lifestyle-related risk factors. Variations in incidence, prevalence, and mortality rates are influenced by risk factor levels, resource availability, and epidemiological transitions, with significant economic impacts. However, effective strategies exist to mitigate this burden.
3. Stephen F. Weng(2017) According to Stephen F. Weng Traditional methods for cardiovascular risk prediction often fail to identify individuals needing preventive treatments while subjecting others to unnecessary interventions. Machine learning offers a transformative approach by analyzing complex interactions between risk factors to improve prediction accuracy. This study evaluates the potential of machine learning in enhancing cardiovascular risk prediction.
4. V. V. Ramalingam(2018)Cardiovascular diseases (CVDs) are among the leading causes of death globally. To address this, machine learning techniques are increasingly being applied to automate analysis and improve diagnostic accuracy for heart diseases. This survey evaluates various supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF), and ensemble models, highlighting their effectiveness and potential in aiding healthcare professionals

III. EXISTING SYSTEM

The amount of data in the medical industry is increasing day by day. It is a challenging task to handle a large amount of data and extracting productive information for effective decision making. For this reason, medical industry demands to apply a special technique which will provide fruitful decision from a vast database. Data mining is an exciting field of machine learning and thus capable of solving this type of problem very well. For solving various kinds of real-world problems, data mining is a novel field for discovering hidden patterns and the valuable knowledge from a large dataset. Because it is very strenuous to extract any useful information without mining large database. In brief, it is an essential procedure for analyzing data from various perspectives and gathering knowledge. However, heath care industry is another field where a substantial amount of data is collected using different clinical reports and patients manifestations.
Heart disease remains one of the most serious health issues of our day. It is said to be the primary motive in death globally. Many times it's difficult for medical professionals to expect a heart disease on time. Nowadays, the health sector contains a lot of precious hidden facts & information which could prove to be very helpful in making predictive decisions especially in the field of medicine.

Disadvantages

- Information Misuse: Sensitive patient data can be vulnerable to breaches or misuse due to inadequate security measures in existing systems. Privacy concerns arise when personal health information is shared without proper consent or stored in unsecure databases.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

- Accuracy Limitations: Many systems rely on outdated or simplistic algorithms, which may lead to false positives or negatives. Inadequate handling of complex interactions between multiple risk factors can result in inaccurate predictions.

IV. PROPOSED SYSTEM

Nowadays, people can face any heart failure symptoms at any stage of a lifetime. But old people face this type of problem rather than the young people. Different classification techniques can discover the hidden relationship along correlated features which plays a consequential role in predicting the class label from a large dataset. By using those hidden patterns along with the correlated features. Then, it will act as an expert system for separating patients with heart disease and patients with no heart disease more accurately with lower cost and less diagnosis time.A decision tree is a versatile machine learning model used for both classification and regression tasks. It is a tree-like structure where each internal node represents a test or decision on an attribute, each branch represents the outcome of that decision, and each leaf node represents a class label or a regression value.

Advantages

- Easy and Fast to Predict: Modern systems provide quick predictions, which is crucial for timely diagnosis and treatment. Automated systems reduce the time required for manual analysis, enabling healthcare providers to focus on patient care.
- Simplicity and Interpretability: Many machine-learning models, especially rule-based systems like Decision Trees, offer clear and interpretable results that healthcare professionals can easily understand and act upon. User-friendly interfaces make these systems accessible even to non-technical users.

V. SYSTEM ARCHITECTURE

The system architecture for heart disease prediction using a decision tree typically consists of several key layers working in sequence to ensure efficient data processing, prediction accuracy, and user interaction. The process begins with data collection from various sources such as patient medical records, wearable devices, or public health databases. This data undergoes preprocessing, including cleaning, normalization, and feature selection, to extract relevant attributes such as age, blood pressure, cholesterol levels, and other risk factors for heart disease.

Once preprocessed, the data is fed into a decision tree algorithm. The decision tree acts as the core predictive model, splitting the data into branches based on feature thresholds to classify whether a patient is at risk of heart disease. The trained model, which is built using historical datasets, evaluates these features and provides an interpretable output. The system incorporates a user-friendly application layer where healthcare professionals or patients can input new data, view predictions, and access visualizations of decision paths. Security measures ensure data privacy and compliance with regulations, while integration with electronic health records (EHR) and wearable devices enables real-time data updates. The system concludes with an output layer that provides risk assessments and actionable insights, such as recommendations for lifestyle changes or further medical testing.

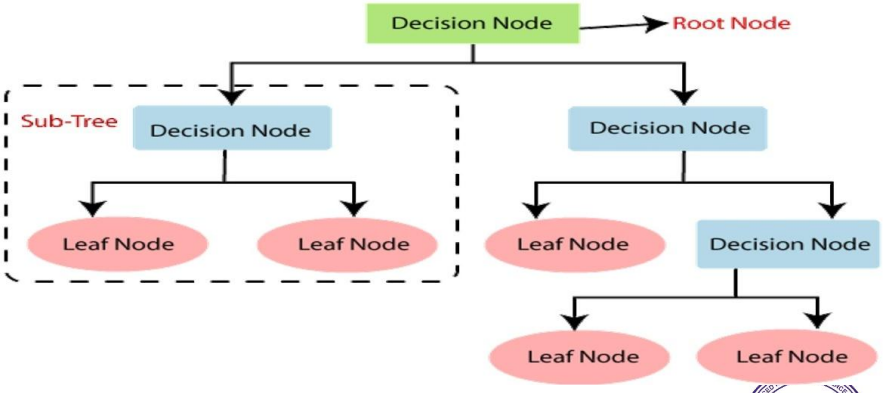

Figure A-SYSTEM ARCHITECTURE

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

VI. METHODOLOGY

The methodology for heart disease prediction using a decision tree begins by defining the objective of identifying patients at risk based on medical and demographic data. Data is collected from sources like electronic health records or public datasets, including attributes such as age, cholesterol levels, and blood pressure. Preprocessing involves cleaning, normalizing, and encoding data while selecting relevant features for analysis. A decision tree model, such as CART or ID3, is then developed by training it on a dataset, optimizing splits based on criteria like information gain or the Gini index. Pruning techniques are applied to prevent overfitting, ensuring the model generalizes well. The trained model is validated using metrics like accuracy, precision, and ROC-AUC. Once validated, the system is deployed through user-friendly interfaces, allowing healthcare providers or patients to input data and receive predictions. The output includes interpretable risk assessments and actionable insights, with ongoing feedback used to refine the model and maintain its accuracy over time.

MODULES

1. Dataset

The Heart Disease dataset is a well-known dataset in the field of machine learning and medical research. It is often used for developing and evaluating models for predicting the presence of heart disease based on various clinical and demographic features.

2. Importing necessary libraries

The code imports libraries required for data manipulation (pandas, numpy), visualization (matplotlib, seaborn), and machine learning (scikit-learn). The dataset is loaded into a pandas DataFrame from a CSV file named 'heart_disease.csv'. Initial exploration of the dataset includes displaying the first few rows, summary statistics, and checking for missing values.

3. Analyzing

To analyze the Heart Disease dataset comprehensively, we will follow a systematic approach that includes Exploratory Data Analysis (EDA), data preprocessing, and model building. This comprehensive approach helps in understanding the dataset, preparing it for analysis, and building models to predict heart disease effectively.

4. Preprocessing

Preprocessing the dataset is a crucial step before building machine learning models. It involves handling missing values, encoding categorical variables, and scaling numerical features. This preprocessing pipeline ensures that the dataset is clean, well-structured, and ready for building machine learning models.

5. Split the data

Splitting the data is a crucial step in machine learning to ensure that we can evaluate the performance of our models on unseen data. Typically, we split the dataset into training and testing sets. This process ensures that we have separate datasets for training and testing, which is essential for evaluating the performance and generalization of machine learning models. Adjust the preprocessing steps according to your specific dataset and modeling needs.

6. Model

Building a model for the Heart Disease dataset involves selecting appropriate algorithms, training the model on the training data, and evaluating its performance on the testing data. Here, I'll demonstrate how to build and evaluate a Logistic model for predicting heart disease presence based on the dataset.

7. Prediction

To make predictions using a trained machine learning model, such as the model we trained earlier on the Heart Disease data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

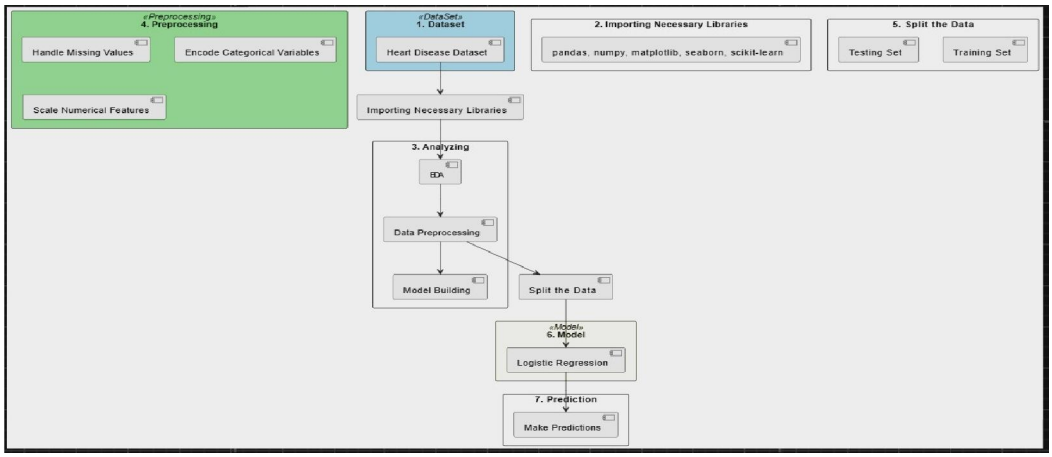

Figure B:-MODULE DIAGRAM

VII. IMPLEMENTATION

Algorithm:-

1. Import all the necessary libraries for the web application, data processing, and machine learning.
2. Create a Flask app instance.
3. Load the pre-trained decision tree model using pickle.
4. Set up routes to serve different pages of the web application like Home route, Login route, Chart route, Upload route, Performance route.
5. Allow users to upload a dataset and preview its content.
6. Create a route for rendering the prediction page.
7. Enter all the data like age, cholesterol, Chest pain type, Fasting blood pressure, Maximum heart rate whether it is up sloping, down sloping or flat sloping
8. Use the model to predict heart disease based on user input.
9. If the prediction according to the user input given is Negative then patient is not suffering from any kind of Heart disease.
10. If the prediction according to the user input given is Positive then patient is suffering from kind of Heart disease.
1. 11.We can check the performance analysis of overall patients who are suffering from heart diseases using the Precision and recall, Confusion Matrix and Pie Chart

VIII. EXPERIMENTAL RESULTS

LOGIN PAGE: Enter the User name and Password in the login page and click on login

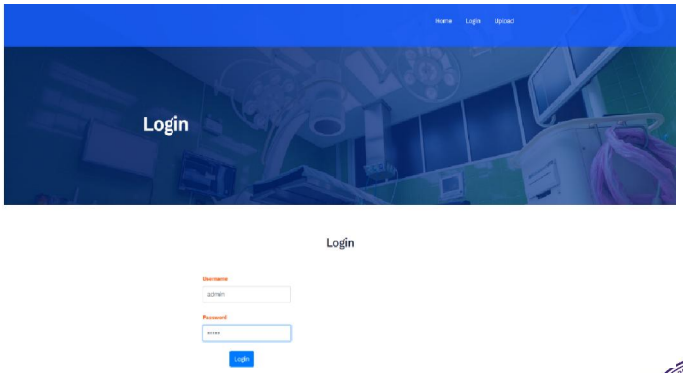

FIGURE:-Login page

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

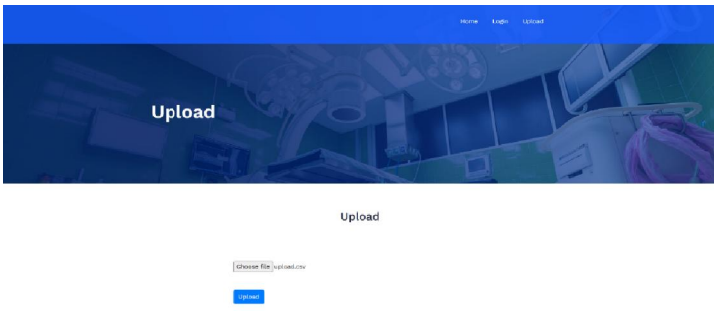

FIGURE:-Upload Page

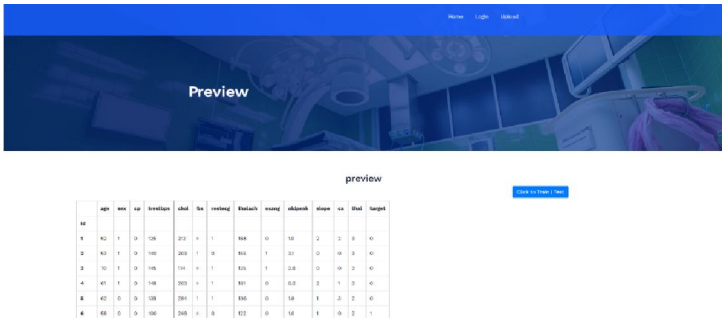

FIGURE:-Preview Page

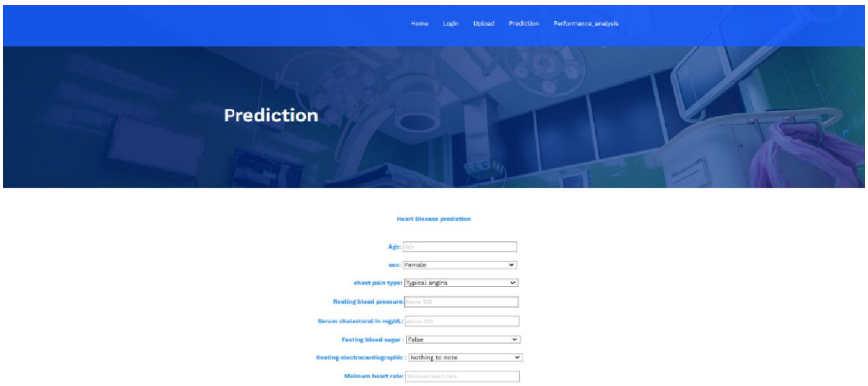

FIGURE:-prediction page

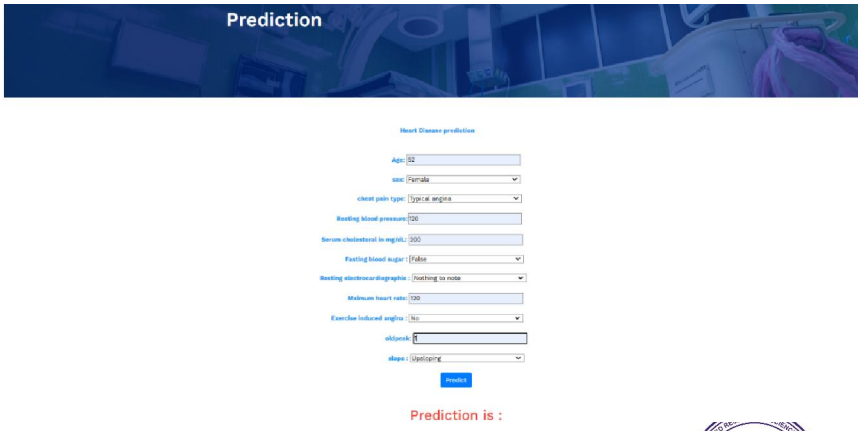

FIGURE:-PATIENTS DATA IN PREDICTION PAGE

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

FIGURE:-PERFORMANCE ANALYSIS

IX. CONCLUSION

As heart disease patients are increasing every year, huge amount of medical data is available. Researchers are applying data mining techniques on this data to diagnosis heart disease. It is analyzed that artificial neural network algorithm is best for classification of knowledge data from large amount of medical data. Population is growing in exponential way. Death rate due to cardiovascular diseases is also increasing. The only solution to control this is to predict the heart disease and medicate it before it gone worse. Our hybrid approach gives higher accuracy rate of 97% of disease detection than earlier proposed method.

X. FUTURE ENHANCEMENT

This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. Integrate real-time monitoring systems with IoT devices like smartwatches, fitness trackers, or ECG monitors to continuously gather data for prediction and early warnings.

REFERENCES

[1] Dewan, A., & Sharma, M. (2015, March). Prediction of heart disease using a hybrid technique in data mining classification. In Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on (pp. 704-706). IEEE.
[2] Dbritto, Rovina, AnuradhaSrinivasaRaghavan, and Vincy Joseph. "Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods." International Journal of Applied Information Systems 11.2 (2016): 22-25.
[3] Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on (pp. 173-177). IEEE.
[4] Lakshmi, K. R., Krishna, M. V., & Kumar, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. International Journal of Scientific and Research Publications, 3(6), 1-10.
[5] Kahramanli, H., &Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1), 82-89.
[6] Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. International journal on recent and innovation trends in computing and communication, 2(10), 3003-3008.
[7] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, 85-94.
[8] Dr.Mohanraj, SubhaSuryaa, Sudha, Sarath Kumar, “Heart Disease Prediction using K Nearest Neighbour and K Means Clustering”, International Journal of Advanced Engineering Research and Science (IJAERS) 2016.
[9] Shinde, R., Arjun, S., Patil, P., &Waghmare, J. (2015). An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm. IJCSIT) International Journal of Computer Science and Information Technologies, 6(1), 637-639.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, December 2024

[10] Kaya, Y., &Pehlivan, H. (2015, November). Feature selection using genetic algorithms for premature ventricular contraction classification. In Electrical and Electronics Engineering (ELECO), 2015 9th International Conference on (pp. 1229-1232). IEEE.

[11] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on. IEEE, 2008.

[12] Boutayeb, A., &Boutayeb, S. (2005). The burden of non communicable diseases in developing countries. Internationaljournal for equity in health.

[13] Jahnavi, Y., Kumar, P. N., Anusha, P., & Prasad, M. S. (2022, November). Prediction and Evaluation of Cancer Using Machine Learning Techniques. In International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology (pp. 399-405). Singapore: Springer Nature Singapore.

[14] Saranya, S. S., Anusha, P., Chandragandhi, S., Kishore, O. K., Kumar, N. P., & Srihari, K. (2024). Enhanced decision-making in healthcare cloud-edge networks using deep reinforcement and lion optimization algorithm. Biomedical Signal Processing and Control, 92, 105963.

[15] Singh, A., Kumar, T. C. A., Mithun, T., Majji, S., Rajesh, M., & Anusha, P. (2021, December). Image Processing Approaches for Oral Cancer Detection in Color Images. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 817-821). IEEE.