Unsupervised Learning for Heart Disease Prediction: Clustering-Based Approach

Janani. Jetty[1], Sajida Sultana. Sk[1*,**], *Ranga Bhavitha*. Polepalle[2], *Vishwitha*. Parusu[3]

[1,2,3]Vignan's Foundation for Science, Technology and Research Vadlamudi, Guntur, Andhra Pradesh, India.

[1*]Assistant Professor, Department of Computer Science and Engineering

[**]Email: sajidashaik550@gmail.com

Abstract. This paper on the prediction of heart disease addresses the application of unsupervised machine learning algorithms, digs up the latent pattern of risk in the data of patients for early diagnosis, and intervenes. We have compared models K-Means Clustering, DBSCAN, Agglomerative Clustering, Gaussian Mixture Model, and Spectral Clustering, wherein K-Means brought out the best result that happened to be 84 percent with the groups formed for patients using nuanced risk indicators. For such insights, the project embeds an HTML web-based interface where healthcare professionals and patients alike can easily read predictions. This approach advances predictive accuracy, yet brings to the medical profession an incredibly powerful tool for a more personalized type of care. Providers would then have the ability to identify ahead of time high-risk people and monitor their care more carefully. It, however, opens up the possibility of unsupervised learning in health analytics and shows how this can be applied to the role of machine learning for early detection and targeted treatment, thereby contributing to better patient outcomes and proactivity in managing heart disease risks.

1. Introduction

Heart disease is one of the major causes of death worldwide, and early detection and prevention are crucial for improving patient outcomes. Traditionally, the diagnosis of heart disease is based on supervised learning models that need labeled data and prior knowledge of risk factors. On the other hand, unsupervised machine learning presents an interesting way in which it automatically discovers the hidden patterns in the patient data that may help identify some previously unknown risk factors without having to have examples labeled. This paper looks at how unsupervised algorithms can be applied in patient data analysis for heart disease prediction: K-Means Clustering, DBSCAN, Agglomerative Clustering, Gaussian Mixture Model (GMM), and Spectral Clustering.

Model comparison reveals that the K-Means Clustering technique was the most efficient

with 84 percent accuracy on the classification of patients on the basis of very subtle risk indicators. Such findings could be beneficial for health professionals in getting a clearer profile of their patients and thus predict those who would have a greater risk. In order to make these predictions accessible to healthcare providers and their patients, we have developed a web-based HTML interface which they can use easily in understanding the results and make in- formed decisions. This is particularly important in health analytics: it demonstrates how unsupervised learning can assist with early detection and personalize the care strategy. Overall, this work points toward a future where unsupervised machine learning can help strengthen preventive healthcare, provide proactive approaches to managing heart disease risk, and open doors for similar applications in other domains of health.

2. Literature Review

Murthy et al. [1] suggested an in-silicon approach to predict the heart disease by unsupervised method analyzing the health factors with K-means clustering and K-Fold Cross Validation approach achieving 82.49. Using machine learning-based unsupervised cluster analysis.

Segar et al. [2] identified three phenogroups in heart failure patients with preserved ejection fraction (HFpEF), each with distinct clinical profiles and outcomes. Such a pho- to mapping approach highlights the heterogeneity of HFpEF and that "personalized treatments" might improve the prognosis of patients by targeting the subgroup-specific characteristics.

Nouraei et al. [3] applied unsupervised machine learning methods: Hierarchical clustering, Kprototype, and partitioning around medoids (PAM) in classifying HFpEF patients into phenogroups. PAM was only the technique that outperformed and identified six phenogroups di The classifications also differ in their clinical characteristics and outcomes. Segar et al. and Kao et al. alternatively applied other clustering methods without comparison across methods.

Bhowmick et al. [4] studied on how the machine algorithms like decision tree (DT), random forest (RF), and logistic regression (LR) might predict heart disease (HD) in 2022. They suggested that the maximum accuracy was 94.7 percent by DT compared to formerly developed models like Bhunia et al. and Anbuselvan, which were 83.87 percent and 86.89 percent, respectively. Further- more, through the usage of DT, the level of precision in the diagnosis of heart disease becomes improved.

Jindal et al. [5] developed a machine learning-based system to predict heart disease with three algorithms named KNN, Logistic Regression, and Random Forest, with accuracy as high as 88.5 percent. In that regard, this model offers the efficient diagnosis of patient conditions at a low cost, which may eventually enhance early detection and medical decision-making for patients at cardiovascular risk.

A Naïve Bayes-based heart disease prediction model with the range of ac- curacy 71-73 percent was developed by Sabri et al. [6]. The proposed system is also made up for early diagnosis and focused on remote regions where health- care facilities remain limited.

Through this web-based application with patient's key data, accessibility to heart disease prediction will now be done considering that there is a shortage of resources for diagnostics and practicing clinicians in resource-constrained environments.

Performance of various supervised learning algorithms for heart disease pre- diction is analyzed by Upadhyay et al. [7]. They focused mainly on K-Nearest Neighbor, Logistic Regression, and Support Vector Machine algorithms. The authors concluded that for heart disease prediction, Logistic Regression obtains the highest accuracy with 87 percent and therefore must be the best model to use. This paper drives a great emphasis on model selection but even more importantly on performance metrics for sensitive healthcare applications such as this one.

Using ten machine learning algorithms that include Random Forest and Extra Trees, Panda et al. [8] was able to develop a heart disease prediction model. One of the good proofs that their study has very high predictive accuracy and robustness of the models is when the medical datasets are classified correctly.

Singh and Kumar [9] evaluated heart disease prediction models using ma- chine learning techniques, including K-Nearest neighbor, Decision Tree, Linear Regression, and SVM. Their results indicate that KNN achieved the highest ac- curacy, with 87 percent, using UCI dataset attributes.

AbdElminaam et al. [10] had suggested a heart disease prediction model, testing six machine learning algorithms. Logistic Regression attained 91.6 percent accuracy whereas Random Forest obtained 98.6 percent on a smaller dataset, thereby highlighting the potential of machine learning in early diagnosis of dis- eases.

Vayadande et al. [11] have developed a machine and deep learning model on basis of predictive models for heart disease with 88.52 percent accuracy in Logistic Regression and Random Forest. This will prove that machine learning can help in the early diagnosis and preventive care for diseases.

Nanehkaran et al. [12] proposed an anomaly detection model of heart disease with a density-based clustering method called DBSCAN using adaptive parameters. Their proposed approach attained 95 percent accuracy, thereby proving the usability of unsupervised methods in the identification of patterns and anomalies of heart diseases.

Bizimana et al. [13] derived those including clustering. Among them was the Gaussian Mixture Models whose outcome in predicting heart disease has its ac- curacy level at 59.51 percent. Utility of Clustering: Clustering can be useful in managing complex medical data, improve the strength of prediction, and assists prevention over the outbreak of heart diseases as described in the study.

Nanehkaran et al. [12] developed an anomaly detection model that identified heart disease through an adaptive parameter density-based clustering method called DBSCAN. The accuracy of the proposed approach was at 95 percent. Such unsupervised methods with

such accuracy can be capable of identifying patterns and anomalies related to heart disease.

Ogunpola et al. [15] research successfully used models like XGBoost, Random Forest, and CNN to predict a 98.50 percent accuracy, and it recommends that the model dealing with an imbalanced dataset would excellently serve. The advance under research for XGBoost in diagnosing heart disease for marked surging precision.

3. Methodology

The proposed methodology will incorporate the following major steps:

3.1. Data Collection

Data gathering forms a very initial part in our approach. We had applied a public data source, which is some form of health information and has, therefore, been applicable to predict heart disease in that study. The features taken up in the data included are age, sex, blood pressure, cholesterol, among other clinical measurements. Import the dataset using a CSV file for easy manipulation and analysis of data. The data should be clean and properly formatted, indicating that no missing attributes are required in the research.

3.2. Data Preprocessing

It represents the last stage in dataset preparation. While considering preprocessing, the following came in view:

3.2.1 Handling missing values:

It is an important preprocessing step in data preparation. In this implementation, we used a Simple Imputer for missing- value handling for numerical features. Here, missing entries are replaced with the average value of the feature concerned. It keeps the general distribution of data but loses the least information. All the categorical features remained as is since most of them need a different strategy on how to handle, like mode imputation or even something more complex, like using predictive modeling techniques to adequately fill missing data. By doing this proper missing-data handling, the integrity of the dataset will be much improved and therefore, good training and performance upon further analysis of the model.

3.2.2 Feature Encoding:

Use one-hot encoding to transform the categorical features into a form that can be used with algorithms. This creates binary columns for every category in categorical features.

3.2.3 Feature Scaling:

StandardScaler scales the encoded numerical features, after which standardizing results in features that have zero mean and unit standard deviation. For most machine learning algorithms to run efficiently, features must have zero mean and unit standard deviation.

3.2.4 Data Splitting:

The dataset was split using Stratified K-Fold cross-validation in order to keep the target variable's distribution spread over the folds. That is, the model should be able to assess it more robustly.

3.3 Data Clustering

This K-Means clustering algorithm had classified heart disease cases while split- ting the data into two sets, training and testing, to ensure that this evaluation was robust. For the application of the model, two clusters were used such that k=2 would classify heart disease cases versus non-disease cases and was initialized with 'k-means++' to set the centroids optimally. It had stability above 50 independent runs but ran up to a maximum of 500 iterations for the algorithm. In this preprocessing part which includes treating missing values, encoding of categorical variables, and also feature scaling, the average accuracy produced by the model remains at 84.78 percent. Hence, it gives a good sign about the performance and whether or not that person suffers from heart disease and prospects K-Means may have in healthcare analytics in the future.

3.4 Methodology in Practice

In this study, several clustering models were evaluated to efficiently analyze the dataset. The models implemented are summarized below:

3.4.1 K-Means:

It's an algorithm for clustering that is based on the concept of centroids, and it's pretty simple yet powerful. It is very handy if data points belong to well-separated clusters with approximately spherical shapes. K- Means performs well when clusters are well separated, improving clusters by iteratively updating centroids until cluster assignments do not change anymore. But it may suffer the drawback of complicated shapes and outliers' sensitivity since it supposes that clusters are roughly of equal density and approximately spherical.

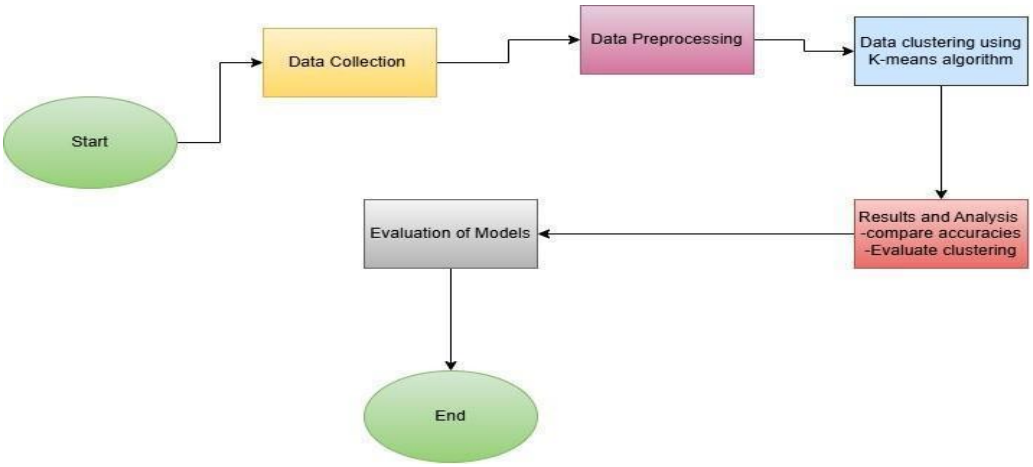

Fig. 1. Visual representation of K-means Clustering

The flowchart (Fig 1) describes a data clustering project by using the K- means algorithm. The process begins with Start and moves forward to Data Collection, followed by Data Preprocessing. Once preprocessed, Data Clustering using K-means Algorithm takes place. After that comes Results and Analysis where the accuracies are compared and the clustering is evaluated. The last is, the process passes through Evaluation of Models, to end.

3.4.2 DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

It is designed such that it identifies clusters with varying shapes and sizes. It groups the closely packed data points into clusters while labeling points that lie in low-density regions as outliers, thus making it resistant to noise. DBSCAN is especially useful when one needs to identify non-convex clusters or should separate noise points. This algorithm is sensitive to the parameter setting, which may influence detecting cluster shape and size.

3.4.3 Agglomerative Clustering:

It is a type of hierarchical clustering where individual data points are taken as clusters themselves and are merged iteratively based on proximity. At the end, this algorithm yields a tree-like structure known as dendrogram; this is especially helpful for data that naturally form some kind of hierarchy. One of the flexibility features in this algorithm is the linkage criteria used to merge the clusters; these criteria are called single, complete, or average. This technique of agglomerative clustering has a disadvantage when large amounts of data are being clustered; it becomes very computationally intensive.

3.4.4 Spectral Clustering:

It Leverages the power of linear algebra to create a similarity matrix for points in data and uses the eigenvalues to reduce dimensions before clustering. It's particularly effective when the shape is complex, non-convex, and the algorithm is frequently used for image segmentation and graph-based clustering. Spectral clustering can be very computationally expensive because of eigenvalue decomposition, making it infeasible for re- ally huge datasets.

3.4.5 Gaussian Mixture Model (GMM): It is a probabilistic model of clustering that describes data as a mixture of a number of Gaussian distributions. GMM offers "soft" clustering, meaning points belong to more than one class with some probability. They are useful when the cluster is overlapping or when their shapes and sizes differ significantly. GMM can model a large number of other distributions but is sensitive to the number of its components and often converges towards local optima, meaning the accuracy of clusters depends on the quality of such optima.

4 Results and Discussion

In this section, we present the results of various clustering models on the dataset.

Model	Precision	Recall	F1-Score	Accuracy
KMeans	0.84	0.84	0.84	0.8478
DBSCAN	0.22	0.34	0.27	0.4146
Agglomerative Clustering	0.36	0.48	0.32	0.6780
Spectral Clustering	0.74	0.59	0.52	0.5161
GMM	0.75	0.51	0.36	0.5951

Fig. 2. Performance comparison - Clustering models

Above table (Fig 2) represents the performance metrics of five different clustering models, including K-Means, DBSCAN, Agglomerative Clustering, Spectral Clustering, and Gaussian Mixture Model (GMM). The metrics include Precision, Recall, F1 Score, and Accuracy.

K-Means: The lowest performance is of DBSCAN with an accuracy of 0.4146. - Agglomerative Clustering Moderate performance. Accuracy 0.6780. Spectral Clustering and GMM are the methods that yield intermediate level results with the respective accuracies of 0.5161 and 0.5951.K-Means performs best of all models for all one of the metrics.

4.1 K-Means Clustering

The K-Means model was tested by its clustering accuracy and detected the existence of different groups within the data set. It shows that the result obtained is very promising in clustering performance; thus, it can well partition the data into meaningful clusters. K-Means clustering tries to minimize the within-cluster sum of squares:

The Euclidean distance between two points A and B in an n-dimensional space is calculated using the formula:

$$d(A, B) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Where:

$d(A, B)$: The distance between points A and B

x_i: The i-th coordinate of point A.

y_i: The i-th coordinate of point B.

$(x_i - y_i)^2$: The squared difference between the i-th coordinates of points A and B.

To find the Euclidean distance:

1. Compute the difference between each pair of corresponding coordinates xi and y_i.
2. Square each difference.
3. Add all the squared differences together.
4. Take the square root of the sum to obtain the distance.

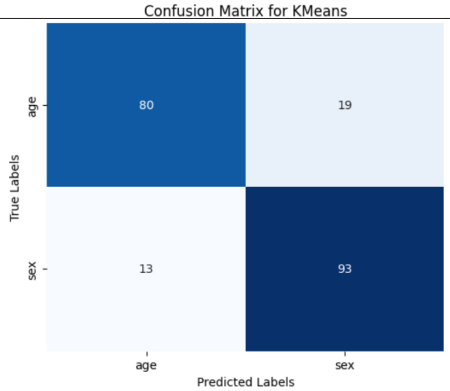

Fig. 3. Confusion matrix for K-Means

The above figure depicts the accuracy of a K-Means clustering model in two categories: "age" and "sex." Diagonal values present the count of correct assignments, including 80 "age" and 93 "sex" instances identified by the model. But misclassified 19, giving "sex" as a classification when the true label is "age," and 13, giving "age" when the true label is "sex." High values on the diagonal show that this model is typically quite good at correctly classifying these labels with only a few misclassifications.

4.2 Comparative Analysis of Models

Besides K-Means, other models were evaluated with their specific performance metric: accuracy, precision, and recall. Further understanding is made possible because of the ability of grouping data and anomalies

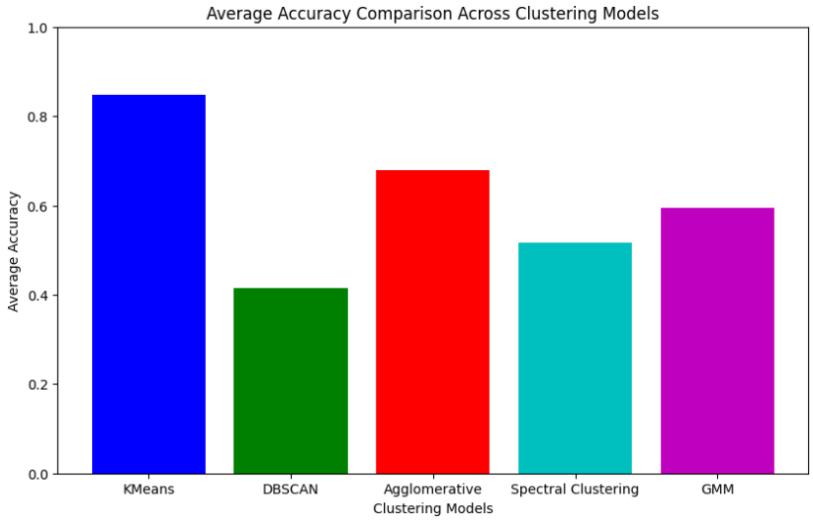

Fig. 4. Performance accuracy Visualization

Above bar plot (Fig 4) compares average accuracy comparing five clustering models: K-

Means, DBSCAN, Agglomerative Clustering, Spectral Clustering, and GMM. K-Means achieved the highest accuracy followed by Agglomerative Clustering and then GMM. Spectral Clustering is of medium accuracy while the lowest one was DBSCAN. Each model is represented with a filled bar over a different color with accuracy values ranging from 0.4 and up to about 0.85.

Input:

Fig. 5. User- Defined Data Input

Fig-5 depicts the medical form interface in the image is meant to receive information from patients, and most probably is employed to measure the risk of suffering from heart disease or to score cardiac health. The form, placed against the light blue background, offers a series of input fields through which users can present demographics and health-related information. These would enable a dropdown menu and numeric input fields, in which data specific to information about heart condition might be inserted. The form would be intuitive enough to be used easily for general estimations of the factors affecting heart disease.

Output:

Fig. 6. Generated result indicating chances of heart disease

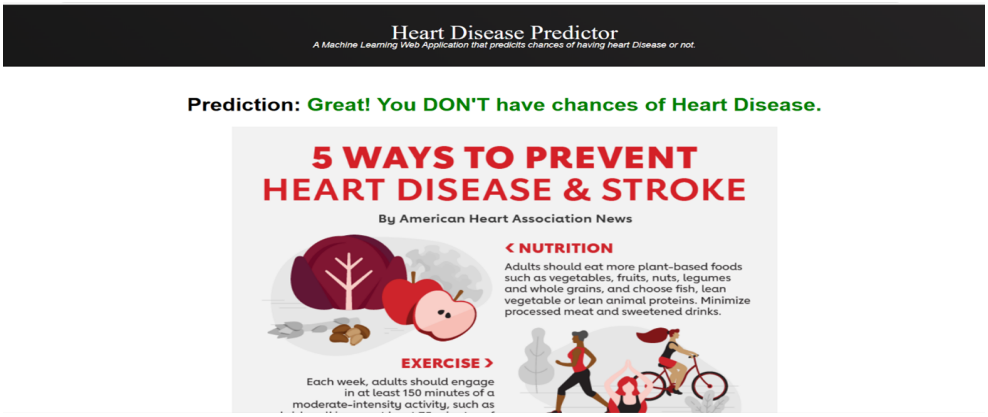

Fig. 7. Generated result showing no signs of heart disease

The result (Fig 6 & 7) for which the user input is given in the form and provided image is the interface of the web application 'Heart Disease Predictor'. It will provide a result for the assessment that is carried out by using machine learning algorithms from a single person's chance to have heart diseases. There are two messages of prediction - 'Great! You DON'T have a risk of heart disease.' and 'Oops! You have a risk of heart disease.

5 Conclusion

This study has demonstrated how one can predict heart disease inexpensively and in large-scale using unsupervised learning algorithms. This model classifies the patients according

to their health data, and this method identifies the possible risk factors of the patient as well by allowing the clustering technique with their unlabeled datasets only. In these three methods, K-Means proves better as an accuracy of around 84.78 percent is noticed that classifies the patient very effectively at very high risk.

Moreover, the system flags up the anomaly detection of outliers that may offer early intervention for such patients with abnormal health profiles sometimes. Real-time monitoring of health using wearable devices enhances the capability together with assessment in real time, alerted.

The model addresses an environment having limited labeled data amply adequate to overcome the drawback of the classical supervised learning system through a scalable method and may have a good prospect in successful improvement of personalized health care and preventive strategies much through the great early detection of heart diseases that reduces the burden upon the healthcare systems.

References

1. Bizimana, P. C., Zhang, Z., Asim, M., El-Latif, A. A. A., & Hammad, M. (2024). Learning-based techniques for heart disease prediction: a survey of models and performance metrics. Multimedia Tools and Applications, 83(13), 39867-39921.

2. Manikandan, G., Pragadeesh, B., Manojkumar, V., Karthikeyan, A. L., Manikandan, R., & Gandomi, A. H. (2024). Classification models combined with Boruta feature selection for heart disease prediction. Informatics in Medicine Un- locked, 44, 101442.

3. Murthy, M. S. N., Vinutna, P. N. S. S., Kumar, P. A., Shravani, P., & Brahmaji, P. (2024). Predicting Heart Disease Cases through Unsupervised Approaches. G. V. P. College of Engineering (A), Journal of Scientific Computing, 13(7), 1-5. DOI: 16.10089.JSC. 2024.V13I7.285311.3000.

4. Nanehkaran, Y. A., Licai, Z., Chen, J., Jamel, A. A., Shengnan, Z., Navaei, Y. D., & Aghbolagh, M. A. (2022). Anomaly Detection in Heart Disease Using a Density- Based Unsupervised Approach. Wireless Communications and Mobile Computing, 2022(1), 6913043.

5. Nouraei, H., & Rabkin, S. W. (2022). Comparison of unsupervised machine learning approaches for cluster analysis to define subgroups of heart failure with preserved ejection fraction with different outcomes. Bioengineering, 9(4), 175.

6. N. Sabri et al., "Heart Inspect: Heart Disease Prediction of an Individual Using Naïve Bayes Algorithm," 2023 IEEE 11th Conference on Systems, Process Control (ICSPC), Malacca, Malaysia, 2023, pp. 350- 354, doi:10.1109/ICSPC59664.2023.10420149.

7. Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2), 144.

8. Radwan, M., Mohamed Abdelrahman, N., Wael Kamal, H., Khaled Abdelmonem Elewa, A., & Moataz Mohamed, A. (2023). ML Heart Disease Prediction: heart disease prediction using machine learning. Journal of Computing and Communication, 2(1), 50-65.

9. Segar, M. W., Patel, K. V., Ayers, C., Basit, M., Tang, W. W., Willett, D., ... & Pandey, A. (2020). Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. Eu- ropean journal of heart failure, 22(1), 148-158.

10. Radwan, M., Mohamed Abdelrahman, N., Wael Kamal, H., Khaled Abdelmonem Elewa, A., & Moataz Mohamed, A. (2023). ML Heart Disease Prediction: heart disease prediction using machine learning. Journal of Computing and Communication, 2(1), 50-65.

11. A. Bhowmick, K. D. Mahato, C. Azad and U. Kumar, "Heart Disease Prediction Using Different Machine Learning Algorithms," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2022, pp. 60-65, doi: 10.1109/AIC55036.2022.9848885.

12. A. R. Panda, M. K. Mishra, M. Kumar Gourisaria, S. Pal, P. K. Pattnaik and S.K. Swain, "Heart Disease Prediction: A Comparative Analysis of Machine Learning Algorithms," 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/NMITCON62075.2024.10698898.

13. Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

14. Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.

15. S. Upadhyay, A. Dwivedi, A. Verma and V. Tiwari, "Heart Disease Prediction Model using various Supervised Learning Algorithm," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 197-201, doi: 10.1109/CSNT57126.2023.10134595.

16. Vayadande, K., Golawar, R., Khairnar, S., Dhiwar, A., Wakchoure, S., Bhoite, S., & Khadke, D. (2022, May). Heart disease prediction using machine learning and deep learning algorithms. In 2022 international conference on computational intelligence and sustainable engineering solutions (CISES) (pp. 393-401). IEEE.