Enhanced classification method for large-scale medical data

Ahmed Hassan
ElSewedy University of Technology
Cairo, Egypt
ahmed.hassan@sut.edu.eg

Khaled M, Fouad
faculty of computer science and engineering
New Mansoura University
Mansoura, Egypt
kmfi@fci.edu.eg
Khaled.foad@nmu.edu.eg

Alaa Eldin Abdalla Yassin
Information System department
Benha University
Benha, Egypt
alaaeldeen.yassin@fci.bu.edu.eg
Alaaeldin.yassin@must.edu.eg

Somia Mohamed
Information System department
Benha University
Benha, Egypt
somia.aboelnaga@fci.bu.edu.eg

Abstract— the medical community has been concerned about how to increase the accuracy of different classification methods with large data that are being generated every day. The traditional KNN method has many limitations, such as dealing with large-scale data, handling uncertain data, and also determining the k parameter for KNN that gives the best result. In this research, three limitations are solved. By using optimized feature selection methods, optimal features are chosen from many irrelevant features. Then, uncertain data that has a conflict with the class label is handled. Finally, the optimal number of k in the KNN method that gives better accuracy is chosen using the ROC curve. The prime objective of this paper is to develop a hybrid optimal model for medical data classification that handles these challenges. The results are evaluated using the accuracy metric. Experimental results show that the enhanced KNN method outperforms the previously used KNN method in used medical datasets

Keywords—classification, KNN, feature selection, optimized method, ROC curve, uncertain data, classification

I. INTRODUCTION

The data is growing very fast, and there are many fields and applications that depend on data for their decisions and work, such as the medical field. It is important to extract important information from data [1]. Huge amounts of data cannot be handled and processed by traditional machine learning [2-3], so we need to find ways to adapt these methods to large-scale data.

Data are becoming bigger not only in terms of the number of data objects but also the dimensionality of features. This can significantly affect the accuracy and efficiency of most learning algorithms [4]. Feature selection methods can be used to identify and remove irrelevant attributes from data that do not affect the accuracy of a classification model [5]. The main goal of feature selection is to find minimal features from a problem domain with high accuracy in representing the original features [6]. By looking at previous studies of using feature selection with classification models, the studies conclude that using feature selection as a preprocessing step not only reduces the number of features but also enhances the performance of the classification model [7].

A classification is a supervised machine learning method that builds models that can assign objects to one of the predefined classes. Data are divided into two sets: a training set and a testing set. The training set is used to build the model, and the testing set is used to validate the model [8]. The k-nearest Neighbor (KNN) classifier is learned by analogy. When the unseen object is presented, the K-Nearest Neighbor classifier searches for the closest k-training objects. The k closest training objects are the k nearest neighbors for the new unseen object. Then, a new object is assigned to the class with the maximum number of samples in the closest neighborhood of the new object [9]. From the previous definition, we can conclude that the implementation of the KNN method is easy to implement and does not require training time. Still, as mentioned previously, the traditional KNN method has some limitations when dealing with scale data. Some research has handled this limitation [9 – 11]. Some limitations of the KNN method that are handled in this paper are choosing the best number of K and enhancing KNN method implementation to deal with uncertain data.

In this paper, an enhanced classification method for medical data is introduced. The presented approach consists of two phases. The first phase uses a hybrid feature selection method as a preprocessing phase to select an optimal feature subset. The second phase is to build an improved KNN classifier for large-scale medical datasets. Our contributions are as follows:

- A hybrid feature selection method with optimized methods is used to generate optimal features that give high accuracy to models.
- An enhanced KNN method is used to deal with uncertain data that has a conflict in the class label by updating both the training set and testing set.
- Finally, ROC curve is used to get the optimal K number for the enhanced KNN method.

The structure of this paper is as follows. The related works are reviewed in Section 2, and then the proposed model is shown in Section 3. The experimental platform is shown in Section 4. The used medical datasets are presented in Section 5. The experimental results and a comparison between other models are discussed in Section 6. Finally, Section 7 concludes this study.

II. RELATED WORKS

Mohamed et al. [12] aimed to improve an ensemble of the traditional KNN algorithm and four new models were generated from it based on the density, single linkage, average linkage, and complete linkage distributions of the dataset. As a result, the enhanced KNN obtained the best results from several ensemble classifiers.

979-8-3503-6777-5/24/$31.00 ©2024 IEEE

Jayasri et al. [13] evaluate a medical database of diabetes patients using innovative techniques that include Hierarchical Decision Attention Network, Association Rules to identify relationships between diseases and symptoms and the authors use MapReduce Framework. The aim of this study was to give the patients the best solutions. The techniques used show improved performance.

Amartya et al. [14] designed a method for the early detection of diabetic retinopathy using the Hadoop Framework to handle big data. The methods consist of three phases: feature extraction, feature reduction, and classification. The Histogram of Oriented Gradients was used for the first phase, PCA was used for the reduction phase, and KNN was used for the final phase. The proposed technique provided better performance than other approaches. Aishwarya et al. [15] used various classification methods for diabetes predictions. The study determined that the AdaBoost classifier gave better accuracy results than other classification methods, such as Logistic Regression, Random Forest Classifier Linear Discriminate Analysis, and Gradient Boost Classifier.

Shafi et al. [16] proposed a method for classified colorectal diagnosis with and without feature selection techniques. The combinations of feature selection and classification methods prove that it gave good results. Other research that used the same combination with different methods is listed [17, 18, and 19]. Mrutyunjaya Panda [3] conducted on ten well-known cancer microarray gene selection datasets obtained from the UCI Machine Learning Repository. The empirical results from the proposed Elephant Search Algorithm based deep learning approach are compared with those from the most recent published studies, assessing its applicability for future bioinformatics research. This analysis provides valuable insights into selecting the best classification model for gene expression data.

Vahidet al., [20] introduced a feature-level aggregation-based ensemble based on overlapped feature subspace partitioning framework for microarray data classification. The proposed framework consisted of three steps. Firstly, vertical portioning in the training data set was performed. Then, a ranker was performed for each portion. Finally, six aggregation functions were executed to combine ranked lists of features. The FLAE-OFSP framework evaluated on seven microarray datasets, measuring performance across four criteria: stability, classification accuracy, runtime, and Modscore. The results demonstrate substantial improvements in runtime efficiency and notable enhancements in classification accuracy and other evaluated measures compared to individual feature selection methods.

Namrata Singh et al. [21] presented a hybrid feature selection method for medical data to improve classification performance. The proposed method is composed of four stages. In the first stage, data was portioned. In the second phase, filter feature methods are used to rank features. The wrapper method was used in the third stage to select the best features. In the final stage, a classification method was used. The introduced approach outperformed other techniques.

Yılmaz Kaya et al. [22] proposed a hybrid model for diagnosing breast cancer, Lymphography, and erythematous-squamous diseases. The model was based on using Factor Analysis (FA) as a preprocessing phase, and then the obtained factors were used as input features for the Extreme Learning Machine (ELM). For the dermatology dataset, the ELM model achieved an average success rate of 96.39%, while the FA+ELM model improved this to 96.94%. The lymphography dataset yielded a success rate of 84.50% with ELM, which was slightly enhanced to 85.10% using the FA+ELM approach. In the case of the Wisconsin Breast Cancer dataset, the ELM model attained a success rate of 97.10%, with the FA+ELM model achieving an impressive 97.25%. The results given by using the hybrid model outperformed the results obtained using ELM only.

S. K. Lakshmanaprabuet al. [23] introduced a method for improving a classification model for the Health Medical Record Collection system. The introduced model used a Map-Reduce framework to reduce the size of trained data. The optimal features were selected using the Improved Dragonfly algorithm. Finally, a Random forest classifier was used for classification. Salem Alelyani [24] proposed an ensemble approach utilizing the bagging technique to improve feature selection stability in medical datasets through data variance reduction. Authors conducted experiments on four microarray datasets characterized by high dimensionality and small sample sizes. For each dataset, five well-known feature selection algorithms were applied to select varying numbers of features. The proposed technique demonstrated a significant improvement in selection stability, with increases ranging from 20% to 50% across all datasets.

III. METHODOLOGY

In this section, we introduce two phases of the proposed algorithm. One challenge for medical data is the small number of samples compared to many features, so relevant features need to be selected. The first phase produces the optimal features that are used as inputs to the second phase, as shown in the following figure. Firstly, the model performs some preprocessing steps into data to get more accurate results. The missing data are removed, and the data sets are scaled to normalize all values. Then, ten subsets of data are generated using 10-fold cross-validation from the data set. These ten subsets are divided equally and randomly. For each subset, the best features (f) are selected using the Hybrid method for feature selection with the optimized method. This hybrid method aims to reduce the number of input features that keep the classification accuracy as high as possible. The hybrid method is performed by first using a filter method, such as information gain, which is used for ranking features, after ranking, employ optimal search algorithms to find the best subset of features. The optimal methods used are Particle Swarm Optimization (PSO), BAT Algorithm, Firefly Algorithm and Elephant Herding Optimization. Finally, the best features are collected using the voting method. The voting method in feature selection is an effective way to enhance the robustness of the final feature set. The given features are used as input to the second phase Fig. 1.

The second phase performs an enhanced k-nearest neighbor's algorithm (k-NN) to deal with uncertain data and also to perform a classification that gives the best accuracy. The Roc curve is used to select the best K for KNN. The steps used to select best K are listed below:

- Data Preparation: Split data into training set and testing set.

- Model Training: The KNN algorithm is fitted multiple times with different K values (1 to 25). Then the model is used to predict probabilities of the positive class on the test dataset.
- ROC curve generation: For each predicted probability, True Positive Rate (TPR) and False Positive Rate (FPR) are computed to construct ROC curve.
- Area Under Curve (AUC) calculation: For each K, AUC is calculated using TPR and FPR.
- Select Best K: Find the KKK value that corresponds to the highest AUC in the list.

Then, the KNN method with selected K is used to handle uncertain data that may cause the mining results to be wrong or unreliable. This is done by updating both the training set and testing set with objects that cause imbalance. Finally, a comparison between the proposed method and previous works is performed in Fig 2.

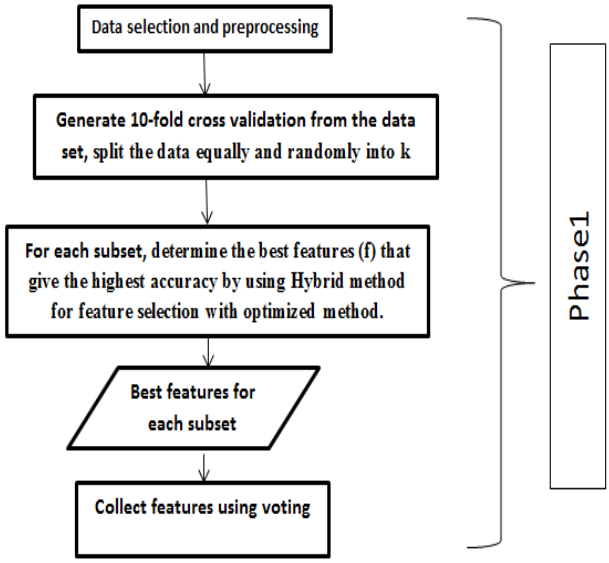

Fig. 1: steps for the first phase in the proposed method

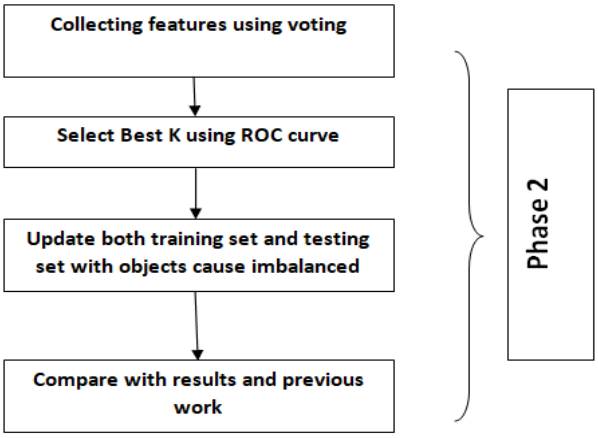

Fig. 2: steps for the second phase in the proposed method

IV. EXPERIIMENTAL PLATFORM

The proposed method has been implemented using R language to generate results running on 64-bit Windows 7 Ultimate with 4 G.B Ram and Intel® Core™ i5 processor.

V. DATA SAMPLE

In this study, medical datasets are used to analyze the performance of the proposed method. The datasets are highly dimensional and have very small sample sizes, and they are publicly available in feature selection @ ASU. The summary of the used datasets is described in Table 1

VI. EXPERIMENTAL RESULTS

In this section, the Results of the proposed framework are analyzed. The results are divided into two parts. The first part analyses the result given in enhanced feature selection with traditional KNN. The second part analyses the result of enhanced feature selection with enhanced KNN.

A. Results from Enhanced Feature selection with traditional KNN

The following analyses describe the result of the first phase using a combination of information gain and optimized search methods to get the features that give the highest accuracy for the used data set. First, some preprocessing steps have been done for the dataset, such as handling missing values and normalizing the dataset. Then, the data are divided into ten subsets, and the best features are selected for each subset. Finally, features using voting are shown in Table 2, Table 3 and Table 4. Table 2 shows the result of the blood dataset, Table 3 shows the result of the colon dataset and table 4 shows the result of the leukemia dataset.

Table 1: Used Data Set

Data set	#objects	#features	Keywords
Blood	89	2759	Discrete, binary
Colon	62	2000	Discrete, binary
leukemia	72	12582	Discrete, binary

Table 2: Blood dataset results

Method used for feature selection	# features
PSO + IG	27
Bat + IG	29
Elephant + IG	29
Firefly +IG	29

Table 3: Colon dataset results

Method used for feature selection	# features
PSO + IG	27
Bat + IG	27
Elephant + IG	26
FireFly +IG	29

Table 4: leukemia dataset results

Method used for feature selection	# features
PSO + IG	194
Bat + IG	205
Elephant + IG	224
Firefly +IG	219

After selecting the best subset of features, the traditional KNN method is used for classification through 10-fold cross-validation. The result of classification accuracy for the four feature methods is described for two datasets, which are shown in Fig 3, Fig 4 and Fig 5.

The results show that the Firefly optimized search method gives the best subset of features that give the best accuracy of the model from the four optimized methods for KNN classification methods.

The results listed in [24] are compared with the obtained result to make the comparison. The previous studies take the random number of features with a threshold =50. Fig 5 shows that the proposed method gives better accuracy.

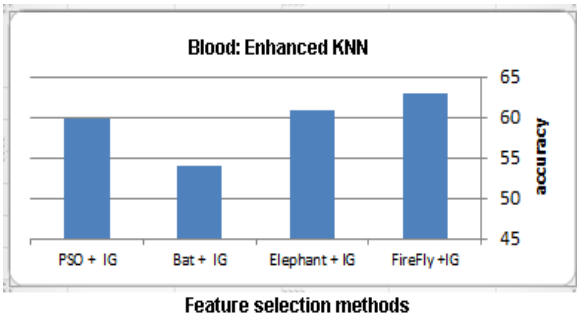

Fig. 3: Blood dataset results

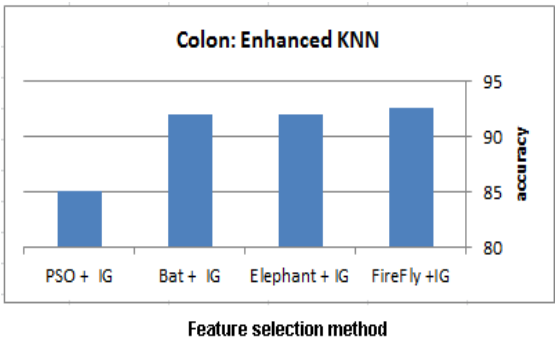

Fig.4: Colon dataset results

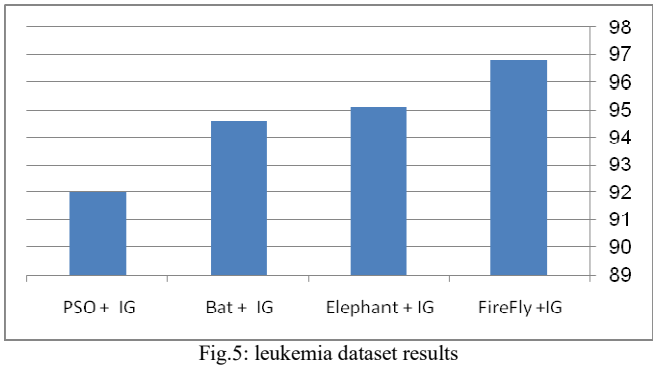

Fig.5: leukemia dataset results

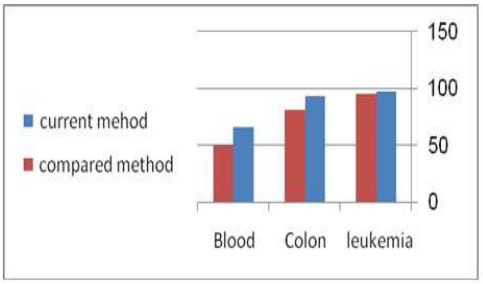

Fig.5: Comparison with the previous method

B. Results for enhanced KNN

In this section, an improvement in traditional KNN is executed to deal with a limitation on the traditional method. The first one is updating the training and testing sets if uncertain objects are found. The second one is that instead of using a random K value, the optimal K value that performs best accuracy is extracted. After selecting the best features from the previous phase, the algorithm is run through 10-fold cross-validation with updated KNN, as shown in Fig 2, and by using the ROC curve to select the best K. ROC is used to select the optimal model using the largest value as shown in Fig 6, Fig 7 and Fig 8.

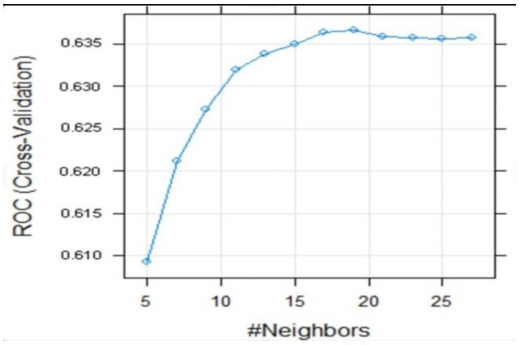

Fig.6: Roc Curve for Blood dataset

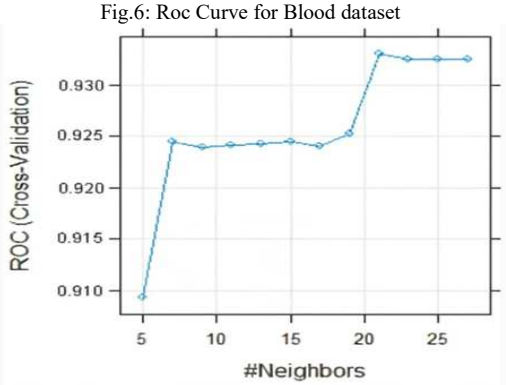

Fig.7: Roc Curve for Colon dataset

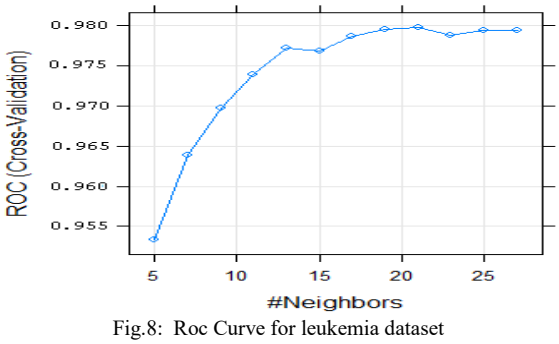

Fig.8: Roc Curve for leukemia dataset

The ROC curve is used to select the optimal model using the largest value from the obtained result. The final value used for the model in the blood dataset is K=19, with model accuracy=63.68%, better than the accuracy obtained from traditional KNN, which is 63%. The final value used for the model in the colon dataset is K=21, with model accuracy=93.3%, which is better than the accuracy obtained from traditional KNN, which is 92.2%. The final value used for the model in leukemia dataset is K=21 with model accuracy =97.9% which the accuracy obtained from traditional KNN =96.2%.

VII. Conclusion

This study proposed a hybrid method to enhance KNN classification method performance in large-scale medical data. The proposed method handles the issues in traditional KNN in dealing with large medical data. The proposed technique consists of two phases. The first phase is to select optimal features from given features. Four optimized methods with a combination of Information Gain are used to select the best feature subset that gives the best performance.

The second phase is to enhance the traditional KNN method. The weighted KNN method is used to handle certain data. Then, to select the optimal K that gives the best model, the ROC curve is used to select K that gives the best model. Also, the proposed model deals with the confusion that can occur with uncertain data; this is done by updating both the training set and testing set with objects that cause imbalance. The results indicate that the proposed framework performs better than other methods.

References

[1] Enas M.F. El Houby, "A survey on applying machine learning techniques for management of diseases," Journal of Applied Biomedicine, Volume 16, Issue 3,2018, Pages 165-174, ISSN 1214-021X, https://doi.org/10.1016/j.jab.2018.01.002.

[2] X. Wu, X. Zhu, G. Q. Wu and W. Ding, "Data Mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, Vol. 26, 2014, Pages 97–107.

[3] Mrutyunjaya Panda, “Elephant search optimization combined with deep neural network for microarray data analysis” , Journal of King Saud University - Computer and Information Sciences, Volume 32, Issue 8, 2020, Pages 940-948, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2017.12.002.

[4] J.C. Miraclin Joyce Pamila, R. SenthamilSelvi, P. Santhi, T.M. Nithya, Ensemble classifier based big data classification with hybrid optimal feature selection, Advances in Engineering Software, Volume 173, 2022, 103183, ISSN 0965-9978, https://doi.org/10.1016/j.advengsoft.2022.103183.

[5] Z. Ma, L. T. Yang and Q. Zhang, "Support Multimode Tensor Machine for Multiple Classification on Industrial Big Data," in IEEE Transactions on Industrial Informatics, vol. 17, no. 5, pp. 3382-3390, May 2021, doi: 10.1109/TII.2020.2999622.

[6] Nicoletta Dessì, Barbara Pes, Similarity of feature selection methods: An empirical study across data intensive classification tasks, Expert Systems with Applications, Volume 42, Issue 10, 2015, Pages 4632-4642, https://doi.org/10.1016/j.eswa.2015.01.069

[7] Newton Spolaôr, Everton AlvaresCherman, Maria Carolina Monard, Huei Diana Led, A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach, Electronic Notes in Theoretical Computer Science, Volume 292, 2013, Pages 135-151, https://doi.org/10.1016/j.entcs.2013.02.010

[8] K. J. D'souza and Z. Ansari, "Big Data Science in Building Medical Data Classifier Using Naïve Bayes Model," 2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Bangalore, India, 2018, pp. 76-80, doi: 10.1109/CCEM.2018.00020.

[9] Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, Shichao Zhang, Efficient kNN classification algorithm for big data, Neurocomputing, Volume 195, 2016, Pages 143-148, https://doi.org/10.1016/j.neucom.2015.08.112.

[10] Xiaojia Song, Tao Xie, Stephen Fischer, Accelerating kNN search in high dimensional datasets on FPGA by reducing external memory access, Future Generation Computer Systems, Volume 137, 2022, Pages 189-200, https://doi.org/10.1016/j.future.2022.07.009.

[11] Haiyan Wang, Peidi Xu, Jinghua Zhao, Improved KNN algorithms of spherical regions based on clustering and region division, Alexandria Engineering Journal, Volume 61, Issue 5, 2022, Pages 3571-3585, https://doi.org/10.1016/j.aej.2021.09.004.

[12] Mohamed A. Mahfouz, Amin Shoukry, Mohamed A. Ismail, EKNN: Ensemble classifier incorporating connectivity and density into kNN with application to cancer diagnosis, Artificial Intelligence in Medicine, Volume 111, 2021.

[13] Jayasri N.P., R. Aruna, Big data analytics in health care by data mining and classification techniques, ICT. Express, Volume 8, Issue 2, 2022.

[14] Amartya Hatua, Badri Narayan Subudhi, Veerakumar T., Ashish Ghosh, Early detection of diabetic retinopathy from big data in hadoop framework, Displays, Volume 70, 2021.

[15] Aishwarya Mujumdar, V Vaidehi, Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science, Volume 165, 2019, Pages 292-299, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.047.

[16] Shafi, A.S.M., Molla, M.M.I., Jui, J.J. et al. Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques. SN Appl. Sci. 2, 1243 (2020). https://doi.org/10.1007/s42452-020-3051-2

[17] ElhamNazari, Mehran Aghemiri, Amir Avan, Amin Mehrabian, HamedTabesh, Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data, Gene Reports, Volume 25, (2021) ,101419,ISSN 2452-0144, https://doi.org/10.1016/j.genrep.2021.101419

[18] Hanaa Salem, Gamal Attiya, Nawal El-Fishawy, Classification of human cancer diseases by gene expression profiles, Applied Soft Computing, Volume 50, 2017, Pages 124-134, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2016.11.026

[19] Md. AlaminTalukder, Md. Manowarul Islam, Md Ashraf Uddin, Arnisha Akhter, KhondokarFida Hasan, Mohammad Ali Moni, Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning, Expert Systems with Applications, Volume 205.

[20] Vahid Nosrati, Mohsen Rahmani, An ensemble framework for microarray data classification based on feature subspace partitioning, Computers in Biology and Medicine, Volume 148, 2022.

[21] Namrata Singh, Pradeep Singh, A hybrid ensemble-filter wrapper feature selection approach for medical data classification, Chemometrics and Intelligent Laboratory Systems, Volume 217, 2021, 104396, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2021.104396.

[22] Yilmaz Kaya, Fatma Kuncan, “A hybrid model for classification of medical data set based on Factor Analysis and Extreme Learning Machine: FA+ELM, Biomedical Signal Processing and Control, Volume 78, DOI:10.1016/j.bspc.2022.104023

[23] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, Abdul Wahid Nasir, V. Vijayakumar, Naveen Chilamkurti, "Random forest for big data classification in the internet of things using optimal features", International Journal of Machine Learning and Cybernetics, 2019

[24] Salem Alelyani, "Stable bagging feature selection on medical data", Journal of Big Data, Volume 8, 2021, https://doi.org/10.1186/s40537-020-00385-8