# Feature Selection Using Data Mining Techniques for Prognostication of Cardiovascular Diseases

Naga Venkata Jashwanth Vanami(✉), Lohitha Rani Chintalapati, Yagnesh Challagundla, and Sachi Nandan Mohanty

School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India
jashwanthvanami0502@gmail.com

**Abstract.** Cardiovascular diseases (CVD) are a major cause of mortality worldwide causing about 17.9 million deaths per year. Cardiovascular illnesses are a group of conditions that affect the heart and blood arteries. These illnesses may have an effect on various parts of the heart and/or blood vessels. CVD encompasses coronary artery disorders (CAD), such as myocardial infarction and angina. To reduce the risk and deaths caused by cardiovascular diseases it is important to predict it at an early stage. It is crucial to be aware of these cardiac disease-related signs in order to forecast outcomes and offer a solid foundation for diagnosis for which data mining and feature selection prove to be useful. However, manual analysis and prediction are laborious and tiring due to the sheer volume of data. In this study, data science is used to predict cardiac problems. The potential method for heart disease prediction is one that analyses the relationships between variables and extracts hidden knowledge from the data. Through a variety of indications, our study attempts to anticipate cardiac disease correctly and promptly. We propose a cardiovascular disease prediction model which uses a dataset obtained from Kaggle on which we perform various data pre-processing techniques on which feature selection is done and the refined data is given to different machine learning models for the prediction of the disease. We obtained the highest accuracy of 99.4% using Random Forest, demonstrating the effectiveness and dependability of the heart disease prediction approach we presented.

**Keywords:** Cardiovascular disease · Feature Selection · Random Forest · Machine learning

## 1 Introduction

A first-sized organ that pumps blood throughout the body, the heart, is situated at the left of the sternum and is shielded by the ribcage. The main component of our circulatory system is the heart. It primarily consists of four main chambers made of muscle that are propelled by electrical impulses that regulate the rhythm and speed of the heart-beat as well as maintain blood pressure. The heart valves, which act as gateway between the chambers of the heart which opens and closes to allow the flow of blood through the


B. Wang et al. (Eds.): ICMTEL 2023, LNICST 532, pp. 344–353, 2024.
https://doi.org/10.1007/978-3-031-50571-3_24

mitral and tricuspid valves, are controlled by the nervous system, which also controls the rhythm and speed of the heart rate.

The neurological nervous system, which transmits impulses instructing our hearts to beat more slowly when we are at rest and more quickly when we are under stress, and the endocrine system, which releases hormones instructing our blood vessels to constrict or relax, are the two main systems of the heart. From the aforementioned information, we can conclude that the heart plays one of the most important roles in a person's ability to survive. However, the truth is that heart conditions are among the most prevalent types of illnesses that people experience, and statistics indicate that 17.9 million people worldwide died from cardiovascular diseases in 2019, accounting for 32% of the deaths caused. 85% of these deaths are caused by heart attacks and strokes [4]. WHO estimates that 31% of worldwide human mortality is caused by cardiovascular diseases. About one person dies from heart disease every minute in the modern era [7]. As per WHO, ischemic and hypertensive heart attacks are one of the leading causes of death. This illness can be lethal if it is not carefully monitored [13].

A blockage that obstructs the flow of blood to heart or brain is the main cause of heart attacks and strokes, which are typically sudden events. Unhealthy diet, physical inactivity, cigarette use, and problematic alcohol consumption are some of the risk factors for heart disease and stroke [8]. Heart attacks are primarily brought on by blood clotting in arteries, which can also cause chest pain or stroke. As a result, some individuals experience unstable angina [16]. When blood vessel function is impaired by heart disease, coronary artery infections occur, which weaken the patient [6]. The underlying blood vessel illness frequently has no symptoms. However, the consequences of behavioral risk factors might manifest in people as high BP, high fasting blood sugar, high blood lipids, overweight, and obesity. An elevated threat of cardiovascular disease death and morbidity exists in those with type 2 diabetes [9]. Mostly low and middle-income countries globally contribute to nearly 75% of Cardiovascular deaths which is heartbreaking because most cardiovascular diseases are preventable and treatable when detected earlier. The inclusion of strategies for managing cardiovascular disease in universal health insurance plans, such as making use of cutting-edge tools like machine learning which are used to determine, track down, and forecast a variety of medical conditions, holds the key to reducing the prevalence of the cardiovascular disease. Our research's main goal is to provide clinicians with a tool for early cardiac issue detection. The advancement of machine learning applications shows that it is possible to detect cardiovascular disease in its early phases using electrocardiogram and patient data, which will help in overcoming the significant challenges in today's modern world that are contributing to huge number of deaths globally [10]. It will be simpler to provide patients with proper care using machine learning techniques' superiority in pattern recognition and categorization on comparison to the conventional statistical approaches, medical professionals and researchers have resorted to them to create screening tools [18], while avoiding serious effects on some of the most exciting applications of AI that lie in utilizing big data and machine learning to identify patients at high risk of developing chronic diseases [19]. Researchers can forecast the chance that susceptible patients may develop cardiac ailments by using data mining techniques [14]. Data science is needed in the healthcare sector to handle the enormous quantity of data that is generated [7]. The objective

interpretation of all outcomes for the same patient made possible by machine learning techniques may enhance the precision of each step's diagnostics [11].

In order to reduce the input variable to our model by using only the pertinent data and eliminating noise, we have also employed the Feature Selection methods [20], which automatically selects relevant features of our dataset that are trained and tested on our machine learning model based on the sort of problem we are trying to solve.

In this study, we have implemented various types of techniques for the prediction of cardiac diseases which include preprocessing the data, feature selection using various data mining techniques and classification and prediction over various data mining models. The rest of the paper is structured as follows: Sect. 2 addresses related work, Sect. 3 discusses methodology, Sect. 4 puts forth the results, and Sect. 5 concludes the paper.

## 2 Related Work

Machine learning algorithms can be used to predict and diagnose heart diseases [15]. Researchers make use of various data mining techniques and computational intelligence models to predict heart diseases. Predicting heart diseases accurately and on time can save lots of lives and can be very useful for healthcare professionals for diagnosis.

There is a need to precisely predict and provide a reliable basis required for further treatment on time to reduce the risk, complications and death rate caused by heart diseases. Improving the ability to predict and identify heart diseases would benefit healthcare professionals and would decrease the risk of death and save lives. Several machine learning algorithms can be used to accomplish this along with feature selection, data mining and various other techniques.

Research by Le, Hung Minh et al. provides a method by combining data mining and feature selection to study the presence of heart disease. The Infinite Latent Feature Selection (ILFS) approach is used to re-select heart disease characteristics based on the rank and weights supplied to them. Support Vector Machine is used to divide a subset of the chosen qualities into various heart disease classes [17]. To produce greater quantities and types of data Synthetic Minority Over-sampling Technique is used. Experimental findings showed that the method distinguished between "no presence" and "presence" and obtained 97.87% accuracy and five separate categories of cardiac disease with a 93.92% accuracy [3].

In a study by Bashir, Saba, et al. Multiple heart disease datasets are utilized in feature selection strategies and algorithms for experimentation analysis and to demonstrate accuracy improvement in predicting heart disease. Using the Rapid Miner tool; applying feature selection techniques like Logistic Regression, Logistic Regression SVM, Random Forest, Decision Tree, and Naive Bayes, and it is observed that the accuracy of the findings is improved [2].

Zhang, Dengqing, et al. put forth a DNN and Linear SVC embedded feature selection method-based heart disease prediction system. Additionally, for the feature selection the Linear SVC method and L1 norm are used to choose the best feature subset. They examined the He normal, random normal, and Xavier weight and concluded that He initialization produces the best outcomes. The suggested approach has a 98.56% accuracy rate, demonstrating the viability and dependability of using deep neural networks and feature selection to predict heart disease [1].

Features importance ranking of two gradient boosting techniques, XGBoost and CatBoost was calculated on SA heart, Statlog heart and Cleveland data sets, by Anuradha et al. and observed that CatBoost fared better than the other classifiers [4].

The proposed work by Boukhatem et al. uses a variety of data mining approaches, including Logistic Regression, Random Forest, Naive Bayes, and Decision Trees to classify patient risk factors and assess the chance of heart disease. The effectiveness of various machine learning algorithms has been compared; According to the trial findings, the Random Forest method has the best accuracy 90.16% [7]. Mohan et al. proposed combining a linear model and a hybrid random forest to predict heart disease which showed improved results with 88.7% accuracy [5]. In a study by El-Hasnony et al. by iteratively selecting the most relevant data to query their labels, five multi-label active learning selection algorithms (AUDI, Random, Adaptive, MMC, and QUIRE) were used to lower the cost of labelling. [15].

In a paper by Πεταρούδας et al. an impressive prediction system referred to Intelligent Heart Disease Prediction System (IHDPS) was proposed, which employs the three widely used data mining approaches of Decision Trees, Neural Network and Naive Bayes. The Decision Trees method is the most accurate, which obtained an accuracy of 89%, Naive Bayes obtained an accuracy of 86.5%, and an accuracy of 85.53% by neural network [14]. Saikumar et al. Applied the DCAlexNet Convolutional Neural Networks technique to carry out the deep learning-based classification for heart disease detection. Performance metrics such as accuracy of 98.67%, the sensitivity of 97.45%, recall of 99.34%, and an F1 score of 99.34% are generated by the feature based fusion based confusion matrix of DCAlexNet-CNN. These numerical comparison results outperform application robustness and compete with present technology [12].

## 3 Methodology

Cardiac diseases can be predicted through various indicators so we considered a dataset from Kaggle which consists of 75 attributes or indicators and the target attribute field which alludes to the patient having heart illness or not, where having heart disease is indicated by 1 and no heart disease is indicated by 0. It consists of 1025 rows i.e. patient records of various ages, where 312 are female, and 713 are male. The considered dataset is made up using four different databases named Cleveland, Hungary, Switzerland, and the VA Long Beach. To predict cardiovascular disease, we follow three major steps that include Data Pre-Processing, Feature Selection using various data mining techniques and Classification using various models.

### 3.1 Data Pre-processing

Real-world data is inadequate, erratic, inaccurate, and sometimes in an undesirable format and in order to make predictions, machine learning models cannot use the obtained data directly. Therefore, data preprocessing is performed and it aids in improving data quality and preparing the raw data for machine learning models by cleaning, formatting, and organizing it. It promotes the data to be mined for insightful information. Data preprocessing in machine learning is a data mining technique that transforms raw data into

readable and understandable format. Data preprocessing includes various procedures like data cleaning, outlier removal, data reduction, data standardization, data normalization, data transformation, binning etc. which have been performed on the considered dataset for the accurate prediction of cardiovascular diseases.

Data cleaning eliminates large portions of irrelevant data, corrects inaccurate data in the train-validation-test dataset and minimizes duplicates.

Outliers should be eliminated or rescaled as an important data preparation steps. Outliers are data points that deviate significantly from the average. They may bias the findings and reduce the precision of the machine learning model. The most popular technique for finding outliers is to utilize standard deviation, however there are other approaches as well. Outliers are data points that deviate more than two standard deviations from the mean.

Data reduction is the process of taking a larger amount of original data and reducing it significantly while preserving its integrity. Data reduction techniques include Dimensionality Reduction, Numerosity Reduction, Data Cube Aggregation, Data Compression, Discretization Operation.

Rescaling the parameters to have a mean of 0 and a variation of 1 is the process of data standardization. Standardization's objective is to reduce all characteristics to a comparable scale without distorting the variations in the value's ranges.

Data translation into a particular range usually [0 1] or simple data transformation onto the unit sphere are both instances of normalizing in machine learning. The process of connecting disparate, soiled, and normalized data into a single, dimensionally modelled, de-normalized, and analysis-ready state is known as data transformation. Without the right technical stack in place, data transformation may be expensive, time-consuming, and difficult. However, conversion will ensure the best data quality, which is necessary for accurate analysis and eventually allowing data-driven decisions.

Binning is used to reduce the impact of minor observational errors. It is sometimes referred to as data discrete binning or data bucketing. The original data values that lie inside a particular narrow interval are replaced with a value representative of that interval, usually mean or median. After successfully preprocessing the data, we move on to feature selection.

### 3.2 Feature Selection

A dataset consists of numerous features or attributes which help in prediction. But there can be huge number of features and many of which might not play an important role in prediction and as a result of the addition of extraneous information during model training the model's overall accuracy is decreased, its complexity is increased, its capacity to be generalized is decreased, and it becomes biased. Filter methods, wrapper methods, and embedding methods are the three main groups of feature selection techniques. Feature selection plays a crucial role in increasing the prediction accuracy by finding the ideal combination of characteristics or features for a machine learning model. Better accuracy is obtained by using lesser data.

Our data which is preprocessed is now undergoing feature selection, our data is now reduced to 7 features on feature selection, keeping the target attribute intact after preprocessing. The considered features play an important role in predicting heart disease

and considering which we obtain higher accuracy and hence now machine learning algorithms are provided the data on feature selection for precise cardiovascular disease prediction.

### 3.3 Classification Using Various Models

KNN:
K-Nearest Neighbour is a supervised learning technique. K-NN algorithm that stores the data and based on similarity it sorts a new data point. It is used for problems like classification and regression, although categorization problems are where it is most commonly employed.

Decision Tree:
Decision Tree is a type of unsupervised machine learning in which the training data is segmented continually based on a particular parameter, on description of the input and the associated output. The two elements that may be utilized to describe the tree are decision nodes and leaves.

SVM:
Both classification and regression employ the Support Vector Machine (SVM), a supervised machine learning technique. The SVM method looks for the optimum line or decision boundary that may split the space into several categories as well as a hyper-plane in an N-dimensional space that unambiguously classifies the input points.

SGD:
Stochastic gradient descent is a simple yet very powerful technique for fitting linear classifiers and regressors under convex loss functions, such as (linear) Support Vector Machines and Logistic Regression (SGD). Gradient descent is a general-purpose optimization technique that may find the best solutions to a range of problems.

Random Forest:
In the Random Forest supervised machine learning method, a "forest" is created by growing a number of decision trees and combining them. It uses a variety of samples to generate decision trees, using the majority of them for classification and the average of them for regression.

Naive Bayes:
A family of classification algorithms built on the Bayes' Theorem are known as naive Bayes classifiers. It is one of the most straightforward and effective Classification algorithms, assisting in the development of rapid machine learning models capable of making prompt predictions.

Logistic Regression:
Guided learning is demonstrated using logistic regression. Assigning data to a discrete set of classes using a classification method allows one to determine or anticipate the likelihood of a binary (yes/no) event occurring.

Gradient Boosting:
Classification and regression tasks are accomplished using a machine learning technique called gradient boosting. It offers a prediction model in the form of several weak prediction models, such as decision trees. It is based on the supposition that the total prediction error is minimized when prior models are coupled with the best potential upcoming model. The main idea is to specify the expected outcomes for this subsequent model in order to minimize inaccuracy.

## 4 Results

The prediction of cardiovascular diseases was performed on a dataset obtained from Kaggle which was preprocessed using various data preprocessing techniques including procedures like data cleaning by imputing missing values and replacing them with average or most frequent value, selecting relevant features using Chi2 Score with 80% proportion of features upon which the dataset now has 10 attributes or features and one target attribute. The features are now normalized and standardized to $\mu = 0$, $\sigma^2 = 1$. We then removed the Sparse features i.e. removed features with too many missing values with a threshold percentage of 5. On removing the sparse features, we now have 8 features that are considered along with the target attribute. We now implement Principal Component analysis where data is normalized, build the covariance matrix, then find eigen values and eigen vectors, arrange the eigen vectors in highest to lowest order and select the number of principal components, in our data upon PCA we obtain 8 Principal Components. We then apply CUR Matrix Decomposition with rank 10 and relative error of 1.00 upon which we have 7 features in our dataset and one target attribute in which 0 means no heart disease and 1 means heart disease. The dataset is divided into a 30% test set and a 70% train set. Repeated Stratified K-Fold = 5 validation of the models KNN, Decision Tree, SVM, SGD, Random Forest, Naïve Bayes, Logistic Regression and Gradient Boosting on the train set containing the selected subset of features is performed on 20 times repeat test/train. We obtain prediction accuracy of KNN 81.9%, Decision Tree 95.4%, SVM 86.3%, SGD 82.3%, Random Forest 99.4%, Naïve Bayes 79.8%, Logistic Regression 82.3%, and Gradient Boosting 97.0% (Figs. 1, 2, 3, and 4).

**Table 1.** Accuracy, Precision, Recall, F1-score of the considered models

| Model | Accuracy (%) | F1 Score (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|
| KNN | 81.9 | 81.8 | 81.9 | 81.9 |
| Tree | 95.4 | 95.4 | 95.4 | 95.4 |
| SVM | 86.3 | 86.3 | 86.5 | 86.3 |
| SGD | 82.3 | 82.3 | 82.7 | 82.3 |
| Random Forest | 99.4 | 99.4 | 99.4 | 99.4 |
| Naïve Bayes | 79.8 | 79.8 | 79.8 | 79.8 |
| Logistic Regression | 82.3 | 82.3 | 82.6 | 82.3 |
| Gradient Boosting | 97.0 | 97.0 | 97.0 | 97.0 |

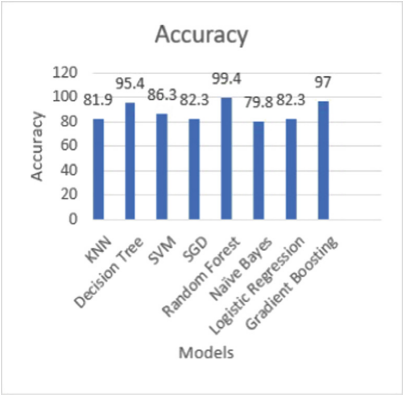


**Fig. 1.** Accuracy of considered models

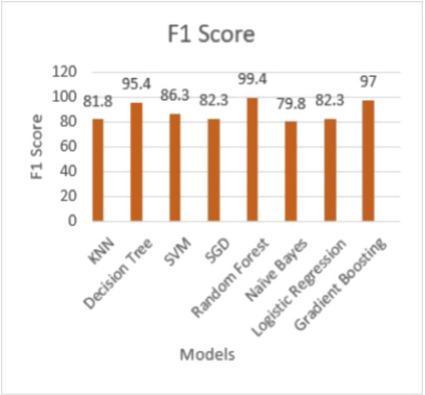


**Fig. 2.** F1 Score of considered models

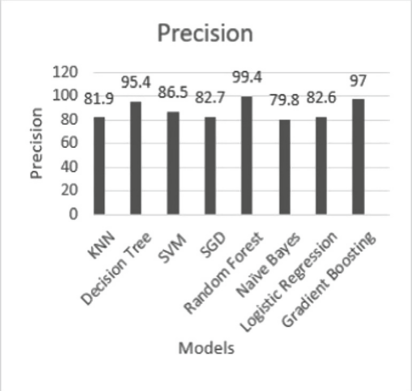


**Fig. 3.** Precision of considered models


**Fig. 4.** Recall of considered models

Precision is a metric used to analyze a model's dependability and its accuracy in categorizing a sample as positive. It is calculated using the ratio of True Positives to True Positives and False Positives.

$$Precision = \frac{True_{positive}}{True_{positive} + False_{positive}} \tag{1}$$

The Recall parameter is used to assess how well the model can identify positive test data. It is the ratio of Positive samples that are rightly labelled as positive to the total number of positive Instances samples. The greater the Recall value, the greater the number of positive samples detected.

$$Recall = \frac{True_{positive}}{True_{positive} + False_{Negative}} \tag{2}$$

Models for classification include the F1 Score. The F1 Score is focused on precision and recall. The Harmonic mean of Precision and Recall is the F1 Score.

$$F1\,score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

From Table 1 we can examine the accuracy of heart illness prognosis using different machine learning algorithms and we have obtained the highest accuracy of 99.4% for Random Forest. Thus, we can conclude that upon data preprocessing and feature selection using data mining methodologies for prognostication of cardiac diseases is useful to predict heart diseases accurately which will help decline the death rate by predicting the risk of the disease and help in diagnosis.

## 5 Conclusions

Accurate cardiovascular disease prediction is crucial in lowering risk and death rate. In this research we proposed a cardiovascular disease prediction model which uses various data mining techniques, feature selection techniques which play a major role in increasing the accuracy by reducing the input variables by considering only relevant data and avoid noise. It reduces the data which in turn reduces the complexity of processing and prediction by the model and thereby increases the accuracy. We performed preprocessing and feature selection which included procedures like imputing missing values, selecting relevant features using Chi2 score with 80% proportion of features, it is now normalized and standardized to $= 0, = 1$, we then removed the sparse features with 5% threshold and then applied Principal Component Analysis and then applied CUR matrix decomposition with rank 10 and relative error of 1.00 upon which the dataset is now processed, refined, transformed, and 7 features have been selected successfully and one target attribute in which 0 indicates no heart disease and 1 indicates the presence of heart disease. The dataset produced is split into 70% train and 30% test set. Further we performed 5-fold cross validation on various machine learning models namely KNN, Decision Tree, SVM, SGD, Random Forest, Naïve bayes, Logistic Regression and Gradient Boosting which produced accuracies of 81.9%, 95.4%, 86.3%, 82.3%, 99.4%, 79.8%, 82.3% and 97.0% respectively. We conclude that Random Forest has shown the highest accuracy of 99.4% with F1 score 99.4%. From the aforementioned results obtained we can conclude that machine learning models with feature selection, Data mining techniques can be useful for prognostication of cardiovascular diseases.

## References

1. Zhang, D., et al.: Heart disease prediction based on the embedded feature selection method and deep neural network. J. Healthc. Eng. **2021**, 1–9 (2021)
2. Bashir, S., et al.: Improving heart disease prediction using feature selection approaches. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST). IEEE (2019)
3. Le, H.M., Tran, T.D., Van Tran, L.A.N.G.: Automatic heart disease prediction using feature selection and data mining technique. J. Comput. Sci. Cybernet. **34**(1), 33–48 (2018)

4. Anuradha, P., David, V.K.: Feature selection and prediction of heart diseases using gradient boosting algorithms. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). IEEE (2021)
5. Mohan, S., Thirumalai, C., Srivastava, G.: Effective heart disease prediction using hybrid machine learning techniques. IEEE Access **7**, 81542–81554 (2019)
6. Riyaz, L., et al.: Heart disease prediction using machine learning techniques: a quantitative review. In: International Conference on Innovative Computing and Communications. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-3071-2_8
7. Boukhatem, C., Youssef, H.Y., Nassif, A.B.: Heart disease prediction using machine learning. In: 2022 Advances in Science and Engineering Technology International Conferences (ASET). IEEE (2022).
8. Ayon, S.I., Islam, M.M., Hossain, M.R.: Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J. Res. **68**(4), 2488–2507 (2022)
9. Wang, M., et al.: Artificial intelligence models for predicting cardiovascular diseases in people with type 2 diabetes: a systematic review. Intell. Based Med. **6**, 100072 (2022)
10. Ahsan, M.M., Siddique, Z.: Machine learning-based heart disease diagnosis: A systematic literature review. Artif. Intell. Med. **128**, 102289 (2022)
11. Kukar, M., et al.: Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artif. Intell. Med. **16**(1), 25–50 (1999)
12. Saikumar, K., Rajesh, V., Babu, B.S.: Heart disease detection based on feature fusion technique with augmented classification using deep learning technology. Traitement du Signal **39**, 1 (2022)
13. Dunbray, N., et al.: An analytical survey on heart attack prediction techniques based on machine learning and IoT. In: Proceeding of International Conference on Computational Science and Applications, pp. 299–312. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-0863-7_24
14. Πεταρούδας, Μιλτιάδης Γεωργίου. Comparative analysis of machine learning techniques in predicting heart attacks. Diss. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης ( 2022)
15. El-Hasnony, I.M., et al.: Multi-label active learning-based machine learning model for heart disease prediction. Sensors **22**(3), 1184 (2022)
16. Aggarwal, R., Kumar, S.: An automated perception and prediction of heart disease based on machine learning. AIP Conf. Proc. **2424**, 1 (2022). AIP Publishing LLC
17. Lakshmanaprabu, S.K., et al.: Optimal deep learning model for classification of lung cancer on CT images. Futur. Gener. Comput. Syst. **92**, 374–382 (2019)
18. Derbali, M., et al.: Water desalination fault detection using machine learning approaches: a comparative study. IEEE Access **5**, 23266–23275 (2017)
19. Al-Darraji, I., et al.: Adaptive robust controller design-based RBF neural network for aerial robot arm model. Electronics **10**(7), 831 (2021)
20. Mohanty, S.N., et al.: Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks. Phys. Commun. **40**, 101097 (2020)