International Journal of Scientific Research and Engineering Development-– Volume 6 Issue 5, Sep-Oct 2023
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Utilizing Machine Learning Algorithms for the Early Prediction of Heart Disease: A Comparative Study

Kaushal Kishor Gupta[1], Dr Rajesh Keshavrao Deshmukh[2]

[1]Research Scholar, Department of Computer Science and Engineering, Kalinga University, Raipur (C.G.)
[2]Assistant Professor, Department of Computer Science and Engineering
Kalinga University, Naya Raipur, Chhattisgarh, India Year 2023

Abstract

The increasing reliance on computer technology in the healthcare industry has led to the accumulation of vast amounts of electronic data. This situation presents challenges for healthcare professionals and doctors, who must navigate complex information to diagnose symptoms accurately and identify diseases in their early stages. The objective of this research is to determine the feasibility of forecasting infectious disease outbreaks in advance through the application of machine learning techniques. The methodology employed adheres to the standards set by the Cochrane Collaboration, as well as the protocols for conducting meta-analyses of observational studies in epidemiology and the established criteria for systematic study and meta-analyses reporting.

Index terms Machine learning; Disease Prediction, Dataset, Healthcare System

Introduction

Predicting human diseases plays a vital role in healthcare. Timely identification of diseases is a critical factor in effective treatment, a responsibility traditionally managed by doctors. The healthcare sector is heavily reliant on continual innovation to enhance efficiency [1]. The heart of the medical field is its capacity for innovation, crucial for developing new treatments, cures, and therapeutic approaches [2]. It keeps the sector up-to-date and significant. There's a broad spectrum for growth in the medical field, with numerous areas needing innovative approaches for advancement, including new disease treatments, enhancing patient care, and streamlining medical procedures [3, 4]. In today's digital era, such innovations often involve digitalizing medical procedures [5]. A significant challenge in healthcare is the heavy workload shouldered by doctors [6] and the high costs of medical consultations [7]. These issues become more apparent in disease prediction, which typically begins with a patient consulting a generalist doctor about their symptoms. The generalist then makes an initial diagnosis and refers the patient to a specialist for further examination [8].Diseases globally pose a significant challenge in healthcare, with increased mortality rates attributed to these conditions. The financial burden of treating these diseases often accounts for more than 70% of a patient's income. Therefore, reducing the risk factors associated with these diseases is crucial. Recent advancements in medical research have simplified the collection of health-related data [4], encompassing patient demographics, medical reports, and disease history. The diversity of diseases can be influenced by regional factors and living conditions, necessitating the inclusion of environmental and habitat data in patient records. The healthcare sector has undergone considerable evolution recently, thanks in part to the incorporation of information technology (IT). The goal of integrating IT into healthcare is to enhance affordability and comfort in individual lives, similar to the impact of smartphones [13]. Examples of this trend include the development of smart ambulances and hospital facilities, which benefit both patients and doctors [12]. Research focused on a specific region

revealed a minimal gender disparity among patients suffering from chronic diseases, with a notable increase in admissions in 2014. The combination of structured and unstructured data yields more accurate results than using structured data alone. Unstructured data includes doctors' notes and patients' descriptions of symptoms and experiences, providing valuable context to structured data like demographics, disease details, living environments, and lab results [5, 6].

II Review of Literature

In the introductory sections, various research papers were reviewed, highlighting multiple models for predicting diseases based on patient symptoms. Among these, some models stand out for their accuracy but also present certain limitations:Cao et al. [1] utilized a Support Vector Machine (SVM) for disease classification based on symptoms. While SVM is effective, it tends to be time-consuming and struggles with enhancing accuracy. Its main limitation lies in using a hyperplane for classification, which is only efficient for binary class separation, not suitable for the multiple disease classes required in medical diagnostics [2]. Hamidiet al. [2] implemented the K-Nearest Neighbors (KNN) algorithm, which assigns a data point to the most common class among its K nearest data points. This method is affected by noisy and incomplete data and has shown lower accuracy when considering factors like age, symptoms, and gender [3]. Similarly, Kashvi et al. [16] also used KNN, achieving high accuracy in certain cases like diabetes and heart risk prediction, but their methodology is limited by the small data set size [8].Pingale et al. [7], employed the Naïve Bayes method for disease prediction. While Pingale et al. focused on a small range of diseases with a limited dataset, MacLeod[14], developed a web application for disease prediction, with accuracy dependent on the dataset quality. Both approaches face challenges in dataset comprehensiveness and accuracy [11].Chhogyal and Nayak [9] also used the Naïve Bayes classifier but obtained poor accuracy, partly due to not utilizing a standard dataset for training [7].Kumar et al. [10] applied the RUSBoost Algorithm, designed to tackle class imbalance. However, its reliance on random under-sampling may result in vital information loss, making it less favorable for comprehensive training [10].These methods reveal various challenges in machine learning for disease prediction, including issues with efficiency, accuracy, limited dataset size, and a narrow scope of symptoms. To address these shortcomings, a more refined and accurate model is proposed, which is detailed in the following section.

III Proposed methodology

In this study, we combined both structured and unstructured data from healthcare sources to evaluate disease risk. We employed a latent factor model to fill in gaps in medical records sourced online. This approach allowed us to analyse major chronic diseases prevalent in certain regions and populations using statistical data. Additionally, we sought insights from hospital experts to identify relevant features for effectively handling structured data.Various techniques are employed in data mining, with machine learning being a prominent approach. The machine learning strategy encompasses methods like grouping, clustering, and summarization. This project specifically utilizes classification, a key process in data mining for categorizing categorical data. The classification process is divided into two stages: training and testing. During the training phase, predefined data along with their corresponding class labels are utilized to perform classification, a process commonly known as supervised learning. The classification process involves two key phases, illustrated in a diagram: the training phase, where training tuples (data sets) are used, and the testing phase, where test data tuples are applied to evaluate the accuracy of the classification rules. If the classification rules demonstrate sufficient accuracy on the test data, they are

deemed suitable for classifying new, unprocessed data.The proposed methodology for the project described above involves a structured approach to data mining using machine learning, specifically focusing on the classification technique. This methodology is outlined in several key steps:

Data Collection and Preparation: Gather both structured and unstructured healthcare data. This data will include a variety of elements, such as patient records, demographic information, and medical history.

Data Integration: Combine the structured and unstructured data to create a comprehensive dataset. This step may involve using a latent factor model to address any missing data, particularly from online medical records.

Feature Selection: Consult with hospital experts to identify the most relevant features for disease classification. This step ensures that the data used in the model is pertinent and impactful.

Classification Process: Divide the classification process into two main phases: training and testing.

- **Training Phase (Supervised Learning)**: Utilize predetermined data sets with associated class labels to train the Random Forest model. This phase involves feeding the model with known data to enable it to learn and recognize patterns.
- **Testing Phase**: Apply the trained model to a set of test data tuples not previously exposed to the model. This phase is crucial for evaluating the effectiveness and accuracy of the classification rules developed during the training phase.

Evaluation of Model Accuracy: Assess the accuracy of the classification rules based on their performance with the test data. The model's ability to correctly classify the test data is a key indicator of its suitability for practical application.

Application to Unmined Data: Once the model demonstrates sufficient accuracy with the test data, it can be applied to classify new, unmined data, thereby aiding in the prediction and analysis of disease risks.

This methodology aims to leverage the strengths of machine learning in the field of healthcare data mining, particularly for disease risk assessment and prediction.

3.1 Heart Disease Prediction

Heart disease encompasses a range of conditions that affect the heart's structure and function.This is the most common type of heart disease and occurs when the arteries that supply blood to the heart muscle become hardened and narrowed due to the buildup of cholesterol and other material, known as plaque, on their walls. The data have taken from [15]. This can lead to chest pain (angina) or a heart attack.The heart can beat too fast, too slow, or irregularly. Arrhythmias can be benign or life-threatening and include conditions such as atrial fibrillation, tachycardia, and bradycardia.

Cardiomyopathyis a heart muscle that makes it harder for the heart to pump blood to the rest of the body. Cardiomyopathy can lead to heart failure.

Heart Valve Disease occurs when one or more of the valves in the heart don't work properly, which can affect the blood flow through the heart to the body. Prevention and management of heart disease involve lifestyle changes such as diet, exercise, quitting smoking, and managing stress. Medical treatments can include medications, surgery, or other interventions depending on the specific type of heart disease. Machine learning models, like the ones discussed in the dataset screenshots you shared, are increasingly used to predict heart disease risk by analysing a variety of health indicators and historical data. These models can assist healthcare professionals in early detection and treatment, potentially improving patient outcomes.

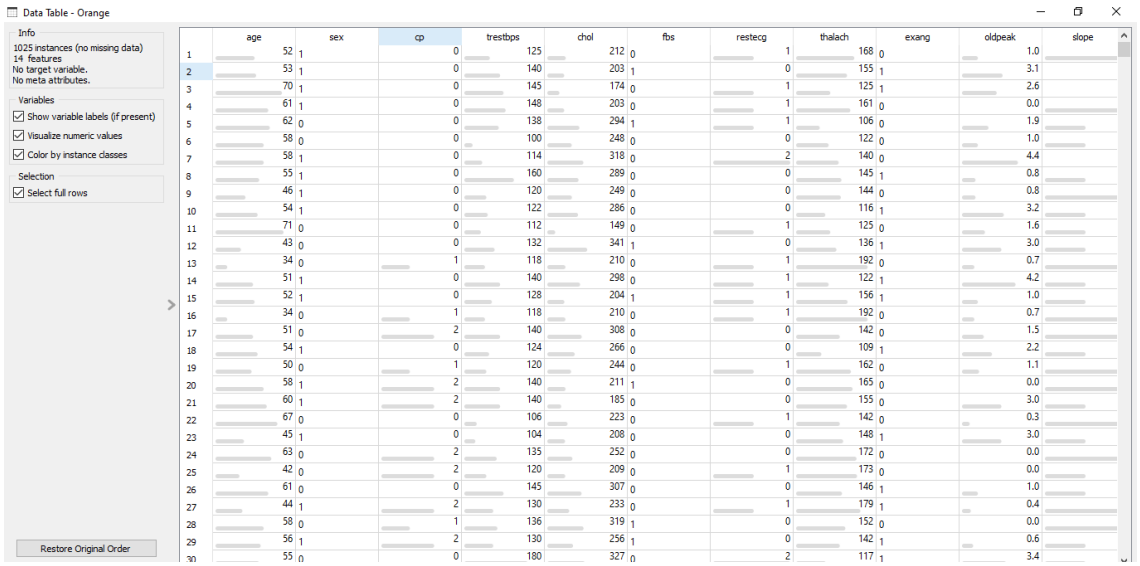


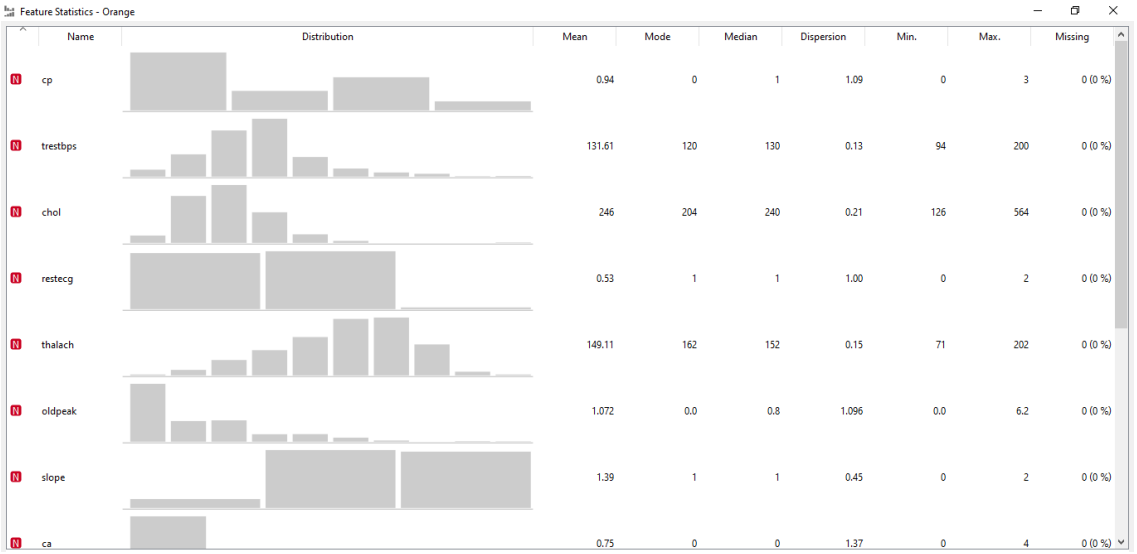
Figure 1: Data table

In above figure appears to be a screenshot of a data table from a software application that is used for data analysis, possibly a tool like Orange Data Mining or a similar data visualization and analysis software. Here's a breakdown of what the various columns in the data table likely represent:

1. **age**: The age of the individual.
2. **sex**: The biological sex of the individual (often coded as 0 and 1, where 0 could mean female and 1 could mean male, but the specific coding should be confirmed by the dataset documentation).
3. **cp** (chest pain type): The type of chest pain experienced by the individual, which is typically a key indicator in diagnosing heart conditions. This is also coded numerically, which likely corresponds to different types of chest pain.
4. **trestbps** (resting blood pressure): The resting blood pressure measurement in mm Hg (millimeters of mercury).
5. **chol** (serum cholesterol): The individual's cholesterol measurement in mg/dl (milligrams per deciliter).

6. **fbs** (fasting blood sugar): Indicates whether the individual's fasting blood sugar is above 120 mg/dl (1 if true, 0 if false).
7. **restecg** (resting electrocardiographic results): The results of the resting electrocardiogram tests.
8. **thalach** (maximum heart rate achieved): The highest heart rate the individual achieved during a stress test.
9. **exang** (exercise-induced angina): Indicates whether the individual experienced chest pain during exercise (1 if yes, 0 if no).
10. **oldpeak** (ST depression induced by exercise relative to rest): The measurement of the ST segment depression on an EKG after exercise compared to the ST segment at rest.
11. **slope**: The slope of the peak exercise ST segment, which is an indicator used in diagnosing myocardial ischemia.

Each row in the data table corresponds to a single patient or individual, with their associated measurements and categorical data. This type of data is often used to train machine learning models to predict heart disease risk. The presence of both continuous variables (like 'age', 'trestbps', 'chol', 'thalach', and 'oldpeak') and categorical variables (like 'sex', 'cp', 'fbs', 'restecg', 'exang', and 'slope') makes it suitable for a range of machine learning techniques, including the Random Forest classifier mentioned earlier. The dataset does not seem to have a specific target variable labeled in this screenshot, which would be necessary for supervised learning tasks such as classification.


Feature Statistics - Orange

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
cp		0.94	0	1	1.09	0	3	0 (0 %)
trestbps		131.61	120	130	0.13	94	200	0 (0 %)
chol		246	204	240	0.21	126	564	0 (0 %)
restecg		0.53	1	1	1.00	0	2	0 (0 %)
thalach		149.11	162	152	0.15	71	202	0 (0 %)
oldpeak		1.072	0.0	0.8	1.096	0.0	6.2	0 (0 %)
slope		1.39	1	1	0.45	0	2	0 (0 %)
ca		0.75	0	0	1.37	0	4	0 (0 %)

Figure 2: feature Statics

Above figure of feature statics 'cp' (which might stand for chest pain type):

- The mean value is 0.94.
- The mode is 0, indicating that 0 is the most common category for this feature in the dataset.
- The median is 1, which suggests that half of the values are below 1 and half above.

- The dispersion is 1.09, which may indicate the average distance of the data points from the mean.
- The minimum value is 0, and the maximum is 3, suggesting that 'cp' is a categorical feature with four possible categories (0, 1, 2, 3).
- There are no missing values for 'cp'.

These statistics are crucial for understanding the data's characteristics and for informing the pre-processing and analysis steps in machine learning or statistical modelling. For instance, high dispersion might indicate the need for data normalization, and the presence of missing values would require a decision on imputation or exclusion of incomplete records.

IV Model Description and Evaluate

The below figure have provided appears to be a screenshot from a data analysis software tool showing a comparison of machine learning models based on their performance metrics. The models tested include a Decision Tree (Tree), k-Nearest Neighbors (kNN), and Support Vector Machine (SVM). Here's a breakdown of the information shown:

- The table at the bottom shows the statistical comparison of the model performance based on MSE. The values represent the probability that the score for the model in the row is higher than that of the model in the column, with small numbers suggesting a negligible difference.
- The diagonals are all 1.000, which is trivial because a model will always have the same performance as itself.
- The off-diagonal values, which are all zeros, indicate that there is a statistically significant difference in performance between the models, given that the 'Negligible diff.' threshold is set to 0.1.

The below figure 3, displays the following values for each of the machine learning models evaluated:

- **Decision Tree (Tree)**
    - MSE (Mean Squared Error): 4.749
    - RMSE (Root Mean Squared Error): 2.179
    - MAE (Mean Absolute Error): 0.556
    - R2 (R-squared): 0.942
- **k-Nearest Neighbors (kNN)**
    - MSE: 43.082
    - RMSE: 6.564
    - MAE: 4.976
    - R2: 0.476
- **Support Vector Machine (SVM)**
    - MSE: 65.354

- RMSE: 8.084
- MAE: 6.587
- R2: 0.205

The Figure 3, below these metrics compares the models by their Mean Squared Error. It shows probabilities indicating the likelihood that the model in the row has a higher MSE than the model in the column. With the negligible difference set to 0.1, the table entries are:

- For the Tree row, all zero probabilities against kNN and SVM, indicating that the Tree model has a significantly lower MSE compared to both kNN and SVM.
- For the kNN row, a 1.000 probability against Tree, and all zero against SVM, suggesting that kNN has a higher MSE than the Tree model but a significantly lower MSE than SVM.
- For the SVM row, a 1.000 probability against both Tree and kNN, which means SVM has a higher MSE than both the Tree and kNN models.

This statistical comparison suggests that the Decision Tree model outperforms the kNN and SVM models for this particular dataset based on the MSE metric. The zero probabilities indicate a significant difference in performance, and the value of 1.000 indicates a worse performance.

V Conclusion

The evaluation of the machine learning models for heart disease prediction reveals that the Decision Tree model outperforms the k-Nearest Neighbors (kNN) and Support Vector Machine (SVM) in accuracy. With the lowest Mean Squared Error and the highest R-squared value, the Decision Tree is the most reliable model according to the provided dataset and chosen evaluation metrics. However, this does not discount the importance of a thorough validation process, including the assessment of potential overfitting and the consideration of other performance metrics that could impact the model's deployment in a real-world scenario. The statistical significance of the Decision Tree's superior performance suggests it as the preferable model for this application, but careful consideration and further testing are advised before finalizing the model for practical use in predicting heart disease.

References

1. Cao, J., Wang, M., Li, Y. and Zhang, Q., "Improved support vector machine classification algorithm based on adaptive feature weight updating in the hadoop cluster environment", PloS One, Vol. 14, No. 4, (2019), e0215136. https://doi.org/10.1371/journal.pone.0215136 12.

2. Hamidi, H. and Daraee, A., "Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases", International Journal of Engineering, Transactions B: Applications, Vol. 29, No. 7, (2016), 921-930.

3. Pisner, D.A. and Schnyer, D.M., Support vector machine, in Machine learning. 2020, Elsevier.101-121.

4. Chen, J., Yu, J., Wen, J., Zhang, C., Yin, Z.e., Wu, J. and Yao, S., "Pre-evacuation time estimation based emergency evacuation simulation in urban residential communities", International Journal of environmental Research and Public Health, Vol. 16, No. 23, (2019), 4599. doi: 10.3390/ijerph16234599.

5. Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., "Disease prediction from various symptoms using machine learning", Available at SSRN 3661426, (2020).

6. Taunk, K., De, S., Verma, S. and Swetapadma, A., "A brief review of nearest neighbor algorithm for learning and classification", in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), IEEE. (2019), 1255- 1260.

7. Pingale, K., Surwase, S., Kulkarni, V., Sarage, S. and Karve, A., "Disease prediction using machine learning", International Research Journal of Engineering and Technology (IRJET), Vol. 6, (2019), 831-833. doi: 10.1126/science.1065467.

8. Ibrahim, I. and Abdulazeez, A., "The role of machine learning algorithms for diagnosing diseases", Journal of Applied Science and Technology Trends, Vol. 2, No. 01, (2021), 10-19. doi: 10.38094/jastt20179.

9. Chhogyal, K. and Nayak, A., "An empirical study of a simple naive bayes classifier based on ranking functions", in AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29, Springer., (2016), 324-331.

10. Kumar, A., Bharti, R., Gupta, D. and Saha, A.K., "Improvement in boosting method by using rustboost technique for class imbalanced data", in Recent Developments in Machine Learning and Data Analytics: IC3 2018, Springer., (2019), 51-66.

11. Gupta D., Khare S., Aggarwal A. A method to predict diagnostic codes for chronic diseases using machine learning techniques. Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA); April 2016; Greater Noida, India. IEEE; pp. 281–287.

12. Chen M., Yixue H., Hwang K., Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IeeeAccess . 2017;5:8869–8879. doi: 10.1109/access.2017.2694446.

13. Ge R., Zhang R., Wang P. Prediction of chronic diseases with multi-label neural network. IEEE Access . 2020;8:138210–138216. doi: 10.1109/access.2020.3011374.

14. MacLeod H., Yang S., Kim O., Kay C., Natarajan S. Identifying rare diseases from behavioural data: a machine learning approach. Proceedings of the 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE); June 2016; Washington, DC, USA. IEEE; pp. 130–139.

15. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

