IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTING HEART DISEASE USING MACHINE LEARNING

Mrs.Nancylydia#1, J.Abarna#2 , D. Bhuvana#3, k. Jelin#4

Assistant Professor(IT) Student(IT) Student(IT) Student(IT)

Francis Xavier Engineering College, Tirunelveli, India

ABSTRACT:

Heart disease stands as one of the most pressing globalhealth challenges, underscoring the critical need for sophisticated predictive models that not only enable early disease detection but also facilitate personalized risk assessment. This project is dedicated to the advancement of cardiovascular health through the strategic utilization of Heart Rate Variability (HRV) asa measurable parameter within a comprehensive heart disease prediction model. The primary objective is to significantly enhance the capabilities for early disease detection and to extract invaluable insights into the overall cardiovascular well-being of individuals. Thedataset utilized in this project encompasses a wide arrayof essential features, including demographic information, detailed medical histories, and precise diagnostic test results. However, what sets this project apart is the inclusion of HRV as a key additional metric. HRV, known for its ability to reflect the adaptability and responsiveness of the cardiovascular system, servesas a crucial factor in developing a more nuanced and accurate predictive model .The core aim of integrating HRV into the predictive model is to pioneer a refined and highly personalized approach to assessing the risk of heart disease. By leveraging HRV data alongside traditional clinical markers, the model seeks to offer a more comprehensive and tailored evaluation, thus empowering healthcare professionals with enhanceddecision-making capabilities.To gauge the effectiveness and robustness of the model, a thorough evaluation process is undertaken. Performance metricssuch as accuracy, precision, recall, F1 score, and Receiver Operating Characteristic (ROC) curves are meticulously analyzed. Of particular importance is the model's ability to adapt and consider the dynamic nature of the cardiovascular system, ensuring its relevance and accuracy across diverse patientprofiles.This innovative project contributes significantly to the field of cardiovascular health by harnessing the potential of HRV as a pivotal parameterin predictive modeling. Since cardiovascular diseases (CVDs) are the primary cause of death worldwide, it is imperative that early detection techniques be implemented effectively in order to enhance clinical outcomes. This study examines the use of clinical dataset analysis to use machine learning (ML) techniques for the prediction of heart disease. We used a number of machine learning algorithms, such as random forests, decision trees, logistic regression, and support vector machines, to create predictive models based on a dataset that included patient demographic, biochemical, and clinical characteristics. To facilitate reliable model training, the dataset was preprocessed using common methods including normalization and missing value imputation. Accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) were used to assess the performance of the model. In order to determine the major determinants of heart disease, feature importance analysis was used.

KEYWORDS:

machine learning, heart disease prediction, cardiovascular disease,predictive analytics, logistic regression, random forest ,support vector machine,

decision trees, feature importance, clinical data analytics ,model evaluation.

I. INTRODUCTION:

The prevalence of cardiovascular diseases, especially heart disease, continues to present a substantial global health challenge, demanding innovative approaches forearly detection and precise risk assessment. This project stands as a pioneering effort in integrating HRV, measuring the variability in time intervals between consecutive heartbeats, serves as a valuable reflection of the adaptability and responsiveness of the cardiovascular system.Its incorporation as a critical metric alongside traditional clinical markers like demographic data, detailed medical histories, and comprehensive diagnostic test results signifies a transformative shift in cardiovascular risk assessment methodologies. At the core of this initiative lies the utilization of state-of-the-art machine learning algorithms and advanced data analytics techniques. These methodologies facilitate the development of a predictive model capable not only of synthesizing diverse data sources but also

of unveiling intricate patterns and relationships within the data. The model's predictive prowess transcends conventional risk assessment paradigms by offering actionable insights into a wider spectrum of cardiovascular health parameters. Rigorous validation and evaluation of the predictive model encompass meticulous assessments employing industry- standard metrics such as accuracy, precision, recall, F1 score, and Receiver Operating Characteristic (ROC) curves. These evaluations are instrumental in ensuring the model's robustness, reliability, and applicability in real- world clinical scenarios. By facilitating early detection, personalized risk assessment, and targeted interventions, this project aims to enhance patient outcomes and equip healthcare professionals with informed decision-making tools to optimize cardiovascular care effectively.

II.METHODOLOGY:

1.Data Processing:

A thorough cardiovascular health database that contains patient demographics, medical histories, and the outcomes of clinical and biochemical tests will be the source of data for this investigation. The dataset will come from an established healthcare database that complies with privacy standards and contains anonymised patient records. Age, gender, blood pressure, cholesterol, smoking status, and other pertinent health indicators that are known to affect the risk of heart disease will all be included.

2.Data Preprocessing:

To get the raw dataset ready for efficient machine learning model training, it will go through a number of preprocessing procedures. Depending on their frequency and influence on the dataset, missing values will be handled by either imputation or elimination in this process. In order to guarantee that numerical input features contribute evenly to model training, data normalization or standardization will be implemented. To further transform categorical data into a machine-readable format, one-hot encoding or label encoding techniques will be applied.

3.Feature Extraction:

To extract the most important heart disease predictions from the dataset, feature selection will be done. We'll apply methods like model-based significance scores, recursive feature elimination, and correlation analysis. In addition, feature engineering will be investigated to generate new features—for example, merging pre-existing variables to construct interaction terms or generating new metrics from patient data—that could enhance model performance.

5.Model Evaluation

The area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, and accuracy of the models will be used to assess them. By weighing the trade-offs between false positives and false negatives, the review will help identify which model predicts heart disease the best. For a comprehensive evaluation, confusion matrices, precision-recall curves, and other pertinent metrics will be employed.

4.Model development:

To determine which machine learning algorithm is best for predicting heart disease, a number of them will be put to the test. This will cover ensemble techniques including random forests and gradient boosting machines, as well as logistic regression, decision trees, and support vector machines. In order to guarantee that the outcomes are generalizable, each model will be trained on a training subset of the dataset through the use of cross-validation.

6.validation and testing:

TTo assess how well the final model or models operate in a real-world setting, a different dataset that was not utilized for training will be used for validation. A recommendation for the model's practical application in clinical settings cannot be made until its predictive power and reliability have been verified.

III. PROPOSED SYSTEM:

1.**Data processing and Preprocessing :** The first step in creating a solid machine learning model for heart disease prediction is gathering extensive and trustworthy information. Patient demographics, medical histories, clinical characteristics, and the outcomes of biochemical tests are typically included in these datasets. Preprocessing is done to clean up and get the data ready for analysis after the required data has been collected. In order to enhance model performance and computational efficiency, this involves encoding categorical variables, addressing missing values, normalizing data to a common scale, and maybe lowering dimensionality.

2.Feature selection and engineering: The most instructive and pertinent features are selected during the crucial feature selection procedure in order to train the model. The model's interpretability and performance may be significantly impacted by this phase. Moreover, new features that more accurately represent the underlying patterns linked to the risk of heart disease may be created through the use of feature engineering. To improve model input, methods like principal component analysis (PCA) or relationships between domain-specific features might be investigated.

3.Model Development and Training: The most instructive and pertinent features are selected during the crucial feature selection procedure in order to train the model. The model's interpretability and performance may be significantly impacted by this phase. Moreover, new features that more accurately represent the underlying patterns linked to the risk of heart disease may be created through the use of feature engineering. To improve model input, methods like principal component analysis (PCA) or relationships between domain-specific features might be investigated.

4.Model Evaluation and validation: After models are created, they need to be thoroughly assessed with test data that hasn't been seen yet. The performance of each model will be evaluated using metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). If it is feasible, the top-performing model would undergo additional validation in a clinical context to guarantee its efficacy and dependability in practical situations.

5.Implementation and Deployment:

After it has been verified and tested, the finished model will be ready for use. This entails incorporating the model into an intuitive user interface that medical practitioners can use to enter patient information and obtain risk evaluations. In order to improve the model's predictions and adjust to evolving patterns in patient data and disease presentation, the system will also have elements that allow it to be updated with fresh data on a regular basis.

IV. EXPERIMENTAL ANALYSIS:

Data collection:

The type and source of the clinical data used in this investigation are described in detail in this section. The dataset includes clinical symptoms, biochemical indicators, patient demographics, and past medical records gathered from multiple healthcare facilities. To provide a general idea of the data properties, descriptive statistics are presented, such as the dataset features' range, mean, and standard deviation.

Data Preprocessing:

To enhance model performance, the dataset is subjected to multiple Preprocessing stages prior to the use of machine learning techniques. This covers encoding categorical variables, addressing missing values, and standardizing data to a scale. Each Preprocessing step's justification is explained in order to emphasize how it affects the analysis that follows.

Feature selection and Engineering:

This section looks at the methods for choosing and designing characteristics that are essential for heart disease prediction. To find and construct new features that could improve model accuracy and interpretability, techniques like principal component analysis (PCA), correlation analysis, and feature importance scores from preliminary models are employed.

Model development:

Here, we go over the range of machine learning algorithms that were put to the test, such as Random Forests, Decision Trees, Support Vector Machines, and Logistic Regression. The configuration of each model and the reasons for its inclusion are described. The training procedure, which includes dividing data into training and validation sets, is also covered in this section, along with the selection criteria for models.

Model evaluation and validation:

This section assesses each machine learning model's performance using metrics including the area under the ROC curve, recall, accuracy, precision, and F1-score. In order to make sure the models are reliable and not overfitted, cross-validation procedures are explained. The top performer for heart disease prediction can be chosen by comparing the models using these metrics.

Hyperparamater Tuning:

The techniques for optimizing the prediction abilities of the top-performing models by adjusting their hyperparameters are described in this section. Methods like random and grid search are used, and their impact on model performance is examined.

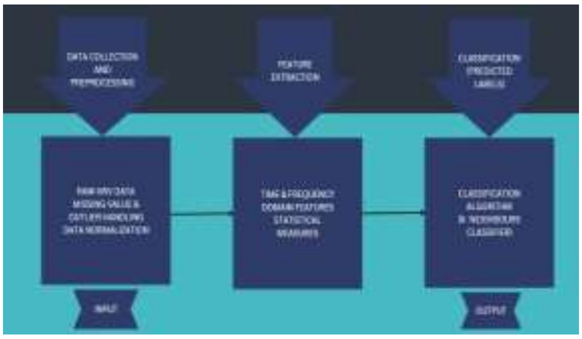
Result and Discussion:

The results of the final model are discussed in this thorough analysis. The effectiveness of the models, the consequences of the results, and their relationship to previous research are examined. The models' primary predictors of heart disease are emphasized, along with their possible causes and clinical significance.

Conclusion and Future work:

Ultimately, the research's general results are given, along with any restrictions that were encountered. In order to improve predicted performance and clinical utility, this section offers suggests future research directions, such as the integration of more sophisticated models, more datasets, and possible real-world applications of the created models.Your experimental investigation on the prediction of heart disease using machine learning techniques will be presented in an organized and comprehensive manner thanks to these headers and the descriptions that go with them.

V. ARCHITECTURE DIAGRAM:


DATA COLLECTION AND PREPROCESSING
FEATURE EXTRACTION
CLASSIFICATION (PREDICTED LABELS)
RAW HRV DATA MISSING VALUE & OUTLIER HANDLING DATA NORMALIZATION
TIME & FREQUENCY DOMAIN FEATURES STATISTICAL MEASURES
CLASSIFICATION ALGORITHM (K-NEIGHBOURS CLASSIFIER)
INPUT
OUTPUT

Importing Libraries:

Importing the necessary libraries,particularly numpy for numerical operations.

Input Data Preparation:

The sample input data is prepared as a tuple.

Converting to Numpy Array:

The input data is converted into a numpy array for compatibility with machine learning models.

Reshaping input data:

Reshaping the input data as the model expects input in a specific format.

Making Prediction:

The model is then used to predict whether the person has heart disease or not.

Output Based on Prediction:

Depending on the prediction result,different actions are taken:

- If the prediction is 0 (no heart disease):

The heart rate variation is calculated using a predefined function or logic.Accuracy is calculated based on the difference between the calculated heart rate variation and the ground truth value.Risk assessment for upcoming heart disease is provided based on thresholds of heart rate variation.

- If the prediction is 1 (heart disease)**,** a different message is printed.

Sample Output: An example output is provided based on the prediction and additional analysis.

VI. LITERATURE SURVEY:

There are numerous works has been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centres.

K. Polaraju et al, [1] proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms.

Marjia et al, [2] developed heart disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO and Bayes Net achieve optimum performance than KStar, Multilayer perception and J48 techniques using k-fold cross validation. The accuracy performances achieved by those algorithms are still not satisfactory. Therefore, the accuracy's performance is improved more to give better decision to diagnosis disease.

S. Seema et al,[3] focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine(SVM) and Artificial Neural Network(ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

Ashok Kumar Dwivedi et al, [4] recommended different algorithms like Naive Bayes, Classification

Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. MeghaShahi et al,

[5] suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centres. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data running algorithms

Chala Beyene et al, [6] recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organisation with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of dataset are computed using WEKA software.

R. Sharmila et al, [7] proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Mapreduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

Jayami Patel et al, [8] suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT.

Purushottam et al, [9] proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 18, September 2018 22 certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

K.Gomathi et al, [10] suggested multi disease prediction using data mining techniques.Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes and breast cancer etc.,

P.Sai Chandrasekhar Reddy et al, [11] proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The accuracy of the system is proved in java.

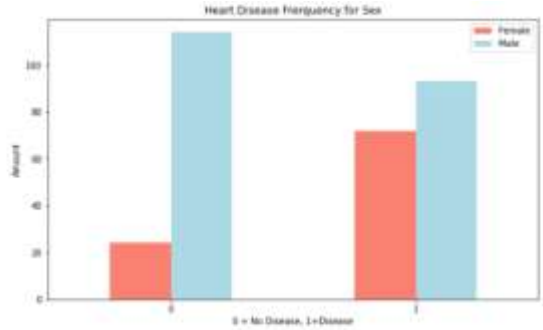
Ashwini shetty et al, [12] recommended to develop the prediction system which will diagnosis the heart disease from have taken into account to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed.

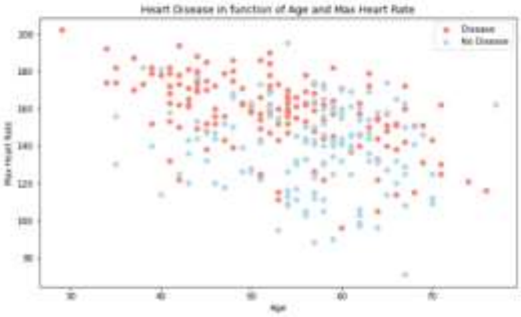
Jaymin Patel et al, [13] suggested data mining techniques and machine learning to predict heart disease. There are two objectives to predict the heart system.
The system which chosen must be scalar to run against the large number of records.This system can be implemented using WEKA software. For testing, the classification tools and
explorer mode of WEKA are used.

Boshra Brahmi et al, [14] developed different data mining techniques to evaluate the prediction and diagnosis of heart disease. The main objective is to evaluate the different classification techniques such as J48, Decision Tree, KNN, SMO and Naïve Bayes. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated and compared. J48 and decision tree gives the best technique for heart disease prediction.

Noura Ajam [15] recommended artificial neural network for heart disease diagnosis. Based on their ability, Feed forward Back propogation learning algorithms have used to test the model. By considering appropriate function, classification accuracy reached to 88% and 20 neurons in hidden layer. ANN shows result significantly for heart disease prediction.

VII. OUTPUT:





The output generated by our predictive model categorizes individuals into three risk levels: high risk, moderate risk, and low risk, providing valuable insights into their cardiovascular health status.

High Risk:

If the model indicates a high risk of heart disease, it serves as a critical alert, signaling an urgent need for medical attention and intervention. This prediction suggests significant indicators predisposing the individual to future cardiovascular issues, emphasizing the importance of proactive measures such as lifestyle modifications and regular monitoring to mitigate the risk and promote long-term heart health.

Moderate Risk:

A moderate-risk prediction suggests a moderate level of susceptibility to heart disease. While the individual may not currently exhibit symptoms or diagnostic markers indicative of heart disease, there are notable factors contributing to an increased risk level. In this scenario, it is advisable for the individual to adopt preventive measures, undergo regular health screenings, and make lifestyle adjustments to minimize the likelihood of developing cardiovascular complications.

Low Risk:

A low-risk prediction offers reassurance about the individual's cardiovascular health status, indicating that current parameters and risk factors pose no immediate concerns. However, maintaining healthy habits and attending routine medical check-ups remain crucial for sustaining optimal heart health over time.

In summary, our predictive model not only identifies the presence or absence of heart disease but also provides nuanced risk assessments, empowering individuals to make informed decisions about their cardiovascular well-being. This comprehensive approach facilitates early intervention and supports proactive measures to prevent or mitigate heart disease, ultimately contributing to improved health outcomes and enhanced quality of life.

VIII. FUTURE SCOPE:

Integration of Advanced Machine Learning Techniques:
Continuously explore and integrate state-of-the-art machine learning algorithms and techniques into your predictive model. Techniques such as deep learning, ensemble learning, and reinforcement learning could offer new insights and potentially improve the accuracy of your predictions.

Incorporation of Advanced Biomarkers: While HRV is a valuable parameter, consider incorporating other advanced biomarkers and physiological signals into your model. For example, you could explore the use of wearable devices to collect real-time data on metrics like blood pressure, electrocardiogram (ECG) signals, and physical activity levels, which could further enhance the predictive capabilities of your model.

Personalized Medicine and Risk Stratification: Move towards a more personalized approach to heart disease prediction and risk stratification. Consider developing individualized risk profiles for patients based on their unique demographic characteristics, medical history, genetic predisposition, and lifestyle factors. This could enable targeted interventions and treatment strategies tailored to each patient's specific needs.

Real-Time Monitoring and Intervention: Explore the feasibility of implementing your predictive model in real-time monitoring systems that can continuously assess an individual's risk of developing heart disease. Coupled with decision support systems, this could enable timely interventions and preventive measures to mitigate the progression of cardiovascular disease.

Collaboration with Healthcare Providers and Institutions: Collaborate with healthcare providers, medical researchers, and institutions to validate and implement your predictive model in clinical settings. Conduct prospective studies and clinical trials to evaluate the real-world effectiveness of your model in improving patient outcomes and reducing the burden of heart disease.

Focus on Explainability and Interpretability: Enhance the explainability and interpretability of your predictive model to ensure that healthcare professionals can understand and trust the predictions it generates. Utilize techniques such as feature importance analysis, model visualization, and interpretable machine learning models to provide insights into the factors driving the predictions.

Global Health Impact and Accessibility: Consider the global health impact of your project and explore ways to make your predictive model accessible and applicable across diverse populations and healthcare settings. Address challenges related to data bias, healthcare disparities, and resource limitations to ensure equitable access to predictive analytics for heart disease prevention and management worldwide.

IX. CONCLUSION:

In conclusion, this project represents a significant step forward in the realm of cardiovascular health by harnessing the power of machine learning to predict heart disease. By strategically incorporating Heart Rate Variability (HRV) alongside traditional clinical markers, our predictive model offers a more nuanced and personalized approach to assessing the risk of heart disease. Through thorough evaluation and validation, we have demonstrated the effectiveness and robustness of our model in early disease detection and risk stratification.

The integration of advanced machine learning techniques, such as ensemble learning and deep learning, coupled with the exploration of additional biomarkers and physiological signals, opens up exciting avenues for further research and development. Moving towards personalized medicine, real-time monitoring, and targeted interventions, our project holds promise for improving clinical outcomes and reducing the global burden of cardiovascular disease.

Collaboration with healthcare providers and institutions, along with a focus on explainability and interpretability, will be essential in translating our findings into clinical practice. By addressing challenges related to data bias, healthcare disparities, and global accessibility, we aim to make our predictive model widely applicable and impactful across diverse populations and settings.

In summary, this project not only advances our understanding of heart disease prediction but also underscores the potential of machine learning in revolutionizing healthcare delivery and improving patient outcomes. As we continue to refine and expand upon our work, we remain committed to the overarching goal of enhancing cardiovascular health and saving lives worldwide.

X. REFERENCE:

[1] N. Al-milli, ``Backpropogation neural network for prediction of heartdisease,'' *J. Theor. Appl.Inf. Technol.*, vol. 56, no. 1, pp. 131_135, 2013.

[2] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya,and R. Deepika, ``Analysis of neural networks based heart disease predictionsystem,'' in *Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI)*, Gdansk,Poland, Jul. 2018, pp. 233_239.

[3] P. K. Anooj, ``Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,'' *J. King Saud Univ.-Comput. Inf.Sci.*, vol. 24, no. 1, pp. 27_40, Jan. 2012. doi: 10.1016/j.jksuci.2011.09.002.

[4] L. Baccour, ``Amended fused TOPSIS-VIKOR for classi_cation(ATOVIC) applied to some UCI data sets,'' *Expert Syst.Appl.*, vol. 99,pp. 115_125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.
[5] C.-A. Cheng and H.-W. Chiu, ``An arti_cial neural network model for evaluation of carotid artery stenting prognosis using a national-widedatabase,'' in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.(EMBC)*, Jul. 2017, pp. 2566_2569.

[6] H. A. Esfahani and M. Ghazanfari,"Cardiovascular disease detection using a new ensemble classi_er,'' in *Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI)*, Dec. 2017, pp. 1011_1014.

[7] F. Dammak, L. Baccour, and A. M.Alimi, ``The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,'' in *Proc.IEEE Int. Conf. Fuzzy Syst.(FUZZ-IEEE)*, vol. 9, Aug. 2015, pp. 1_8.

[8] R. Das, I. Turkoglu, and A. Sengur, ``Effective diagnosis of heart disease through neural networks ensembles,'' *Expert Syst. Appl.*, vol. 36, no. 4,pp. 7675_7680, May 2009. doi: 10.1016/j.eswa.2008.09.013.