IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION: UNVEILING INSIGHTS WITH K-NEAREST NEIGHBOURS ALGORITHM

[1]C.SAMEERA ANJUM,[2]C.BABY,[3]K.NAVITA, [4]P.SHAHINAZ, [5]B.UMME SALMA, [1]STUDENT,[2] STUDENT,[3]STUDENT, [4]STUDENT, [5]STUDENT
[1]Department of Computer Science and Engineering,
1 Sri Venkateswara Institute of Technology(9F), Ananthapur, A.P

Abstract : Heart disease remains a significant health challenge in India, with millions affected annually, highlighting the pressing need for accurate diagnosis and effective management strategies. This chronic condition, characterized by inadequate blood supply to the body due to coronary artery narrowing or blockage, presents diverse symptoms including fatigue, swelling, and shortness of breath. Despite advancements in medical technology, diagnosing and treating heart disease remains complex, particularly in underserved regions. Cardiovascular diseases continue to be a leading cause of mortality in India, underscoring the urgency for improved healthcare delivery.

In parallel, healthcare systems in countries like Indonesia face challenges due to a shortage of medical professionals, prompting research into innovative solutions for patient monitoring. Data mining techniques have emerged as valuable tools in healthcare diagnosis, offering insights into disease prediction and management. Various algorithms, including CART, ID3, Naive Bayes, and KNN, have been explored in predicting heart disease with varying degrees of accuracy.

This study proposes a novel approach to heart disease prediction using the K-nearest neighbors (KNN) algorithm with instant measurement parameters. By leveraging body vital signs that can be instantly measured, this approach eliminates the need for invasive procedures, offering real-time monitoring particularly beneficial for home-based patient care. This innovative strategy aligns with advancements in patient monitoring technologies, aiming to enhance healthcare outcomes and reduce the burden of heart disease on the Indian healthcare system.

IndexTerms **- Component,formatting,style,styling,insert.**

I. INTRODUCTION

INTRODUCTION

1.NEED OF THE STUDY.

Heart disease remains a significant and life-threatening chronic condition in India, posing substantial challenges to healthcare systems nationwide. This condition arises when the heart fails to adequately supply blood to the body, often due to the narrowing or blockage of coronary arteries, essential for cardiac blood flow.Common symptoms of heart disease include fatigue, swelling in the feet, shortness of breath, and overall weakness. Several factors contribute to its prevalence, including smoking, poor diet, high blood pressure, high cholesterol, lack of physical activity, and other lifestyle habits.

Coronary artery disease (CAD) stands as the most prevalent form of heart disease, leading to chest pain, heart attacks, and strokes. Other types of heart ailments include congenital heart defects and various cardiovascular complications.

Despite advancements in medical technology, diagnosing and managing heart disease remains challenging in many parts of India, particularly in rural and underserved areas. Recent data highlights that cardiovascular diseases continue to be a leading cause of mortality in the country. Current estimates suggest that millions of Indians are affected by heart disease annually.

2.LITERATURE REVIEW

In recent years, there has been a notable shift towards developing non-invasive and predictive healthcare systems to address the challenges associated with traditional invasive methods for detecting heart problems. These systems leverage the predictive capabilities of machine learning techniques, including Random Forest, Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), and others. The adoption of such innovative approaches has contributed to a reduction in the mortality rate associated with heart disease.

One commonly used dataset in heart disease research is the Cleveland heart disease dataset, which has been extensively utilized by researchers. For instance, a study by [1] employed a logistic regression classification system to diagnose cardiac disease, achieving a commendable accuracy of 77.1%. Similarly, [2] presented the use of neural networks for diagnosing and predicting blood pressure and heart disease, achieving promising results.

In another approach, [3] employed Naive Bayes (NB) and Decision Tree (DT) for heart disease diagnosis and prediction, achieving reasonable accuracies of 82.7% and 80.4% respectively. Additionally, [4] developed a predictive system using NB approaches for categorizing data and the AES Advanced Encryption Standard (AES) algorithm for secure data transportation, effectively predicting risk factors related to heart diseases.

Moreover, [5] explored Naive Bayes and support vector machine (SVM) for classification on the UCI dataset, with SVM outperforming Naive Bayes in terms of classification accuracy. Similarly, [6] investigated the Multi-Layer Perceptron (MLP) Neural Network for predicting cardiac illness with high accuracy of 93.39% using significant attributes.

The combination of different machine learning techniques has also been explored, such as in [7] where a hybrid neural network consisting of ANN and fuzzy neural network resulted in an accuracy of 82.4% in heart disease classification. Furthermore, [8] developed a decision support system involving SVM and ANN for heart disease diagnosis, achieving an accuracy of 80%.

A survey conducted by [9] evaluated several classification algorithms for predicting the risk level of patients based on various parameters, including Naive Bayes, KNN, Decision Tree, and ANN. The authors found that the accuracy of risk prediction increases with the number of parameters.

Additionally, [10] proposed a hybrid recommender system for detecting heart disease, while [11] developed a machine learning-based framework achieving an accuracy of 86% in distinguishing between healthy individuals and those with heart problems.

Furthermore, [12] employed data mining methods, including Naive Bayes, Decision Tree, and bagging, achieving accuracies ranging from 82.31% to 85.03% for heart disease diagnosis. Meanwhile, [13] investigated deep learning networks for heart disease detection, claiming a high accuracy of 99% and a Matthews correlation coefficient (MCC) of 0.98.

Finally, [14] suggested a system utilizing data mining techniques with the MapReduce algorithm, achieving an accuracy of 98% in predicting heart disease risks, highlighting the potential of such techniques in clinical settings.

3.PROPOSED SYSTEM

Present time, data has been scattered as Statistics, Reports, and Forms and so forth. It is a vast benefit which allows the making of outcome in genuine time conditions. In spite of that, a group of study has been conducted in various areas, health care has a wide extension to utilize officially accessible information and determine results which will be available to the world. Cardiovascular illnesses consist of Heart and blood vessel sicknesses that comprise of many problems, lot of which are linked to an operation termed atherosclerosis. When a material termed plaque accumulates in the walls of the arteries and evolves that case is termed atherosclerosis. This accumulates and tightens the arteries making harder for blood to flow out of the arteries. The term Myocardial Infarction or stroke is when the blood becomes clot which can also cause a heart attack.

The proposed approach in this paper uses neighbouring based outlier detection mechanism to attain high prediction rate of heart disease from the used dataset.

A.DATASET: The dataset selected for the prediction of cardio vascular diseases (heart diseases) is collected from Statlog and Cleveland. Data set description is as given in Table1.

Table 1. Description of the Attributes in Dataset

Sno	Attributes	Data Type	Description
1	Age	Integer	Age in years
2	Sex	Binary	Gender of the individual (1: women, 2: men)
3	Cp	Categorical	Cp: chest pain type 0 : asymptomatic 1 : atypical angina 2 : non-anginal pain 3 : typical angina
4	Trestbps	Continuous	Resting blood pressure (in mm Hg)
5	Chol	Continuous	Serum cholesterol in mg/dl
6	Fbs	Binary	(Fasting blood sugar > 120 mg/dl) (1 =true;0=false)
7	Restecg	Categorical	Resting electrocardiographic results 0: showing probable or definite left ventricular hypertrophy by Ester' criteria; 1: normal; 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of>0.05mV)
8	Thalach	Continuous	Maximum heart rate achieved

9	Exang	Binary	Exercise induced angina (1 = yes;0= no)
10	Oldpeak	Continuous	ST depression induced by exercise related to rest
11	Slope	Categorical	The slope of the peak exercise ST segment 0: downsloping; 1: flat; 2: upsloping
12	Ca	Integer	Number of major vessels (0-3) colored by fluoroscopy *(4 missing values)
13	Thal	Categorical	Thallium stress test result 1 =fixed defect; 2= normal; 3= reversable defect) *(2 missing values)
14	Target	Binary	Presence of heart disease: 0 = disease, 1= no disease

The implementation is one of the most important tasks in the project is the phase where precautions are required as all the efforts taken during the project will be very interactive. Implementation is the key step-in reaching a good scheme and offering users trust that the new system is functional and efficient. At the moment of growth, each programmed, by means of the sample information, is evaluated separately and checked, in the mannerindicated in the program specification, that these programs connect to one other.

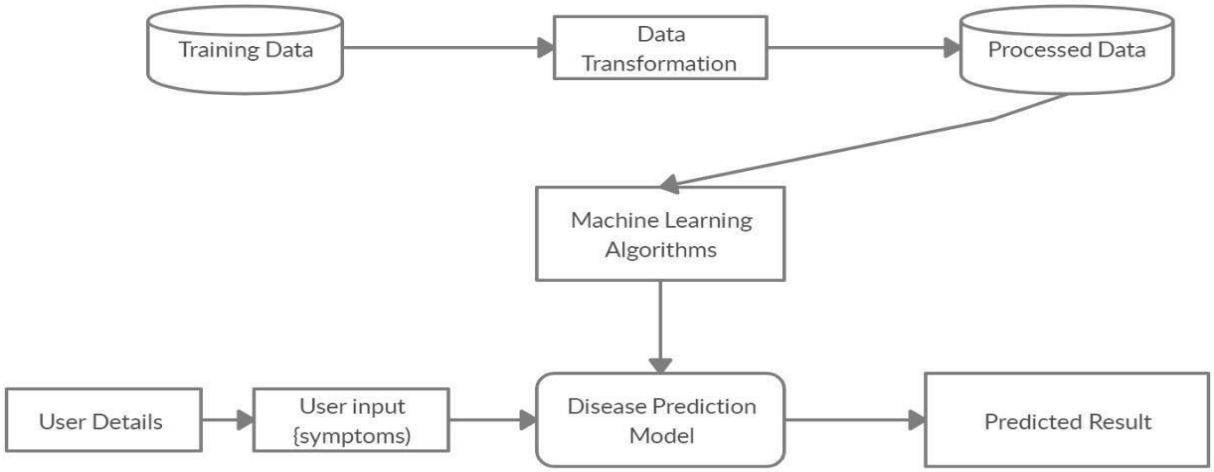

Figure 1: System Architecture

As shown in figure 1, the architecture of the system disease prediction using machine learning consist of various datasets through which compare the symptoms of the user and predicts it, then the datasets are transformed into the smaller sets and from there it gets classified based on the classification algorithms later on the classified data is then processed into the machine learning technologies through which the data gets processed and goes in to the disease prediction model using all the inputs from the user that is mentioned above. Then after user entering the above information and overall processed data combines and compares in the prediction model of the system and finally predicts the disease.

B. PRE-PROCESSING: Coronary illness data is pre-processed after the collection of different records. The dataset covers an aggregate of 303 patient histories, where 6 records are for certain missing qualities. Those six records have been taken out from the dataset and the leftover 297 patient records are utilized in pre-preparing.

In the case of the patient having a coronary illness, the worth is set to 1, else the worth is set to 0 signifying the shortfall of heart illness in the patient. The preparing of information is done by changing over clinical records into determination esteems. The consequences of information pre-handling for 297 patient records demonstrate that 137 records show the estimation of 1 setting up the presence of heart illness while the excess 160 mirrored the estimation of 0 showing the shortfall of coronary illness.

K NEAREST NEIGHBOUR ALGORITHM: The K-Nearest Neighbour classification algorithm is one of the many supervised classification algorithms. K-NN classifies object based on the similarity measures, which could be the distance functions. The K-NN algorithm is a non-parametric algorithm; no assumptions are required on the underlying data. K-NN classification works by calculating the distance between a new input with all the observations in the dataset, and then the algorithm classifies the new input with its nearest neighbour. There are multiple calculation methods to calculate the distance between points in a graph.

4. Experimental Results

Cleveland dataset is used for the study. Several classifiers were employed, namely Random Forest, Decision Tree, Naive Bayes, SVM, and KNN. To assess the performance of these algorithms, evaluation measures such as Accuracy score, Precision (P), Recall (R), and F-measure were used. Precision, specifically, evaluates the accuracy of positive predictions.

During simulations, the preprocessed dataset was utilized to conduct experiments and evaluate the performance of the aforementioned classification approaches using the mentioned metrics.

Table 2: Classification results on full feature space

	Algorithm	Accuracy	Precision	Recall	F-1 Score
1	Naive Bayes (NB)	86.7	83.7	91.1	87.3
2	Linear SVM	81.97	90.0	88.0	89.0
3	RF	86.81	88.0	88.0	88.0
4	DT	81.97	84.5	82.3	83.5
5	KNN	94.16	93.7	88.2	93.16

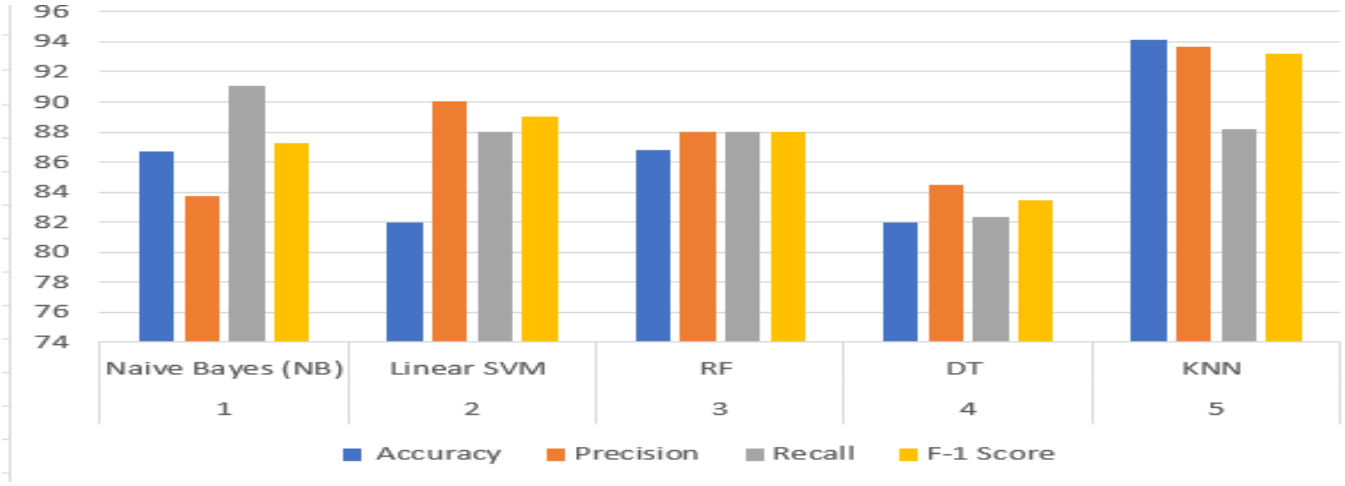

Fig2: Performance of Proposed Method

The performance of each classifier was further analyzed using a Confusion Matrix, which provides insights into the model's performance. Using the parameters derived from the confusion matrix, the performance measures mentioned earlier were calculated. Table2 and Fig2 showcases the accuracy and other metric scores calculated for the classification algorithms. Through empirical evaluation, the aim was to determine the success rates of all classifiers and understand their performance in detail.

Conclusion

The significant burden of heart disease in countries like India underscores the critical need for precise diagnosis and effective management strategies. Despite advancements in medical technology, challenges persist, particularly in underserved regions where cardiovascular diseases remain a leading cause of mortality. Similarly, countries like Indonesia grapple with healthcare system challenges due to a shortage of medical professionals, driving the exploration of innovative solutions for patient monitoring.

In this landscape, data mining techniques have emerged as valuable tools, offering insights into disease prediction and management. Various algorithms, including CART, ID3, Naive Bayes, and KNN, have been explored for predicting heart disease, each demonstrating varying levels of accuracy.This study introduces a novel approach to heart disease prediction, employing the K-nearest neighbors (KNN) algorithm with instant measurement parameters. By leveraging readily available body vital signs, this approach eliminates the necessity for invasive procedures and offers real-time monitoring, particularly advantageous for home-based patient care. This innovative strategy aligns with advancements in patient monitoring technologies, aiming to enhance healthcare outcomes and alleviate the burden of heart disease on healthcare systems.Looking ahead, there are several avenues for future work in this field. Integrating additional data sources, such as genetic information and lifestyle factors, could enhance the accuracy of predictive models. Advanced machine learning techniques, including deep learning and ensemble methods, hold promise for refining predictive models and uncovering complex patterns in heart disease data. Personalized medicine approaches tailored to individual patient characteristics represent another promising avenue for future research, as they can improve treatment outcomes by accounting for patient-specific factors.Furthermore, the integration of remote monitoring and telemedicine technologies into heart disease management strategies presents opportunities for improving patient care. Longitudinal studies tracking patients' health outcomes over time can provide valuable insights into disease progression and treatment effectiveness. Addressing barriers to healthcare access and implementing policy interventions to promote preventive care and early intervention are also crucial aspects of future work.By pursuing these

avenues for future research, we can advance our understanding of heart disease and develop more effective strategies for its diagnosis, management, and prevention. This, in turn, can lead to improved patient outcomes and reduced strain on healthcare systems globally.

REFERENCES

[1] Zhong Xiao, Ma Shaoping, et al. "Survey of Data Mining [J]", fuzzy recognition and artificial intelligence, 2001,01.

[2] Jiawei Han, MichelineKamber, "Data Mining Concepts and Techniques", [M]Second Edition,2007,3. page3-4.

[3] Agrawal. R.Datamining: Crossing the chasm[R/OL]. http://www.almaden.ibm.com.2002-11-20.

[4] John Durkin, CaiJingfeng, CAI Zixing, "Decision tree technology and its current research direction [J]", Control Engineering, 12 (1) 2005.

[5] J. Han and M. Kamber, "Data Mining: Concepts and Techniques". MorganKaufmann, 2000.

[6] T.Imielinski and H. Mannila, "A database perspective on knowledge discovery". Communications of ACM, 39:58-64, 1996.

[7] M. S. Chen, J. Han, and P. S. Yu, "Data mining: An overview from a database perspective". IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[8] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining". AAAI/MIT Press,1996.

[9] Marcos D. Assunção et.al. "Big Data computing and clouds: Trends and future directions", Journal of Parallel and Distributed Computing Volumes 79–80, May 2015.

[10] Xindong Wu, XingquanZhu, et al. "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, January 2014.

[11] Xingqanhu, lan Davidson, "Knowledge Discovery and Data Mining: challenges and Realities", ISBN 978-1-59904- 2521 Hershey, New York, 2007.

[12] Grabmeir. J, Rudolph, "Technique of clustering Algorithms in Data Mining",Data Mining and Knowledge Discovery, 2002.

[13] Gupta GK, "Introduction to data mining with case studies PHI", New Delhi,2012.

[14] Kumar R, Kapil AK, Bhatia, "A Modified tree classification in data mining". Global Journals Inc. 12, 12: 58-63, 2012.