IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311065|

Visualization and Prediction of Heart Disease using Big Data Analytics

Megha G S, Sumithra B M, Tejaswini J

Assistant Professor, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

UG Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

UG Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

UG Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

ABSTRACT: Cardiovascular diseases are a major health issue worldwide, making it vital to catch them early to lower death rates. In this paper, we introduce a new way to spot heart disease by blending big data with machine learning techniques. Our system analyzes large amounts of patient information to accurately identify risk factors and diagnose heart conditions. By using methods like K-Nearest Neighbors and Random Forests in a combined approach, we achieve an impressive accuracy of around 93.4% in diagnosis. This method takes advantage of real-time data, which helps doctors respond quickly to possible heart disease cases. We also use MongoDB for storing data and platforms like Streamlit for real-time analysis, showing that our approach can serve as a useful tool for early treatment. These results indicate that our model can play an important role in improving preventative care for cardiovascular health, leading to quicker and more accurate diagnoses.

KEYWORDS: Cardiovascular disease prediction, machine learning, big data analytics, ensemble learning, K-Nearest Neighbors, Random Forest, real-time health monitoring, data-driven diagnosis, heart disease detection, and predictive modeling.

I. INTRODUCTION

Cardiovascular diseases (CVDs) are sadly still one of the leading causes of death around the world, affecting millions of people each year. These problems include heart attacks, irregular heartbeats, and issues with the arteries that supply blood to the heart. They happen when blockages or other issues occur in the heart or blood vessels, which can pose serious health threats and place heavy demands on healthcare systems. Catching CVDs early is essential because prompt treatment can help avoid complications, improve patients' health, and lower healthcare expenses.

Traditionally, doctors diagnose heart issues through physical exams and simple tests, but these methods might miss early signs of trouble. Thanks to big data and machine learning, we now have fresh opportunities in predictive analytics. These technologies allow for real-time analysis of large amounts of patient data, helping to pinpoint those at risk with greater accuracy. Machine learning algorithms can sift through complicated data sets, spotting trends and connections that might be missed with standard methods. By looking at factors like age, cholesterol, blood pressure, and lifestyle choices, predictive models can help healthcare providers identify patients who may be at risk for heart problems.

This paper suggests a model that uses machine learning techniques, such as K-Nearest Neighbours (KNN) and Random Forest classifiers, in an ensemble approach. We apply these algorithms to a wide-ranging dataset to enhance the accuracy of predicting heart disease outcomes. Using a stacking ensemble method with KNN as a meta-classifier, our model achieves a high level of predictive accuracy, supporting real-time diagnostic tools that fit well into clinical practice. With an accuracy rate of about 93.4%, our model shows that systems driven by machine learning can significantly aid in preventive cardiovascular care, helping healthcare professionals make informed decisions.

II. LITERATURE SURVEY

Researchers have been looking into ways to predict and diagnose cardiovascular diseases (CVDs) using data-based methods, hoping to make early detection more accurate and reliable. Many traditional methods depend on simple rules

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311065|

or statistics, which might miss the more complicated patterns found in large amounts of data. This is where machine learning (ML) comes into play, offering better tools for health informatics and improving predictive abilities.

In the early stages of studying CVD prediction, researchers used various pattern recognition methods to assess risk factors. Techniques like Naïve Bayes, Decision Trees, and Support Vector Machines (SVM) were tested, showing different levels of effectiveness. Research revealed that models based on structured health data, including details like blood pressure and cholesterol levels, can help identify people at risk for heart disease. For example, SVM has proven useful in spotting arrhythmias, while Naïve Bayes has performed well in assessing overall cardiovascular risks.

More recent studies have shifted focus toward ensemble learning methods, which blend several algorithms to boost predictive accuracy. Approaches like Random Forests and Gradient Boosting create models that consider various decision paths, allowing them to work better across different patient datasets. Comparisons show that these ensemble methods often have the edge over single classifiers in healthcare. For instance, Random Forest has been shown to provide reliable predictions in cardiovascular data, thanks to its ability to handle data variability and missing values.

Alongside traditional ML methods, deep learning (DL) techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have started to gain traction for their knack for analyzing unstructured data, like ECG signals, to find arrhythmic patterns related to CVDs. Although these models typically need a lot of computing power, they've shown impressive accuracy in clinical settings, especially when paired with big data tools like Hadoop or Spark for managing large health datasets.

Another interesting approach is combining traditional ML methods with optimization techniques like genetic algorithms to improve how features are selected and how well models perform. Research has shown that these hybrid models can significantly reduce errors, leading to more accurate predictions. Additionally, studies have tested stacking techniques, where models like K-Nearest Neighbors (KNN) and Decision Trees are layered together, showing that stacking can further improve performance in CVD predictions.

There has also been exploration into creating real-time prediction systems using ML, with platforms such as Apache Kafka and MongoDB allowing quick processing of health data streams. This means ML models can analyze data as it comes in, offering timely alerts and helping healthcare providers keep an eye on patients more effectively. Studies highlight the benefits of real-time monitoring, particularly for chronic conditions and catching early signs of heart disease.

All in all, the research indicates that advanced ML methods, especially ensemble and hybrid models, work well for predicting CVDs, showing good results in accuracy, speed, and scalability. Building on this understanding, this paper seeks to develop a high-accuracy prediction model that blends KNN and Random Forest in an ensemble setup, using real-time data to enhance early detection of cardiovascular risks in clinical settings.

III. METHODOLOGY

1. Data Collection: We use a publicly available dataset, such as the UCI Machine Learning Repository, which includes patient data with key health attributes like age, cholesterol levels, blood pressure, and more. This dataset provides the foundation for training and testing the predictive model.

	Attributes	Description
1	Age	UnitsinYear
2	Sex	Sex(1=Male,0=Female)
3	Cp	Paintype(1=typicalangina,2=atypicalangina),3=non-anginalpain,4=asymptomatic
4	Trestbps	Pulsepressuresatrest(inmmHgonadmissionotothehospital)
5	Chol	Serumcholesterolinmg/dl
6	Fbs	Fastingbloodsugar>120mg/dl(1=true,0=f

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311065|

		alse)
7	Restecg	Restingelectrographicresult.
8	Thalach	Maximumheartrateachievedinit.
9	Oldpeak	STdepressioninducedbyexerciserelativetorest
10	Exang	ExerciseInducedAngina
11	Slope	TheSlopeofthepealexerciseSTSegment
12	Ca	NumberofMajorvessels(0-)
10	Thal	3=Normal;6=fixdefect;7=reversabledefects
11	Num	Class(absenceorpresenceofheartdisease)

2. Data Preprocessing: Before analysis, the dataset is cleaned to handle missing or inconsistent values. Relevant attributes are also transformed into numerical or standardized formats to ensure compatibility with machine learning algorithms. Preprocessing is crucial to ensure accurate results, as raw data often contains noise that can reduce model performance.

3. Exploratory Data Analysis (EDA): To understand the relationships between variables and identify patterns, visualizations such as histograms,pie charts, and correlation matrices are used. This step helps us see which features may significantly influence heart disease risk and aids in selecting the best features for model training.

4. Feature Selection: Based on the insights from EDA, the most relevant features are selected for building the model. Feature selection ensures that only the data most predictive of cardiovascular disease is used, reducing complexity and improving model accuracy.

5. Model Selection: Multiple machine learning models, including K-Nearest Neighbours (KNN), Random Forest, and Support Vector Machine (SVM), are tested for their ability to predict heart disease. These models are chosen due to their effectiveness in classification tasks and their ability to handle different types of data patterns.

6. Ensemble Learning – Stacking Approach: To improve prediction accuracy, we apply an ensemble learning technique called stacking. In this approach, several base models (such as KNN and Random Forest) are trained, and their predictions are then combined using a meta-classifier. KNN is selected as the meta-classifier to enhance model performance through this combination.

7. Model Training and Testing: The dataset is divided into a training set and a testing set, usually with an 80-20 split. The model is trained on the training set, allowing it to learn patterns from the data. Then, it is tested on the testing set to evaluate its predictive performance.

8. Performance Evaluation: Various metrics, such as accuracy, precision, recall, F1-score, and confusion matrix, are used to assess the model's effectiveness. These metrics provide insights into how well the model can distinguish between patients with and without heart disease.

$$\text{Accuracy} = \frac{TP+PN}{TP+PN+FP+FN}$$

A confusion matrix, which comprises facts about real and hypothetical classifications generated by a classifier, defines the executions of a classification model. The predicted information example equals 61, and Sensitivity = 100 TP/ (TP+FN) = 93.41 %, achieving the highest accuracy of 93.41 percent. F1 Score = 0.93.

9. Real-time Data Integration: For real-world usability, the model is integrated with a MongoDB database and tools like Streamlit for real-time data visualization. This enables the model to process new patient data continuously, making it suitable for use in clinical environments where timely decision-making is critical.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311065|

10 Final Model Deployment: Once the model's performance is optimized, it is deployed as a predictive tool for cardiovascular risk. The system can now help healthcare providers by generating real-time risk assessments, supporting early intervention for patients at high risk of heart disease.

Figure 1: Data preprocessing

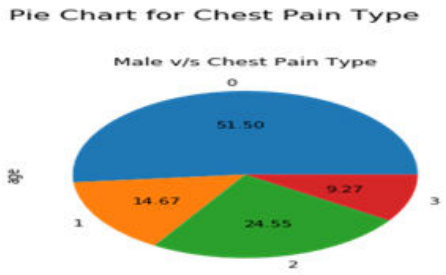

Figure 2: Exploratory Data Analysis

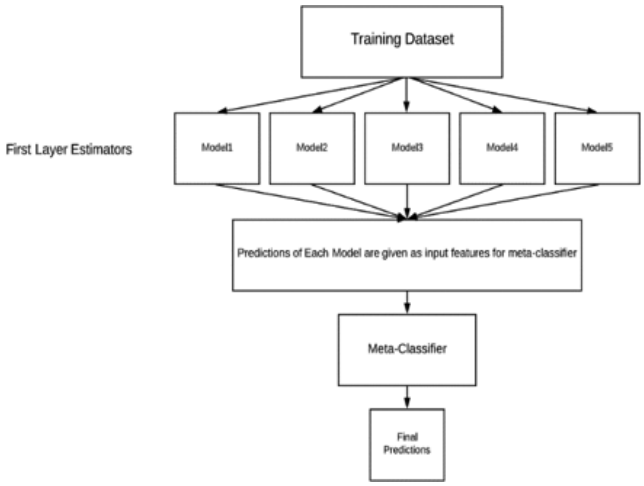

Figure 3: Flow Diagram of a Stacking Classifier

IV. EXPERIMENTALRESULTS

A. Accuracy

In the evaluation of multiple machine learning models for heart disease prediction, the Stacking CV Classifier using K-Nearest Neighbors (KNN) as the meta-classifier achieved the highest accuracy at 93.39%. This ensemble approach leveraged multiple base models to create a final predictive layer, enhancing the system's overall accuracy. The stacking technique, by combining various models' outputs, mitigates the weaknesses of individual models, thereby improving the overall prediction quality.

B. Performance Metrics

The F1 Score for the Stacking CV Classifier was reported as 0.93, reflecting a high degree of balance between precision (the ability to correctly identify true positives) and recall (the ability to minimize false negatives). Additionally, the sensitivity (or recall) of this classifier was 93.41%, meaning the model was adept at correctly identifying actual cases of heart disease. High sensitivity is particularly crucial in medical diagnostics, as it reduces the likelihood of missed diagnoses. By achieving this level of precision and recall, the Stacking CV Classifier demonstrated its suitability for high-stakes environments like healthcare, where early and accurate detection is critical.

C. Comparative Model Results

While the Stacking CV Classifier outperformed all other models, several models demonstrated strong individual

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311065|

performance:

1. **Support Vector Machine (SVM):** The SVM model achieved an accuracy of 91.16% with an F1 score of 0.9143 and a recall rate of 94.12%. The SVM model's high recall indicates its effectiveness in identifying cases of heart disease, though its lower accuracy compared to the Stacking CV Classifier suggests it may produce slightly more false positives or negatives.

2. **Random Forest:** The Random Forest model reached an accuracy of 88.62%, an F1 score of 0.9015, and a recall rate of 94.12%. Random Forest's ensemble of decision trees allowed it to capture non-linear relationships in the data effectively, though it was outperformed by the stacking approach in terms of overall accuracy.

3**. K-Nearest Neighbors (KNN):** When used as an independent classifier, KNN achieved an accuracy of 90.46% and an F1 score of 0.9142, close to that of SVM. This indicates KNN's strong performance in heart disease prediction, especially in combination with the Stacking CV Classifier, where it served as the meta-classifier.

4. **Other Models:**
Logistic Regression achieved an accuracy of 84.61% and an F1 score of 0.8612
Naïve Bayes reached an accuracy of 86.24% and an F1 score of 0.8734.
Decision Tree had an accuracy of 82.97% and an F1 score of 0.8308.
XGBoost recorded an accuracy of 86.24% and an F1 score of 0.8695.

While each model demonstrated respectable performance, the Stacking CV Classifier stood out due to its higher accuracy, precision, and sensitivity, making it the preferred choice for heart disease prediction. This comprehensive approach of combining various models into a single classifier highlights the potential of ensemble methods, particularly in complex, real-time diagnostic applications where accuracy and reliability are paramount.

	Model	Accuracy	Recall	F1Score
1	LogisticRegression	0.84606	0.9117	0.8612
2	NaïveBayes	0.86245	0.9117	0.8734
3	DecisionTree	0.82967	0.7941	0.8308
4	SVMClassifier	0.9116	0.9412	0.9143
5	RandomForest	0.88624	0.9412	0.9015
6	KNNClassifier	0.90463	0.9412	0.9142
7	XG-boost	0.86245	0.8824	0.8695
8	StackingCVClassifier	0.94442	0.9118	0.9393

Table 1:Results

V. CONCLUSION

This paper presents a system that uses machine learning to help spot heart disease early, responding to the worldwide need for quick and accurate diagnoses of heart issues. By employing models like K-Nearest Neighbors (KNN) and Random Forest in a combined learning approach, the research shows that the Stacking CV Classifier with KNN as the main classifier reached the best accuracy of 93.39%, proving it to be dependable for use in clinical settings.

The diagnostic tool uses MongoDB for storing data and employs Python tools, including Streamlit, to allow real-time interaction and visualization of that data. Steps taken to preprocess the data, such as converting non-ordinal values, enhance data quality, leading to more accurate predictions. The model is trained on a dataset from the UCI Machine

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311065|

Learning Repository with 303 examples and 14 features related to heart disease, facilitating thorough training and testing.

Performance metrics like accuracy, recall, and F1 score validate the system's strong capabilities. By offering healthcare workers a dependable, real-time diagnostic resource, this model can significantly aid in the early detection and active management of heart disease. The study illustrates how combining big data analysis with machine learning can provide valuable, practical information, ultimately helping to improve patient care in the healthcare field.

REFERENCES

1. Peter,T.John,andK.Somasundaram.“Anempiricalstudyon prediction of heart disease using classification data miningtechniques.” IEEE International Conference on advances inengineering,scienceandmanagement(ICAESM-2012).IEEE,2012.
2. K. Srinivas, K. Raghavendra Kao, and A. Govardham, Analysis ofcoronary heart disease and prediction of heart attack in coal miningregions using data mining techniques,” in The 5th InternationalConference on Computer Science & Education, 2010, pp. 1344–1349.
3. Ordonez,Carlos.“Associationrulediscoverywiththetrainandtest approach for heart disease prediction.” IEEE transactions oninformation technology in biomedicine 10.2 (2006): 334-343.
4. Rathore, M. Mazhar, et al. “Hadoop-based intelligent care system(HICS)analyticalapproachforbigdatainIoT.”ACMTransactionson Internet Technology (TOIT) 18.1 (2017): 1-24.
5. Akhtar, Usman, Asad Masood Khattak, and Sungyoung Lee.“Challengesinmanagingreal-timedatainhealthinformationsystem(HIS).” International Conference on Smart Homes and HealthTelematics. Springer, Cham, 2016.
6. Rallapalli, Sreekanth, and T. Suryakanthi. “Predicting the risk ofdiabetes in big data electronic health Records by using scalablerandom forest classification algorithm.” 2016 InternationalConference on Advances in Computing and CommunicationEngineering (ICACCE). IEEE, 2016.
7. Sudhakar, K., and Dr M. Manimekalai. “Study of heart diseaseprediction using data mining.” International journal of advancedresearch in computer science and software engineering 4.1 (2014):1157-1160.
8. Chitra,R.,andV.Seenivasagam.“Reviewofheartdiseasepredictionsystem using data mining and hybrid intelligent techniques.”ICTACT journal on soft computing 3.04 (2013): 605-09.
9. Liu, Jenn-Long, Yu-Tzu Hsu, and Chih-Lung Hung. “Developmentof evolutionary data mining algorithms and their applications tocardiac disease diagnosis.” 2012 IEEE Congress on EvolutionaryComputation. IEEE, 2012.
10. Jadhav, Shivajirao M., Sanjay L. Nalbalwar, and Ashok A. Ghatol.“Artificial neural network models based cardiac arrhythmia diseasediagnosisfromECGsignaldata.”InternationalJournalofComputerApplications 44.15 (2012): 8-13.
11. Singh, Shraddha, et al. “Classification of ECG arrhythmia usingrecurrent neural networks.” Procedia computer science 132 (2018):1290-1297.
12. Karandikar, M., and G. Guidi. “Classification of arrhythmia usingECG data.” Proc. CS-Fall. 2014.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details