Original Article

A Survey On Effective Heart Disease Diagnosis For Machine Learning Algorithms

Mrs. P. Sasikala[1] , Dr. K. G. Dharani[2]

[1]Assistant Professor, Department of Electronics and Communication Engineering, Faculty of Engineering, Karpagam Academy of Higher Education, Coimbatore, India

[2]Associate Professor, Department of Electronics and Communication Engineering, Faculty of Engineering, Karpagam Academy of Higher Education, Coimbatore, India

DOI: 10.47750/pnr.2022.13.S07.200

Abstract

One of the most difficult problems in medical data analysis is predicting cardiac disease. In the medical field, especially in the field of cardiology, seasonable and precise identification of cardiac disease is censorious. This paper employs machine-learning approaches to evolve an efficient and veracious grouping for the identification of cardiac disease. Supporting transmitter organization, stipulation reasoning, Unreal group, K-nearest individual, Nave bays, and Choice thespian are among the classification algorithms misused, with authoritative features selection algorithms such as Assuagement, utilized to remove moot and prolix features and also presented a crossbreed SVM model to calculate the job of characteristic pick. The characteristic option improves sorting quality and reduces the categorization group process experience. When compared to the different categorization methods, Varied SVM is the best strategy for rising temperature disease foretelling. Furthermore, the proposed coming can be old in attention to detect cardiac problems.

Keywords: Heart disease prediction, Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbour, Naïve bays, Decision tree, Intelligent systems

I. Introduction

Temperament disease is a series of health problems that have affected more people all over the world [1]. Breathlessness, physical bodily weakness, and swelling feet are all classic indications of cardiac disease [2]. Research students are working to create a practical method for current cardiac disease prediction techniques that are inefficient in early identification for several ideas, including accuracy of output and output execution time, disposition disease is a serious issue [3].

It is very difficult to diagnose and treat cardiac disease at the point when current innovations and clinical specialists are not accessible [4]. With the proper identification and treatment, several lives are spared [5]. Around 26 million people have heart illness, with 3.6 million new cases being identified each year, according to Disposition disease [6].

Heart disease is often diagnosed by a doctor primarily based totally on the patient's scientific history, the effects of the bodily examination, and analysis of any alarming symptoms [7]. The results of this approach to diagnosis, however, fall short in terms of identifying patients with heart disease. Additionally, analysis is expensive and computationally difficult [8]. A painless finding technique given by machine learning (ML) algorithm is being developed to overcome these challenges. Due to the use of an expert

decision-making framework, the fatality rate has dropped, and every disease is a cardiac condition that is difficult to diagnose. [9] and [10]. Numerous researchers [11] and [12] made use of the existence of viscus illness Need for a machine learning prediction model with the right information for training and testing [13]. A machine-learning model can perform better when training and testing are conducted on balanced datasets.

The model's prediction abilities can also be enhanced by using pertinent and pertinent data attributes. Selection of features and data balance is therefore essential for enhancing technique effectiveness. In the paper, numerous researchers have given a variety of diagnostic techniques, but these methods do not accurately diagnose cardiac disease. To boost the predictive power of machine learning models, data pre-processing is required for data standardization [14].

This research proposed a diagnosis method utilizing machine learning. Heart illness is the ruling ailment. Heart disease is diagnosed using machine learning predictive models such as ANN, LR, K-NN, SVM, DT, and NB [15].

II. Literature Review

In the written paper, scientists have suggested various machine learning-based prediction algorithms to identify heart illness. This research study shows a variety of machine learning-based diagnosis tools that are currently in use to explain the significance of the proposed work. Based on machine learning classification techniques, Detrano et al. [16] developed a Heart disease classification system, and the system's accuracy was 77%.

The ANN-DBP algorithm performed exceptionally well when used in conjunction with the FS technique. Palaniappan et al. [17] discussed a method for detecting heart disease using professional medical diagnosis. The system was created using artificial neural networks, navies bays, and decision trees as machine learning predictive models. The DT classifier had an accuracy of 80.4%, NB had an accuracy of 86.12%, and ANN had an accuracy of 88.12%. Olney et al. [18] created a three-step process based on an artificial neural network strategy for angina heart disease prediction that had an accuracy of 88.89%.

The categorization algorithm was 87.4% accurate. With a sensitivity of 80.09 percent, an 89.01 percent accuracy, and a 95.91 percent specificity, an ensemble-based diagnostic using ANNs stem for heart disease was created using the statistical measuring system enterprise miner [19]. built a comprehensive medical decision assistance system for the detection of cardiac disease constructed using fuzzy AHP and artificial neural networks.

Together with the global evolutionary strategy and the features selection technique, the Cleveland dataset was employed. An 80.41 percent accurate diagnosis system for heart disorders was developed by Kumar et al. [21] and Gudadheet et al. [22] utilizing several layers of perception and classifier for a support vector machine. A categorization method for cardiac illness based on neural networks and fuzzy logic integration was created by Humar et al. [23]. Akil et al. [24] created a system for diagnosing cardiac illness based on machine learning.

The Performance of the suggested method was 91.10 percent. A classification system for heart disease was proposed by Liu et al. [25] using the relief and rough set approaches. Classification accuracy for the suggested method was 92.32 percent.

According to Haq et al. [26], the performance of the classifier K-nearest neighbour (K-NN) was assessed on both the entire set of features and a subset of them. It was discovered that the suggested method was incredibly accurate. In additional research, Mohan et al. [27] created a strategy for anticipating cardiac illness utilizing combined machine learning methods. He suggested one more innovative approach for identifying crucial features in data so that machine learning classifiers can be trained and tested. They were accurately categorized, 88.07% of the time.

Geweid et al. [28] created methods for diagnosing heart illness using an improved SVM-based duality optimization method.

III. Materials and Method

a) Data set

The Cleveland heart disease dataset [29] is utilized in this work for testing reasons. 303 occurrences and 75 attributes were present when this data set was created, but only a subset of 14 of these were used in all published studies. [30] Six samples were eliminated from the data set in this study due to missing values after pre-processing it. The dataset has 297 samples remaining, 13 attributes, and one output label. Two classifications on the output label are used to explain the lack of cardiac disease and its identification. The retrieved data are then arranged into a 297*13 features matrix.

b) Pre-processing of the data set

The dataset is pre-processed for the necessary usage for effective representation. The dataset has undergone pre-processing utilizing methods including Min-Max Scalar, standard Scalar (SS), and attribute missing values elimination.

IV. Machine Learning Technique Implementation

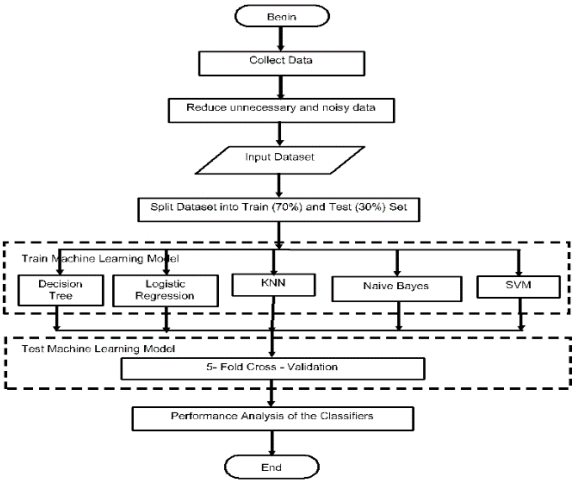

Fig 1: Flow Chart of the Machine Learning Techniques

The completed dataset was split into two sets: training and testing. 70% of the data was still needed to train the machine-learning model. The final model will be tested using the remaining 30% of the data after the model has been trained. Afterward, we applied machine learning (ML) utilizing these divided datasets using the Python programming language. In place of regression, we applied classification methods utilizing supervised machine learning techniques because the results of our experiment are true or false values rather than continuous numerical values. We employed five classification machine learning approaches to train the proposed model. The following are the algorithms: 2. Logistic Regression, 3. Decision Tree, 3. Naive Bayes, 4. K Nearest Neighbor (KNN), and 5. Support Vector Machine (SVM).

V. Theory and Methodology

With roots in fields like processing, knowledge analytics, statistics, algebra, and more, machine learning (ML) is a large, multidisciplinary field. It makes it difficult to come up with an entirely new term. This algorithm may be a specific AI technique because it gathers data from training sets. It is at the inspiration of the tree and has many subsections and branches, but it is not telling the machines where to look during this learning process. According to Fig 2, machine learning is divided into the following categories.

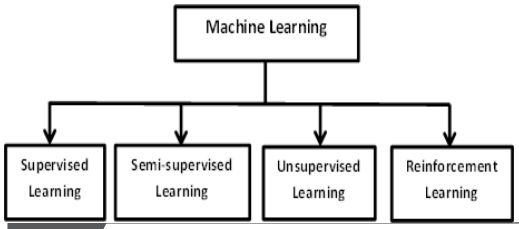

Fig 2: Machine Learning Classifications Algorithms

1. Supervised Learning
2. Semi-supervised Learning
3. Unsupervised Learning
4. Reinforcement Learning

1. Support Vector Machine

The fundamental goal of classification is to identify a partitioning hyperplane in the sample space using a training set D=(x1,y1), (x2,y2),...(xm,ym), where y1-1,+1 denotes the separation between the different sample classes.

The aforementioned linear equations can be used to describe how hyperplanes are distributed throughout the sample space:

$$W^T x + b = 0 \quad (1)$$

Assume that the data meet the following conditions under the linear separable condition:

$$y_i(w.x_i + b) \geq +1 \quad (2)$$

When it comes to relaxation variables, the data must meet the following requirements:

$$y_i(w.x_i + b) + \varepsilon_i \geq +1 \quad (3)$$

Finally, the following is the optimization equation:

$$\min w, b\|w\| 22 \text{ s.t.g } i(w) = -[\, y_i\, (wx_i + b) - 1] \geq 0 \quad (4)$$

Logistic Regression

It is possible to determine the Logistic Regression equation using the Linear Regression equation. The mathematical procedures to create Logistic Regression equations are as follows:

We are aware of the following ways to represent the straight-line equation:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n \quad (5)$$

Divide the preceding equation by (1-y) because the range of y in a Logistic Regression is only 0 to 1:

$$\frac{y}{1-y}; 0 \text{ for } y = 0, \text{and infinity for } y = 1 \quad (6)$$

But we need a range from -[infinity] to +[infinity], in which case the equation's logarithm is

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n \quad (7)$$

This completes the Logistic Regression equation.

2. Artificial neural network

The general model of ANN is depicted in the below diagram, then its processing comes next.

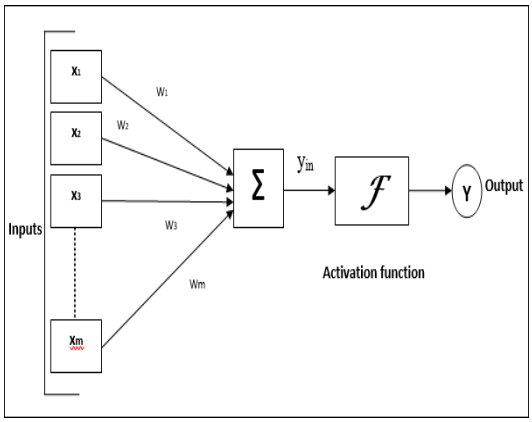

Fig 3: Artificial Neural Network

For the general artificial neural network model mentioned above, the net input can be calculated as follows:

$$y_{in} = x_1.w_1 + x_2.w_2 + x_3.w_3 \ldots x_m.w_m \quad (8)$$

i.e., Net input $$y_{in} = \sum_i^m .w_i x_i \quad (9)$$

The activation function can be applied to the output and is calculated using the net input.

$$Y = F(y_{in}) \quad (10)$$

3. K-nearest neighbour

The K-NN algorithm can be used to explain how it functions:

- Step 1: Select the neighbor's number K.
- Step 2: Find the distance in Euclidean geometry between K neighbours.
- Step 3: Find the K nearest neighbours using the calculated Euclidean distance.
- Step 4: The K nearest neighbours should be located using the calculated Euclidean distance.
- Step 5: The category with the most neighbours should receive the additional data points.
- Step 6: The model algorithm is finished.

4. Naïve bays

The Bayes' Theorem is used in Naive Bayes, which presupposes that all predictors are independent $(c \mid x) = P(x \mid c)\, P(c) / P(x)$

$$P(c \mid x) = P(x_1 \mid c) \text{ x } P(x_2 \mid c) \text{ x} \ldots P(x_n \mid c) \text{ x } P(c) \quad (11)$$

Based on the predictor (x), P (c|x) is the posterior probability for the class (c). P(c) is the prior probability for the class, P(x) is the prior probability for the predictor, and P(x|c) is the probability for the particular class for the predictor (c).

5. Decision tree

An all-purpose predictive modelling technique called decision tree analysis has numerous uses. Decision trees are often constructed using an algorithm that chooses among various methods to segment a data collection based on specific criteria. It is among the most well-liked and useful supervised learning algorithms. A supervised non-

parametric acquisition technique called resolve trees can be used for all categorization and regression problems. The objective is to generate a simulation that forecasts the regard of a train inconstant by processing over simplified selection data features from rules.

If-then-else sentences are commonly used as decision rules. As the tree gets deeper and the model is more precise, the rules get more complicated.

Before we go any further, let's familiarise ourselves with some of the terms:

• Attribute: A quantity that describes an instance
• Instances: A list of the traits or features making up the input area.
• Hypothesis Class: A collection of all potential functions
• Concept: The process of converting input into output.
• Target Concept: The role that we're trying to determine, i.e., the real answer
• Sample: A set of inputs with a label designating the exact output (also known as the Training Set).
• Candidate Notion: A notion that, in our opinion, is comparable to the target notion.
• Testing Set: This set, which is used to assess the candidate's concept and performance, is comparable to the training set.

VI. Result and Discussion

To comprehend the significance of our recommended methodology, Table 1 summarises the disadvantages and benefits of the suggested methods for diagnosing heart disease based on the literature that came before it. The early stages of heart disease are detected using several techniques in all of these contemporary treatments.

But when it comes to the prediction of heart disease, all of these methods have poor prediction accuracy and lengthy computation times. The Table 1 shows that to assess heart disease while travelling and offer better care and recovery, the accuracy of heart disease detection technologies needs to be further enhanced. Due to the use of unnecessary features in the dataset, the main shortcomings of older techniques include low accuracy and lengthy computation durations. New methods for accurately diagnosing heart illness are needed to solve these difficulties. There is a significant research gap and challenge in improving prediction accuracy.

Table 1: Summary of the earlier techniques

Ref	Technique	Disadvantages	Advantages	Accuracy (%)
R. Detrano et al.[11]	This technique is Heart disease diagnosis using machine learning classifiers	The accuracy of the suggested procedure is rather poor.	less difficult to compute.	77.00
M. Gudadhe, et al.[22]	This method is MLP+SVM	mathematically challenging	The proposed technique has a high level of effectiveness in terms of prediction accuracy.	80.41
H. Kahramanliet al.[23]	This technique is ANN + Fuzzy Logic	Results generation takes more time during execution.	Precision is high	87.04

S. Mohan et al.[27]	This model is a Hybrid ML method	poor precision	quick computation	88.07
S. Palaniappanet al.[17]	This prediction model is a Heart disease diagnosis system based on NB, DT, and ANN	Poor performance is shown with decision trees and naive bases.	ANN displayed good accuracy performance.	88.12
E. O. Olaniyiet al.[18]	This model is Three phase technique based on ANN	Intensive computation	more precision	88.89
R. Das et al. [19]	This method is ANN aggregation-based diagnosis scheme	mathematically challenging	High precision	89.01
O.W. Samuel et al.[20]	This method is ANN-Fuzzy –AHP	computationally demanding	great precision was attained	91.01
X. Liu et al.[25]	This paper is a Relief-Rough set based method for heart disease detection	Calculation time is more lengthy	Great precision because the model was trained and tested with the right features.	92.32

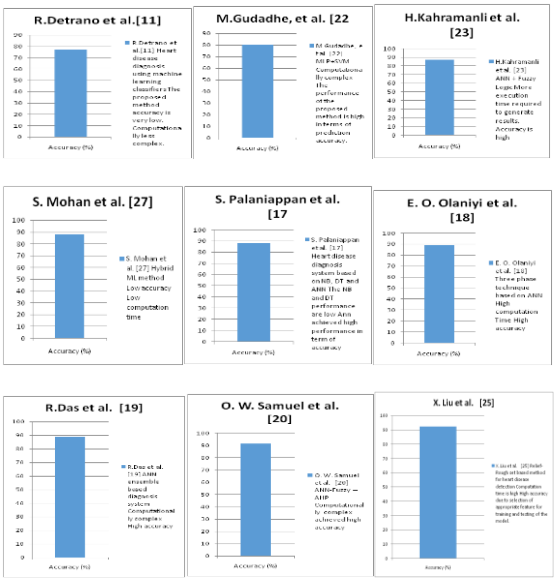

Fig 4: Graphical Representation of data

The above demo delegacy of accumulation, Fig 4 is R. Detrano et al. [11], Disposition disease identification using tool learning classifiers technique is used, the advantage of this classifier is computationally less complex and limitation of this classifier is accuracy, which is very low. The percentage of accuracy is 77%.

The above pictorial representation, Fig 4 is M. Gudadhe, et al .[22], MLP+SVM technique is used, the advantage of this classifier is the action of the proposed approach scores well in terms of precise forecasting, and the limitation of this classifier is accuracy which is computationally complex. The percentage of accuracy is 80.41%.

The above image representation of data, Fig 4 is H. Kahramanliet al. [23], ANN + Fuzzy Logic technique is used, the advantage of this classifier is accuracy that is high and the limitation of this classifier is more execution time required to generate results. The percentage of accuracy is 87.04%.

The above g written performance of accumulation, Fig 4 is S. Mohan et al. [27], A hybrid ML technique is used, the advantage of this classifier is low computation time and also the limitation of this classifier is computation time which is high. The percentage of accuracy is 88.07%.

The above graph representation of data, Fig 4 is S. Palaniappanet al.[17], Utilizing the NB, DT, and ANN techniques, a grouping of diseases based on the disposition is used, the advantage of this classifier is ANN displayed good accuracy performance. and limitation of this classifier is Low NB and DT performance is observed. The percentage of accuracy is 88.12%.

The above illustration performance of accumulation, Fig 4 is E. O. Olaniyietal.[18], three-state model based on the ANN technique is used, the advantage of this classifier is High accuracy, and the limitation of this classifier is high computation Time. The percentage of accuracy is 88.89%.

The above result of Fig 4 is R. Das et al. [19], ANN accumulation-based diagnosis system technology is used, the advantage of this classifier is High accuracy and limitation of this classifier computationally complex. The percentage of accuracy is 89.01%.

The above output graph of Fig 4 is O.W. Samuel et al.[20], ANN-Fuzzy –AHP technique is used, the advantage of this classifier is achieved high accuracy, and the limitation of this classifier is computationally complex. The percentage of accuracy is 91.01%.

The above graph result of Fig 4 is X. Liu et al.[25], A method based on a relief-rough set for temperament disease sleuthing technique is used, the advantage of this classifier is highly accurate as a result of choosing the right feature for the model's training and testing and the limitation of this classifier is computationally complex. The percentage of accuracy is 92.32%.

Table 2: Machine Learning Algorithms

Reference	Classifier	Description
[36],[39]-[41]	Logistic regression	In a binary categorization issue, LR is used to forecast when the y [0,1] negative class, the value of the y variable is 0 and the y [0,1] positive class is 1. When y[0,1,2,3] it also employs multiple classifications are used to calculate y's value.
[26],[41]-[46]	Support Vector Machine	SVM algorithms are frequently employed to solve classification issues. Numerous applications mostly used SVM due to its outstanding classification performance.
[47]	Naïve Bayes	The NB method is used to classify the issue in question. the training set of data that NB uses to Given a, ascertain the conditional probability of the vectors in a particular class. Each vector's value of the conditional probability is assessed, and the conditionality probability of the new vectors class is then evaluated.
[41],[50]	Artificial Neural Network	The neural network algorithm (ANN) integrates neurons that convey messages. An ANN classifier has a hidden layer, an outlier, and a signaling layer. The input layer receives input values that are employed in the network's training process. The output of the ANN is determined for the known class. The error gap between the projected and actual class values is used to reassess the weight.
[41],[48],[49]	Decision Tree	The DT shape resembles a tree with leaves or decision nodes. Both internal and external nodes

		are connected in a DT. Internal nodes make decisions and send a child node to the next nodes as part of the decision-making process.
[41]	K-Nearest Neighbour	K-NN uses them to determine the class label of fresh input, and compares the new input to its input samples in the training set. If the training set's samples do not match the samples in the training set and the new input is the same. The performance of the K-NN classification is subpar.

VII. Conclusion

In this paper, an effective machine learning-related diagnosis apparatus to identify suspicious diseases was built. The problem can be solved by preparing an unstructured machine learning model from raw healthcare data. This system was designed using classifiers for machine learning such as LR, K-NN, ANN, SVM, NB, and DT. After diagnosis, the decision Tree accuracy is 87.546%, Logistic Regression accuracy is 86%, K-Nearest-Neighbor (KNN) accuracy is 87%, Naive Bayes accuracy is 84% and Connection Agent Machine (SVM) accuracy is 91%. I will therefore keep working on disease treatment and rehabilitation in the future, even for serious illnesses like diabetes, breast, and heart disease.

REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, ‘‘Epidemiology and risk profile of heart failure,’’ Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.

[2] M. Durairaj and N. Ramasamy, ‘‘A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,’’ Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255–260, 2016.

[3] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, ‘‘Decision making in advanced heart failure: A scientific statement from the American heart association,’’ Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.

[4] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ‘‘Innovative artificial neural networks-based decision support system for heart diseases diagnosis,’’ J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013, Art.no. 35396.

[5] Q. K. Al-Shayea, ‘‘Artificial neural networks in medical diagnosis,’’ Int. J. Comput.Sci. Issues, vol. 8, no. 2, pp. 150–154, 2011.

[6] J. Lopez-Sendon, ‘‘The heart failure epidemic,’’ Medicographia, vol. 33, no. 4, pp. 363–369, 2011.

[7] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, D. M. Lloyd-Jones, S. A. Nelson, G. Nichol, D. Orenstein, P. W. F. Wilson, and Y. J. Woo, ‘‘Forecasting the future of cardiovascular disease in the united states: A policy statement from the American heart association,’’ Circulation, vol. 123, no. 8, pp. 933–944, 2011.

[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, ‘‘Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,’’ J. Roy. Soc. Interface, vol. 8, no. 59, pp. 842–855, 2011.

[9] S. I. Ansarullah and P. Kumar, ‘‘A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method,’’ Int. J.Recent Technol. Eng., vol. 7, no. 6S, pp. 1009–1015, 2019.

[10] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, ‘‘Fuzzy logic based decision support system for component security evaluation,’’ Int. Arab J. Inf. Technol., vol. 15, no. 2, pp. 224–231, 2018.

[11] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J.Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, ‘‘International application of a new probability algorithm for the diagnosis of coronary artery disease,’’ Amer. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989.

[12] J. H. Gennari, P. Langley, and D. Fisher, ‘‘Models of incremental concept formation,’’ Artif.Intell., vol. 40, nos. 1–3, pp. 11–61, Sep. 1989.

[13] Y. Li, T. Li, and H. Liu, ‘‘Recent advances in feature selection and its applications,’’ Knowl. Inf. Syst., vol. 53, no. 3, pp. 551–577, Dec. 2017.

[14] J. Li and H. Liu, ‘‘Challenges of feature selection for big data analytics,’’ IEEE Intell. Syst., vol. 32, no. 2, pp. 9–15, Mar. 2017.

[15] L. Zhu, J. Shen, L. Xie, and Z. Cheng, ‘‘Unsupervised topic hypergraph hashing for efficient mobile image retrieval,’’ IEEE Trans. Cybern., vol. 47, no. 11, pp. 3941–3954, Nov. 2017.

[16] S. Raschka, ‘‘Model evaluation, model selection, and algorithm selection in machine learning,’’ 2018, arXiv:1811.12808. [Online]. Available: http://arxiv.org/abs/1811.12808

[17] S. Palaniappan and R. Awang, ‘‘Intelligent heart disease prediction system using data mining techniques,’’ in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 108–115.

[18] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, ‘‘Heart diseases diagnosis using neural networks arbitration,’’ Int. J. Intell.Syst. Appl., vol. 7, no. 12, p. 72, 2015.

[19] 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.

[20] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, ‘‘An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,’’ Expert Syst. Appl., vol. 68, pp. 163–172, Feb. 2017.