JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CARDIO DYSFUNCTION PREDICTOR USING MACHINE LEARNING

[1]Sneha Narayahni.SB[1], [2]Mahalakshmi.L[2] , [3]Mrs.B. Arunmozhikalanchiyam[3]
Department of Information Technology, Meenakshi Engineering College , Tamil Nadu
(Affiliated to Anna University Tamil Nadu)

Abstract :

Heart disease of the most significant causes of morality in the world today . Prediction of cardio vascular disease is a critical challenge in the area of clinical data analysis.Machine learning(ML)has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry.

In order to focus the research on real-world datasets rather than merely theoretical approaches and simulations, it would be highly desired to extend this work further. The suggested hybrid HRFLM approach combines the benefits of the Linear Method (LM) and Random Forest (RF) techniques. When it came to predicting heart disease, HRFLM turned out to be fairly accurate.

The connection between mental health and heart health is significant and multifaceted.Strategies such as therapy, stress management techniques, regular physical activity, and social support can all play a role in improving mental health and reducing the risk of heart disease.

I. INTRODUCTION

The main goal of the system is to use machine learning (ML) techniques to advance the field of cardiovascular disease prediction. The goal is to improve the efficiency and accuracy of heart disease prediction by detecting important elements from unprocessed healthcare data. Mental health conditions like sadness and anxiety might make it more likely for people to take prescribed medications as directed or to start unhealthy habits like smoking or leading a sedentary lifestyle. This is because individuals with mental health disorders may find it more difficult to adopt healthy lifestyle choices that lower their risk of heart disease since they may have fewer effective coping mechanisms for stressful situations.

In order to optimize prediction models and provide healthcare practitioners with useful insights for patient care, the system places a strong emphasis on the identification of key features and the integration of several machine learning approaches. By achieving these goals, the suggested method hopes to significantly increase the accuracy of heart disease prediction, which will eventually improve patient care and cardiovascular health outcomes. The system also aims to address the requirement for adaptability and scalability in managing real-world healthcare datasets, encouraging the creation of predictive models that are more precise and flexible.

II. EXISTING SYSTEM

Globally, heart disease is the leading cause of death. Owing to growing complexity, it is imperative that diagnoses be made more quickly in order to provide patients with quality care. Long-term sufferers of depression, anxiety, stress, and even PTSD may notice physiological changes in their bodies, including elevated cortisol levels, decreased blood supply to the heart, and elevated cardiac reactivity (heart rate and blood pressure). These physiological changes have the potential to cause heart disease, metabolic disorders, and artery-clogging calcium accumulation over time.

Traditional methods take a lot of time, and accuracy is a big issue. Accurate and quick prediction is the main challenge. Techniques for machine learning can assist with this issue. It is not necessary to have all of the characteristics in order to forecast

heart diseases. The dataset's redundant and superfluous features will be eliminated using the evolutionary algorithm, which will also identify the optimal feature combination for increased prediction accuracy.

When compared to other conventional algorithms with GA, SVM provides superior accuracy. This paper shows that utilizing GA for feature selection and SVM for classification, it would be possible to improve the accuracy of cardiovascular disease diagnosis.

III. PROPOSED SYSTEM

Predicting heart disease is a difficult but crucial task in medicine.However, if the condition is discovered early and preventative measures are taken as soon as feasible, the death rate can be significantly reduced. In order to focus the research on real-world datasets rather than merely theoretical approaches and simulations, it would be highly desired to extend this work further.

The suggested hybrid HRFLM approach combines the benefits of the Linear Method (LM) and Random Forest (RF) techniques. When it came to predicting heart disease, HRFLM turned out to be fairly accurate.Understanding how to interpret unprocessed medical data about the heart can save lives in the long run and aid in the early identification of irregularities in cardiac diseases.

In this experiment, raw data was processed and cardiac disease was identified using machine learning techniques. Predicting heart disease is a difficult but crucial task in medicine.Nonetheless, if the illness is identified in its early stages and preventative actions are taken as soon as feasible, the death rate can be significantly reduced. In order to focus the research on real-world datasets rather than merely theoretical approaches and simulations, it would be highly desired to extend this work further.

IV. SYSTEM SPECIFICATION

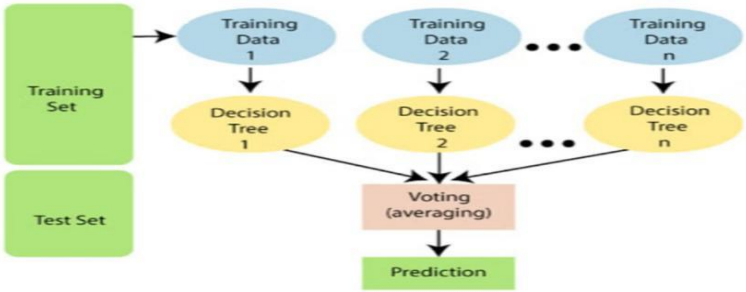
HARDWARE CONFIGURATION:

Processor	-	I5
Speed	-	3 GHz
RAM	-	8 GB(min)
Hard Disk	-	500 GB
Key Board	-	Standard Windows Keyboard
Mouse	-	Two or Three Button Mouse
Monitor	-	LCD

SOFTWARE CONFIGURATION

Operating System: Linux, Windows/7/10
Server: Anaconda, Jupyter,pycharm
Front End: tkinter |GUI toolkit
Server side Script: Python , AIML

V. RANDOM FOREST AND LINEAR METHOD TECHNIQUES



The following steps explain the working Random Forest Algorithm:

Step 1: Select random samples from a given data or training set.

Step 2: This algorithm will construct a decision tree for every training data.

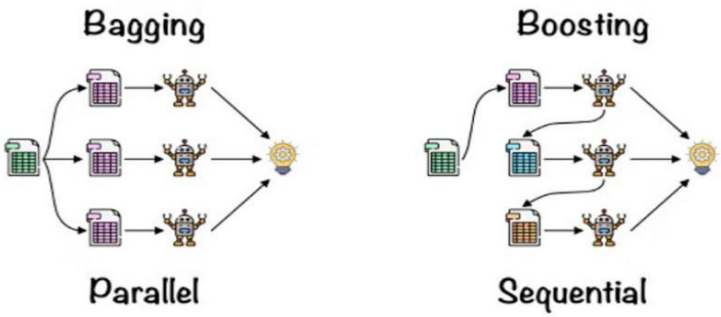
Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.

This combination of multiple models is called Ensemble. Ensemble uses two methods:

Bagging: Creating a different training subset from sample training data with replacement is called Bagging. The final output is based on majority voting.

Boosting: Combing weak learners into strong learners by creating sequential models such that the final model has the highest accuracy is called Boosting. Example: ADA BOOST, XG BOOST.



Bagging: From the principle mentioned above, we can understand Random forest uses the Bagging code. Now, let us understand this concept in detail. Bagging is also known as Bootstrap Aggregation used by random forest. The process begins with any original random data. After arranging, it is organised into samples known as Bootstrap Sample. This process is known as Bootstrapping.Further, the models are trained individually, yielding different results known as Aggregation. In the last step, all the results are combined, and the generated output is based on majority voting. This step is known as Bagging and is done using an Ensemble Classifier.

Key Elements of Random Forest:

Random: Every tree has a distinct quality, range, and characteristic in relation to other trees. Trees differ from one another.The curse of dimensionality does not apply to trees because they are conceptual concepts and do not need characteristics to be taken into account. As a result, there is less feature space.

Parallelization: Since each tree is built independently from distinct data and features, we can produce random forests by utilizing the entire CPU.

Train-Test split: Since the decision tree in a Random Forest never sees 30% of the input, we don't need to separate the data for training and testing.

Stability: The outcome is determined by bagging, which uses the average or majority vote to determine the outcome.

By analyzing unprocessed medical data, the suggested algorithm, Hybrid HRFLM (Random Forest and Linear Method), seeks to improve the forecast accuracy of heart disease. In order to evaluate and identify trends in the data, this program makes use of machine learning techniques, which could aid in early diagnosis and even save lives.

With the knowledge that early detection and prompt preventative interventions can effectively limit death rates, the significance of precise cardiac disease prediction in the medical industry is underscored.The goal of the Hybrid HRFLM strategy is to combine the advantages of the Linear Method (LM) and Random Forest (RF) techniques.

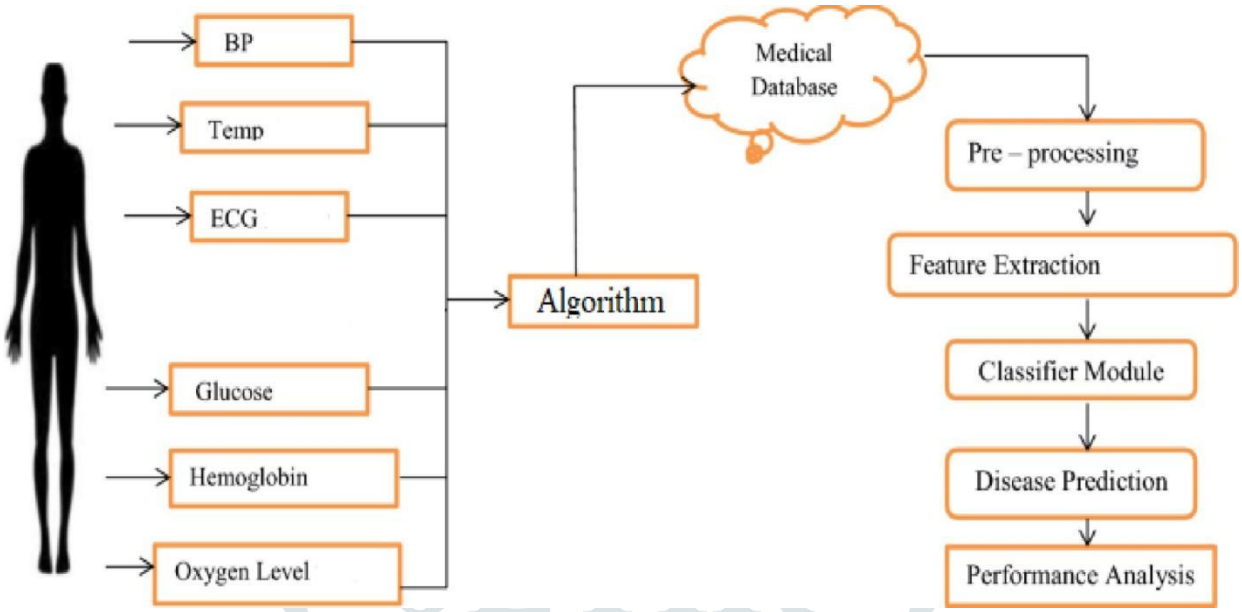
The goal of this hybridization is to provide a more thorough and precise predictive model for heart disease by leveraging the variety of machine learning approaches. The algorithm's potential to improve cardiovascular health outcomes has been proved by its capacity to produce accurate forecasts, demonstrating its usefulness. One of the proposed algorithm's benefits is that it makes use of a variety of machine learning techniques, which strengthens prediction methodology.

The algorithm also concentrates on finding important elements in the data, which eventually improves the accuracy of heart disease prediction models. Consequently, the Hybrid HRFLM algorithm is a promising development in the field of cardiovascular health predictive analytics, with the potential to have a major influence on patient outcomes and clinical practice.

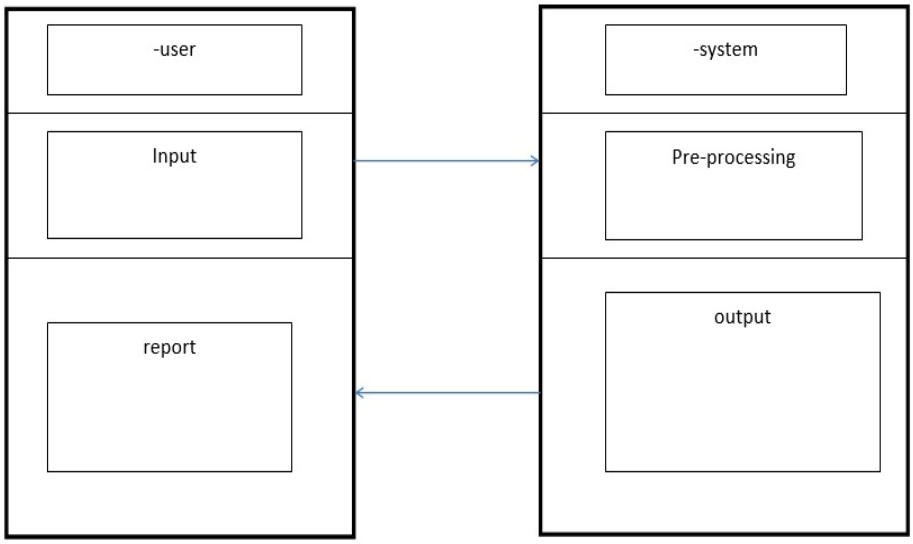
5.1 ADVANTAGE OF PROPOSED ALGORITHM

- **Diverse Machine Learning Techniques:** The algorithm leverages a hybrid approach by combining the strengths of Random Forest (RF) and Linear Method (LM).
- **Improved Prediction Techniques:** By integrating different characteristics of RF and LM, the algorithm aims to overcome the limitations of individual methods, providing a synergistic effect that enhances the overall prediction accuracy.
- **Identification of Significant Features:** The algorithm focuses on identifying and utilizing significant features within the raw healthcare data.
- **Increased Performance in Heart Disease Prediction:** Through the combination of RF and LM characteristics, the Hybrid HRFLM algorithm has demonstrated efficacy in accurately predicting heart disease.
- **Potential for Real-world Application:** The algorithm's success in simulations suggests its potential applicability to real-world datasets.

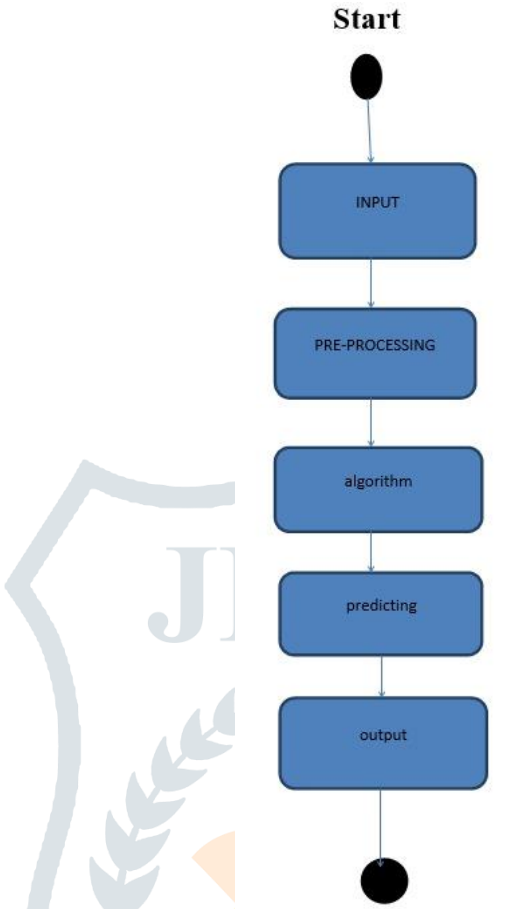
VI. ARCHITECTURE DIAGRAM



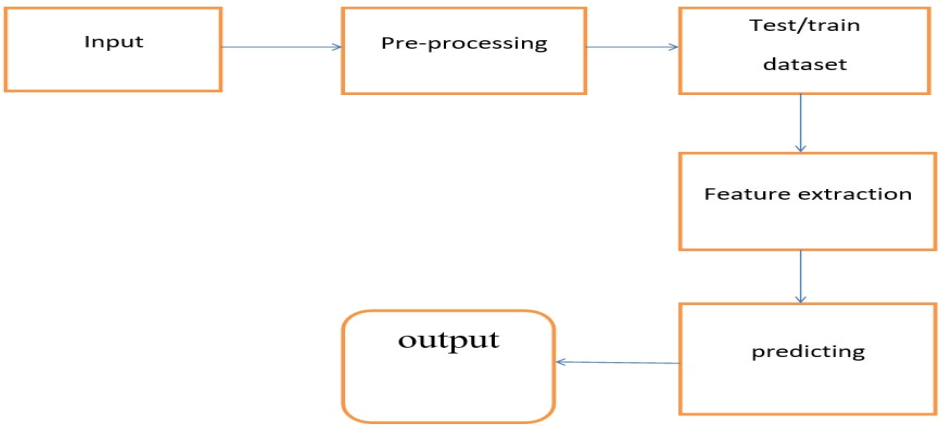
6.1 CLASS DIAGRAM



6.2 ACTIVITY DIAGRAM



6.3 DATA FLOW DIAGRAM



DATASET MODULES

A dataset is a group of data that has been organized into a specific order. Any type of data, from an array series to a database table, can be included in a dataset. A tabular dataset is similar to a database table or matrix, where each row represents the dataset's fields and each column represents a specific variable. "Comma Separated File," or CSV, is the file type that is most commonly supported for tabular datasets. However, we can use the JSON file more effectively to store a "tree-like data." Data types found in datasets

Numerical data: Temperature, property price, etc..
Categorical data: such True/False, Yes/No, Blue/Green, etc.

Ordinal data: These can be measured through comparison, but they are comparable to categorical data.
Requirement for a dataset
We require a vast amount of data to work on machine learning projects because ML/AI models cannot be trained without it. When developing an ML/AI project, gathering and preparing the dataset is one of the most important steps.If the dataset is not correctly prepared and pre-processed, the technology used in any machine learning project will not function as intended.The datasets are the only resource available to the developers while they work on the machine learning project. When creating machine learning apps, datasets are separated into two categories:
Training Dataset
Test Collection

VII. IMPLEMENTAION

There are various processes involved in putting a machine learning-based heart disease prediction system into practice. This is a high-level summary of the procedure:

Data collection: Assemble a dataset with pertinent characteristics (blood pressure, cholesterol levels, age, sex, etc.) and labels designating whether or not each person has heart disease. The Framingham Heart Study dataset, the Cleveland Heart Disease dataset, and other common datasets are used for this work.

Data Preprocessing: Clean up the data, deal with any missing values, and prepare it so that machine learning algorithms can use it. Normalization, feature scaling, and categorical variable encoding may be required for this.
Feature engineering and selection: Examine the dataset to determine which features are most crucial for heart disease prediction. To choose pertinent features or generate new features, you can employ strategies like feature importance ranking, correlation analysis, and domain expertise.
Selecting a Model: Select the right machine learning methods for the categorization. Neural networks, support vector machines (SVM), decision trees, random forests, and logistic regression are popular options. Try out various models and see which one works best for your dataset.
Training Models: Divide the dataset into sets for testing and training. Utilizing the training data, train the chosen machine learning models.
Model Evaluation: Use measures like F1-score, ROC-AUC, accuracy, precision, and recall to assess each model's performance. Select the model that performs the best on the testing set.
Hyperparameter tuning: To further enhance the selected model's performance, adjust its hyperparameters. Hyperparameter optimization can be accomplished by methods such as random or grid search.
Implementation: After your model has been trained and fine-tuned, apply it to a practical application. This could be an integrated part of a healthcare system or a web or mobile application.
Monitoring and Maintenance: To guarantee the deployed model's accuracy and applicability throughout time, keep a close eye on its performance and update it with fresh data on a regular basis.
Moral Aspects to Take into Account: Throughout the development and deployment process, keep in mind ethical considerations like prejudice, fairness, and data privacy.
To create a precise and trustworthy heart disease prediction system, it's critical to record your procedure, try out various strategies, and refine your plan accordingly.

VIII. DOMAIN OF THE PROJECT

PYTHON

Python is a high-level, interpreter-based, object-oriented programming language featuring dynamic semantics. Its dynamic typing and dynamic binding, along with its high-level built-in data structures, make it an appealing language for Rapid Application Development and for usage as a scripting or glue language to join existing components. Because of its straightforward, basic syntax, Python promotes readability, which lowers software maintenance costs. Python's support for packages and modules promotes code reuse and program modularity. The large standard library and the Python interpreter are freely distributable and accessible for free on all major platforms in source or binary form.

Python's increased efficiency is one of the main reasons programmers fell in love with it. The edit, test, and debug cycle is extremely quick because there is no compilation step. Python program debugging is simple because segmentation faults are never caused by bugs or incorrect input. Rather, the interpreter raises an exception when it finds a mistake. The interpreter prints a stack trace if the application fails to catch the exception. Setting breakpoints, evaluating arbitrary expressions, inspecting local and global variables, stepping through the code one line at a time, and other features are all possible with a source level debugger. The fact that the debugger is developed in Python attests to the language's capacity for introspection.

However, adding a few print statements to the source code is frequently the fastest way to debug a program since it creates a short edit-test-debug cycle. This straightforward method is also highly effective.It includes anything from basic automated jobs to web development, gaming, and even sophisticated corporate systems.

On the other hand, adding a few print statements to the code is frequently the fastest way to debug a program because of how quickly the edit-test-debug cycle may be completed.Simple automation jobs, web development, games, and even sophisticated enterprise systems are all included.

PYCHARM

An easy-to-use environment for efficient Python, web, and data science development is created by PyCharm, an IDE specifically designed for Python developers. It offers a large selection of necessary tools for Python developers.

For details, see the editions comparison matrix.

Supported languages

To start developing in Python with PyCharm you need to download and install Python from python.org depending on your platform.

PyCharm supports the following versions of Python:

Python 2: version 2.7

Python 3: from the version 3.6 up to the version 3.10

Additionally, the Professional edition allows for the development of Pyramid, Flask, and Django applications. Additionally, it supports HTML (including HTML5), CSS, JavaScript, and XML to the fullest extent possible. These languages are included in the IDE by default and are packed with plugins. Plugins can be used to add support for additional languages and frameworks; to learn more about them or set them up on the first IDE run, go to Settings | Plugins or PyCharm | Preferences | Plugins for macOS users.

IX. RESULT ANALYSIS

For the purpose of predicting cardiac disease, machine learning methods such as Random Forest can provide insightful results that may even enhance patient outcomes. Here's a summary of possible ways to interpret the findings:

Accuracy Metrics: Start by assessing the Random Forest model's accuracy. Metrics like recall, accuracy, precision, and F1-score can be used to evaluate how well the model works in accurately identifying heart disease cases. Additionally, to make sure the results are robust, you might employ methods like cross-validation.

Feature Importance: Random Forest offers a feature importance score that shows how much each feature (or prediction) influences the way the model makes decisions. Finding the most important factors in heart disease prediction can be aided by feature importance analysis. Understanding the underlying correlations between variables and setting priorities for treatments or additional research are made easier with the help of this knowledge.

Visualization: Understanding how the Random Forest's decision trees are represented might help one better understand how the model generates predictions. The links between predictors and the result variable can be better understood with the use of techniques like partial dependence plots and tree visualization.

Interpretability of the Model: Random Forest models are sometimes regarded as "black-box" models, but methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) values can shed light on specific predictions and improve the model's interpretability.

Validation and Generalization: To make sure the model performs well when applied to previously unseen data, evaluate its performance on a separate dataset (validation set). Validating the model's performance on data that it hasn't seen during training is crucial because overfitting is a major problem in machine learning.

Clinical Relevance: Lastly, it's critical to assess the model's predictions' clinical relevance. To confirm that the model's conclusions align with clinical recommendations and medical knowledge, think about collaborating closely with healthcare specialists. Evaluate the model's usability and potential advantages for improving patient care as well as the possible effects of applying it to actual clinical situations.

To guarantee the dependability and applicability of the model in clinical practice, a thorough examination of the outcomes of heart disease prediction using Random Forest should comprise evaluations of accuracy, feature importance, interpretability, validation, and clinical relevance.

X. CONCLUSION

Early diagnosis of cardiovascular disease can help with lifestyle modifications for high-risk patients, which can lower complications and be a significant medical milestone. Through the use of backward elimination and RFECV in feature selection, this project was able to accurately forecast heart disease.

After testing six different machine learning algorithms, we discovered that Random Forest was the most accurate. To determine how well this model performs, you should test it using the test set. Random Forest served as the model. To further improve it, we can utilize more advanced models and train on models to forecast the different kinds of cardiovascular problems and provide advice to users.

XI. REFERENCES

[1] A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary HEART DISEASE using random forest classifier,’’ in Proc. Int. Conf.Recent Trends Comput.Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ‘‘Using PSO algorithm for producing best rules in diagnosis of HEART DISEASE ,’’ in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306– 311.

[3] N. Al-milli, ‘‘Backpropogation neural network for prediction of HEART DISEASE ,’’ J. Theor.Appl.Inf.Technol., vol. 56, no. 1, pp. 131–135, 2013.
[4] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ‘‘Analysis of neural networks based HEART DISEASE prediction system,’’ in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.
[5]P.K.Anooj,‘‘Clinical decision support system: Risk level prediction of heartdiseaseusingweightedfuzzyrules,’’J.KingSaudUniv.-Comput.Inf. Sci.,vol.24,no.1,pp.27–40,Jan.2012.