Recursive Feature Elimination with Feature Ranking-Based Feature Selection for Healthcare Data

[1]Prof. Arati K. Kale, [2]Dr.Dev Ras Pandey
[1,2]*Kalinga University,Raipur*

Abstract-The feature selection method is a learning acquisition technique that selects the most pertinent features. Feature selection aims to reduce computational overload and improve the classifier by enumerating essential features. The presence of replicate and unimportant features in an extensive data set can dramatically reduce the efficiency of machine learning models. This paper proposes an efficient feature selection algorithm for healthcare data using recursive feature elimination with feature ranking (RFE_FR) to improve classification performance. The different feature ranking methods, including info gain, relief, and correlation-based ranking approaches, are used to compute the feature rank. It reduces the feature dimension and selects features efficiently. This approach improves performance by removing associated and redundant features from the dataset. To investigate the RFE_FR performance, the machine learning classification algorithms KNN, NB, RF, and SVM are employed. The accuracy of the suggested method performance is assessed using the real-time healthcare dataset. This study demonstrates that, when comparing the model efficiency with and without feature selection, the RFE_FR strategies selected have a significant favourable effect on the model performance.

Keywords: Feature Selection, Recursive Feature Elimination, Ranking, health data, machine learning

1 INTRODUCTION

Physicians now have more opportunities to diagnose patients more accurately because of the expansion of medical data collecting. Practitioners have used computers more frequently in recent years to enhance decision-making support. Machine learning is emerging as a vital tool in the healthcare sector to help with patient diagnostics [1]. Machine learning is a useful analytical tool when dealing with vast and challenging-to-programme tasks, like translating medical records into knowledge, making pandemic predictions, and analyzing genomic data [2].

The massive size of the dataset is one of the main issues with machine learning. The weighting features [3] improve algorithm performance by reducing redundant data and processing time, which is caused by the analysis of many features, which uses a lot of memory and causes overfitting [4], [5]. A limited number of features can characterize different disorders of health management and genomic expression. While feature selection reduces the dataset by eliminating superfluous features, dimensionality reduction employs feature extraction to simplify and modify data [6].

Three primary categories are used to distinguish the feature selection process [7]: filter, wrapper, and embedding approach. The primary distinction is that wrapper techniques feature a subset evaluation stage and include a classification algorithm. A wrapper [8] uses the classification algorithm to assess the effectiveness of the selected features. No correlation exists between a classification algorithm and the filter feature selection procedure. Filter algorithms [9] are frequently more broad and less expensive computationally than wrapper algorithms. But wrappers assess the feature subsets according to the classification performance, while filters ignore the performance of the chosen features on a given algorithm. This means that wrappers typically outperform filters for specific algorithms. Embedded approaches combine feature selection and classifier learning into a single procedure [10]. This paper uses filter and wrapper-based feature selection approaches to select the best features from the medical dataset.

Recursive Feature Elimination (RFE) is a wrapper feature selection technique using machine learning models to determine the relevant scores of features [11]. RFE calculates a relevance score for every feature after training a model with the complete feature set. The feature with the lowest relevance score is ignored in the following stage, and the model is

retrained to determine new feature relevance scores. This procedure is repeated until the feature set has the required features. This paper proposes an efficient feature selection algorithm which uses recursive feature elimination with feature ranking. The key objectives of this research paper are:

- This work presents a feature selection algorithm with a filter- and wrapper-based approach.
- The recursive feature elimination is used to find the optimal feature set.
- The different feature ranking methods, including info gain, relief and correlation-based ranking approach, are used within RFE.
- The proposed feature selection approach is evaluated using real-time medical data.

The remaining part of this research paper is organized as follows: Section 2 describes the related work of feature selection methods. Section 3 presents the proposed feature selection methodology. The result and discussion are presented in section 4, and section 5 concludes the research paper.

2 RELATED WORKS

This section presents some existing machine learning-based feature selection approaches. Gárate-Escamila et al. [12] suggested a dimensionality reduction method for identifying heart disease features using a feature selection strategy. This method enhances machine learning model prediction by combining a chi-square with principle component analysis (PCA). In most classifiers, the chi-square and PCA combination provides better results.

A new multi-objective feature selection strategy based on PSO is proposed by Rostami et al. [13]. There are three stages to the process. The initial stage displays the original features as a model of a graph structure. Phase two involves calculating attribute centralities for each node in the graph. Phase three involves final feature selection through an enhanced PSO-based search procedure.

A model that effectively predicts cardiac disease is proposed by Ghosh et al. [14]. The author employed effective data collection, pre-processing, and transformation techniques to produce precise training model data. The Relief and Least Absolute Shrinkage and Selection Operator (LASSO) approaches are used to choose appropriate features. Combining bagging and boosting techniques with traditional classifiers creates new hybrid classifiers for heart disease classification.

The artificial flora algorithm (AFA)--based feature selection method for diabetes classification is proposed by Nagaraj et al. [15]. AFA was used to select patients' electronic health records information, including demographics, health status, laboratory results, and prescriptions. The patient cases were categorized as either type I, type II, or gestational diabetes mellitus using the GBT-based classification model.

Using classifier ensemble approaches, Nagarajan et al. [16] designed a hybrid GA-ABC approach for feature selection and classification. This algorithm is based on genetics and simulates an artificial bee colony. It achieves a classification accuracy improvement of over 90%. Nagarajan et al. [17] created a hybrid genetic-based crow search technique to choose features and classify them using deep convolution neural networks. The model attains a classification accuracy of above 94%.

An efficient classification strategy for the diagnosis of chronic kidney disease (CKD) is proposed by Senan et al. [18]. The most effective significant features of CKD were selected using the RFE method. The classification algorithms SVM, KNN, decision tree, and random forest were used to classify the CKD. Every classifier parameter was adjusted to achieve optimal classification, leading to encouraging outcomes for every method.

Priscilla et al. [19] offer a novel two-phase feature selection strategy that combines filter and wrapper techniques to find the essential feature subsets. Mutual Information (MI) was used in the first step to rank the features according to their importance. Recursive Feature Elimination (RFE), a wrapper technique used with 5-fold cross-validation, is thus implemented in a second phase to remove the redundant features. By modifying the class weights, eXtreme Gradient Boosting (XGBoost) is selected as the estimator for RFE. Four boosting methods were employed with the best features derived from the procedure to examine the classification performance.

A technique called recursive feature elimination-based gradient boosting is presented by Theerthagiri et al. [20]. Features are selected for hyperparameter optimization using a stochastic gradient boosting approach. An ensemble technique based on gradient

boosting has been devised to predict cardiovascular disease. The hybrid filter-genetic feature selection strategy proposed by Ali et al. [21] improves the accuracy of cancer classification by addressing the challenge of high-dimensional microarray datasets. Filter feature selection techniques like information gain, information gain ratio, and Chi-squared are utilized to identify the most important features of malignant microarray datasets. A genetic algorithm has been used to optimize and refine the selected features and increase the suggested method performance for cancer classification. Table 1 provides a summary of related work.

Table 1 Related Work Summary

Ref.	Algorithms / Techniques	Dataset	Results
[12]	Chi-square with PCA	Heart Disease with 74 features	It selects 13 features and attains more than 98% accuracy.
[13]	Multi-objective PSO	Medical datasets: Colon, Leukemia, Prostate tumour and Lung cancer	It achieves more than 85% accuracy for different classifiers (SVM, Naïve Bayes and AdaBoost)
[14]	Relief and LASSO with different classifiers	Heart Disease dataset with 13 features	Random Forest Bagging with Relief feature selection achieves 99.05% accuracy.
[15]	Artificial flora algorithm and Gradient-boosted tree	Diabetes (Type I, Type 2 and gestational diabetes)	It increases the classification performance with more than 90% accuracy.
[16]	Genetic and Artificial Bee Colony	Heart, Dermatology, Lung cancer, Hepatitis, Diabetes etc.,	The model accurately predicted 88.78% of the original features and 92.34% of the extracted features.
[17]	Genetic-Based Crow Search Algorithm	Heart, Dermatology, Lung cancer, Hepatitis, Diabetes etc.,	The accuracy of the GCSA model was 95.34% for extracted features and 88.78% for all original features.
[18]	RFE with SVM, KNN, Decision Tree and Random Forest	CKD dataset	It achieves 96.67%, 98.33%, 99.17%, and 100% accuracy for SVM, KNN, decision tree, and random forest.
[19]	Mutual Information and eXtreme Gradient Boosting with RFE	Credit card fault	It achieves 85% accuracy.
[20]	RFE with gradient boosting	Cardiovascular disease	Compared to SVM, NB and LR algorithms, it increases accuracy by 14.12% to 30.12%.
[21]	Filtering and Genetic algorithm	Lung cancer, Breast cancer, Brain cancer and Central Nervous System	It eliminated roughly 50% of the redundant features and only kept the relevant ones.

3 PROPOSED METHODOLOGY

This section explains the proposed pre-processing techniques to improve the data quality. Figure 1 shows the proposed architecture diagram.

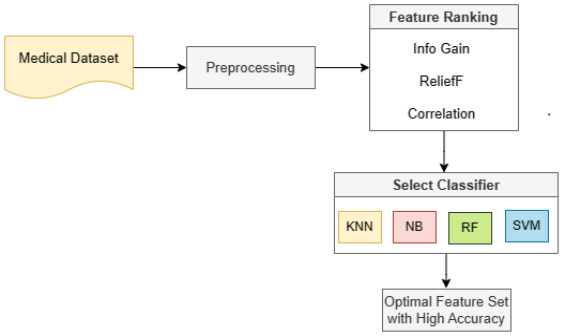

Figure 1 Proposed Work Architecture

1.1 Feature Ranking

Feature ranking is a crucial approach that arranges all the features in a dataset so that many significant features indicate extensive information.

This paper uses Info Gain, Relief, and correlated-based feature ranking approaches.

Info gain:

Info gain is one of the most used feature selection techniques based on mutual information. It is easy to use and efficiently computes. It measures the quantity of information in bits regarding the class prediction in the presence of that feature. It knows the matching class distribution or the data between the jth feature (f_j) and the class labels (C). The following formula is used to determine the information gain of a feature:

$$InfoGain(X, f_j) = E(X) \sum_{av=values(f_j)} \frac{|X, f_j = av|}{|X|} E(X, f_j = av) \quad (1)$$

Where X is the dataset, $\frac{|X,f_j=av|}{|X|}$ is the portion of the sample with f_j having the attribute value av, and E(X) is the entropy given by:

$$E(X) = -\sum_{i=1}^{NC} p(c_i)\, log_2\big(p(c_i)\big) \qquad (2)$$

where NC is the number of classes and $p(c_i)$ is the likelihood that class c_i will be found in training set X. A high information gain indicates that a feature is more relevant.

ReliefF: this technique uses a feature's capacity to distinguish between similar occurrences. The closest hit and closest miss cases are located for a random sample taken from the training set. Next, depending on the values of the hit-and-miss instances, the algorithm modifies the weight of every feature. A feature with a higher weight value is better at differentiating across instances of the same class.

$$w[a] = w[a] - \frac{diff(A, R_i, Hit)}{m} + \frac{diff(A, R_i, Miss)}{m} \qquad (3)$$

R_i is for an instance that was chosen at random. Relief looks for its two closest neighbours: closest miss, who is from the opposite class, and closest hit, who is from the same class. It modifies the w[a] consistency calculation for feature 'a' based on Ri, Miss, and Hit values. The performance rating w[a] is decreased if there is a significant discrepancy between R_i and Hit. However, if a feature 'a' has a substantial difference between R_i and Miss, 'a' might be utilized to identify between various classes, in which case the weight w[a] is raised.

Correlation-based Ranking:

The correlation-based approach is a traditional filter method that selects features based on the output of an evaluation function that is heuristic (correlation-based). This function prefers to choose subsets whose attributes are highly associated with the class but not with one another. Repetitive features are selected because they have a strong relationship with at least one of the other features. In contrast, minor features with minimal connection with the class should be disregarded for the same reason. A feature's recognition will depend on how well it predicts classes in portions of the instance space not currently expected by other features.

$$F_s = \frac{t\overline{r_{cf}}}{\sqrt{t + t(t-1) + \overline{r_{ff}}}} \qquad (4)$$

Here F_s is the heuristic assessment for a feature subset of t features. $\overline{r_{cf}}$ is the average correlation value among features and the class label, $\overline{r_{ff}}$ is the mean inter-correlation value between features.

1.2 Recursive Feature Elimination with Feature Ranking

Recursive Feature Elimination (RFE) is a wrapper feature selection technique that uses machine learning models to determine the features' relevance scores. RFE calculates an importance score for every feature after training a model with the complete feature set. The model is retrained to calculate new feature significance scores once the feature with the lowest relevance score is ignored in the following stage. This process is repeated until the required number of features remains in the feature set.

This paper proposes a feature selection algorithm which uses recursive feature elimination with feature ranking approaches. Algorithm 1 explains the proposed RFE_FR algorithm. The RFE_FR is to select the optimal features based on feature ranking after iteratively building the model. The classification algorithm trains the chosen features and determines the classification accuracy. This process is then repeated for the remaining features. Figure 2 shows the workflow of the RFE_FR approach.

Algorithm-1 RFE_FR Feature Selection

Input: Dataset D = {(X_1,Y_1), (X_2,Y_2),…., (X_n,Y_n)}, NF = Total no of Features, Classifiers Cls ={KNN, NB, RF, SVM}, cls_acc = 0;
Output: Optimal Feature Set with Accuracy
Step01: While (NF != 0)
Step02: Apply InfoGain to compute InfoGain_FR using Eq.(1)
Step03: Apply ReliefR to compute Relief_FR using Eq. (3)
Step04: Apply Correlation to compute Corr_FR using Eq. (4)
Step05: Select Cls and Apply the Wrapper approach to compute Wrap_FR
Step06: $FeatRank = (InfoGain_{FR} + Relief_{FR} + Corr_{FR} + Wrap\_FR)/4$
Step07: $Rank_{Thr} = \frac{\sum FeatRank}{Length(FeatRank)}$
Step08: Initialize sFeat and rFeat
Step09: For i = 1 to Length(FeatRank)
Step10: $If\ (FeatRank(i) < Rank_{Thr}$ then
Step11: rFeat.add(FeatRank(i))
Step12: else
Step13: sFeat.add(FeatRank(i))

```
Step14:    EndFor
Step15:    D' = Remove rFeat from dataset D
Step16:        acc = Apply classification algorithm to
compute classification accuracy
Step17:    If acc > cls_acc then
Step18:       cls_acc = acc
Step19:       OptimalFeat = sFeat
Step20:       NF = Length (OptimalFeat)
Step21     EndIf
Step22:    If NF <= 3|| rFeat is Empty
Step23:       break;
Step24:    EndIf
Step25: EndWhile
Step26: Return OptimalFeat and cls_acc
```

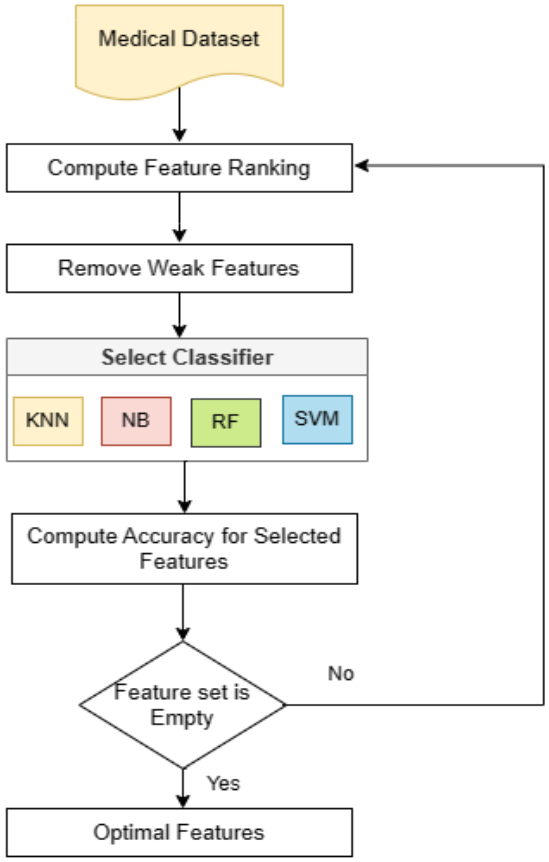

Figure 2 RFE_FR workflow diagram

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

This section explains the performance evaluation of the research activity. The proposed feature selection was implanted and tested using Java with the suggested experimental setup. Real-time medical datasets from Kaggle and UCI were used to evaluate the recommended feature selection strategy. Table 2 describes the dataset.

Table 2 Dataset Description

Dataset	# Instances	# Features	# Class
Cirrhosis	412	17	4
Heart	303	13	2
Kidney	400	24	2
Lung Cancer	32	56	3

The following metrics are used to analyze the performance of the proposed pre-processing approach.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

Where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

The classification algorithms Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF) and Support Vector Machine (SVM). Table 3 shows the number of features selected for different algorithms and datasets.

Table 3 Selected Number of Features for different datasets

Dataset	# Features	Selected Features			
		KNN	NB	RF	SVM
Cirrhosis	17	11	11	11	11
Heart	13	4	8	4	8
Kidney	24	10	10	10	10
Lung Cancer	56	4	9	4	20

From Table 3, all the algorithms select the same number of features for the Cirrhosis and Kidney dataset.

Table 4 and Figure 3 show the accuracy of the dataset without feature selection.

Table 4 Accuracy without Feature Selection

Algorithms	Cirrhosis	Heart	Kidney	Lung Cancer
KNN	47.816	75.248	97.75	53.125
NB	51.699	83.498	97.75	62.5
RF	60.68	81.188	98.75	65.625
SVM	53.155	83.828	97.75	50

From Figure 3, the RF algorithm achieves high accuracy for the Cirrhosis, Kidney and Lung cancer

datasets. The SVM attains high accuracy for the heart dataset.

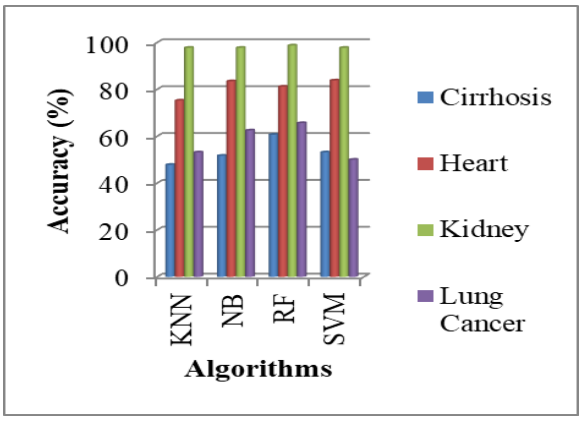

Figure 3 Comparison of accuracy for medical dataset (Without Feature Selection)

Table 5 and Figure 4 shows the accuracy comparison with feature selection

Table 5 Accuracy with Feature Selection

Algorithms	Cirrhosis	Heart	Kidney	Lung Cancer
KNN	54.612	81.518	98.75	81.25
NB	55.583	84.818	97.75	71.875
RF	59.466	81.518	98.25	84.375
SVM	50.243	84.488	98.5	78.125

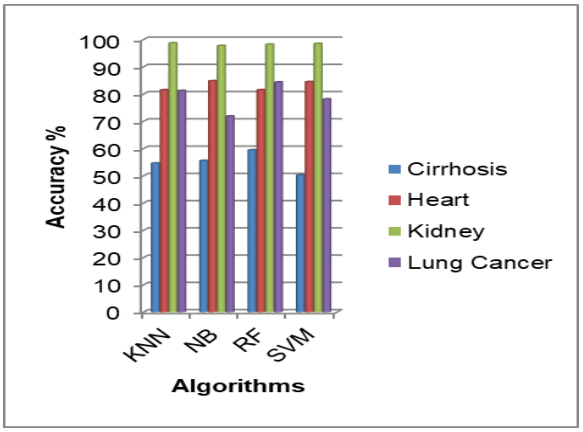

Figure 4 Comparison of accuracy for medical dataset (With Feature Selection)

Table 6 Precision, Recall and F1-measure comparison

Dataset	Algorithms	Metrics		
		Precision	Recall	F1-Measures
Cirrhosis	KNN	53.974	54.612	54.193
	NB	57.8	55.583	55.839
	RF	59.213	59.466	58.97
	SVM	42.1	50.243	42.666
Heart	KNN	81.716	81.518	81.384
	NB	85.006	84.818	84.727
	RF	81.556	81.518	81.447
	SVM	84.707	84.488	84.385
Kidney	KNN	98.75	98.75	98.749
	NB	97.754	97.75	97.751
	RF	98.249	98.25	98.249
	SVM	98.5	98.5	98.5
Lung Cancer	KNN	82.5	81.25	81.046
	NB	71.843	71.875	71.526
	RF	87.121	84.375	84.148
	SVM	78.671	78.125	78.283

5 CONCLUSION

Feature selection is a crucial approach, and its significance becomes even more apparent when the dataset being worked with has many features. This paper proposes the feature selection algorithm, which uses recursive feature elimination with feature ranking to improve the classification model. The feature rank is calculated using several feature ranking techniques, such as info gain, relief, and correlation-based ranking approaches. It effectively chooses features and minimizes the feature dimension. This method enhances performance by eliminating related and unnecessary information from the dataset.

REFERENCE

[1] Ghorbani, R., Ghousi, R., Makui, A., & Atashi, A. (2020). A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset. IEEE Access, 8, 141066-141079.

[2] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3)

[3] Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 34(4), 1060-1073.

[4] Gao, W., Hu, L., & Zhang, P. (2020). Feature redundancy term variation for mutual information-based feature selection. Applied Intelligence, 50, 1272-1288.
[5] Nagarajan, G., & Dhinesh Babu, L. D. (2021). A hybrid feature selection model based on improved squirrel search algorithm and rank aggregation using fuzzy techniques for biomedical data classification. Network Modeling Analysis in Health Informatics and Bioinformatics, 10(1)
[6] Agrawal, P., Abutarboush, H. F., Ganesh, T., & Mohamed, A. W. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). Ieee Access, 9, 26766-26791.
[7] Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2015). A survey on evolutionary computation approaches to feature selection. IEEE Transactions on evolutionary computation, 20(4), 606-626.
[8] Sahebi, G., Movahedi, P., Ebrahimi, M., Pahikkala, T., Plosila, J., & Tenhunen, H. (2020). GeFeS: A generalized wrapper feature selection approach for optimizing classification performance. Computers in biology and medicine, 125, 103974.
[9] Siddiqi, M. A., & Pak, W. (2020). Optimizing filter-based feature selection method flow for intrusion detection system. Electronics, 9(12), 2114.
[10] Afrin, S., Shamrat, F. J. M., Nibir, T. I., Muntasim, M. F., Moharram, M. S., Imran, M. M., & Abdulla, M. (2021). Supervised machine learning based liver disease prediction approach with LASSO feature selection. Bulletin of Electrical Engineering and Informatics, 10(6), 3369-3376.
[11] Gunduz, H. (2021). An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination. Financial innovation, 7(1)
[12] Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19
[13] Rostami, M., Forouzandeh, S., Berahmand, K., & Soltani, M. (2020). Integration of multi-objective PSO based feature selection and node centrality for medical datasets. Genomics, 112(6), 4370-4384.
[14] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R. and De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 19304-19326.
[15] Nagaraj, P., Deepalakshmi, P., Mansour, R. F., & Almazroa, A. (2021). Artificial flora algorithm-based feature selection with Gradient boosted tree model for diabetes classification. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 14.
[16] Nagarajan, S. M., Muthukumaran, V., Murugesan, R., Joseph, R. B., & Munirathanam, M. (2021). Feature selection model for healthcare analysis and classification using classifier ensemble technique. International Journal of System Assurance Engineering and Management, 1-12.
[17] Nagarajan, S. M., Muthukumaran, V., Murugesan, R., Joseph, R. B., Meram, M., & Prathik, A. (2022). Innovative feature selection and classification model for heart disease prediction. Journal of Reliable Intelligent Environments, 8(4), 333-343.
[18] Senan, E.M., Al-Adhaileh, M.H., Alsaade, F.W., Aldhyani, T.H., Alqarni, A.A., Alsharif, N., Uddin, M.I., Alahmadi, A.H., Jadhav, M.E. and Alzahrani, M.Y., 2021. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering, 2021.
[19] Priscilla, C. V., & Prabha, D. P. (2021). A two-phase feature selection technique using mutual information and XGB-RFE for credit card fraud detection. Int. J. Adv. Technol. Eng. Explor, 8, 1656-1668.
[20] Theerthagiri, P. (2022). Predictive analysis of cardiovascular disease using Gradient boosting based learning and recursive feature elimination technique. Intelligent Systems with Applications, 16.
[21] Ali, W., & Saeed, F. (2023). Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. Processes, 11(2)