An Efficient SVM-Based Feature Selection Model for Cancer Classification Using High-Dimensional Microarray Data

PASSENT EL KAFRAWY, (Senior Member, IEEE), HANAA FATHI, MOHAMMED QARAAD, AYDA K. KELANY, AND XUMIN CHEN

1Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Shibin Al Kawm 32511, Egypt
2School of Information Technology and Computer Science, Nile University, Giza 16453, Egypt
3Computer Science Department, Faculty of Science, Amran University, Amran, Yemen
4Department of Computer Science, Faculty of Science, Abdelmalek Essaadi University, Tetouan 93000, Morocco
5Department of Genomic Medicine, Cairo University, Giza 12613, Egypt
6Department of Nephrology, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China

Corresponding author: Xumin Chen (chenxumin860930@163.com)

This work was supported in part by the Natural Science Foundation of Zhejiang Province under Grant LQ21H050008, in part by the New Technologies and Products Projects of Zhejiang Health Committee under Grant 2021PY054, and in part by the Basic Scientific Research Projects of Wenzhou Science and Technology Bureau under Grant Y2020026.

ABSTRACT Feature selection is critical in analyzing microarray data, which has many features (genes) or dimensions. However, with only a few samples the large search space and time consumed during their selection make selecting relevant and informative genes that improve classification performance a complex task. This paper proposed a hybrid model for gene selection known as (SVM-mRMRe), the proposed model provides a framework for combining filter-based, ensemble, and embedded methods to select the most relevant and informative genes from high-dimensional microarray data by fusing embedded SVM coefficients (features ranking) with ensemble mRMRe. Eight of the most commonly used microarray datasets for various types of cancer were used to evaluate the model. The selected subset feature is evaluated by four different types of classifiers: random forest (RF), multilayer perceptron (MLP), k-nearest neighbors (k-NN), and Support Vector Machine (SVM). The experimental results show that the proposed model reduces time consumption and dimensionality and improves the differentiation of cancer tissues from benign tissues. Furthermore, the selected genes for the brain cancer dataset are biologically interpreted, and it agrees with the findings of relevant biomedical studies and plays an important role in patient prognosis.

INDEX TERMS Cancer classification, feature selection, genomic microarray data, support vector machine, ensemble minimum redundancy–maximum relevance.

I. INTRODUCTION

One of the leading causes of death worldwide is cancer [1] Microarray-based gene expression profiling has proven to be an effective technique for cancer diagnosis, prognosis, and treatment [2]. DNA microarray technology is a significant tool that enables researchers to track the level of gene expression in an organism [3]. Microarrays measure the interactions of thousands of genes simultaneously and create a global picture of cellular function [4]. However, analyzing DNA microarray data is difficult for a variety of reasons. First, DNA microarray experiments usually produce many features for a small number of patients, resulting in a dataset with a high dimension. A small number of samples contains several hundred or even thousands of genes (features). Second, the classification of the microarray data, computationally complex and so requires efficient and fast classification algorithms.

Third, Gene expression data is highly complex; genes are directly or indirectly correlated with one another, making classification a difficult task that typically necessitates the use of a powerful and accurate feature selection technique. A robust feature selection method and enough classifiers are required for gene recognition or disease diagnosis using DNA microarray data to overcome these limitations.
The goal of gene (feature) selection is to reduce the complexity of the feature space while also identifying a small subset of distinct genes from a larger set, resulting in not only classification accuracy performance but also biologically meaningful insights [5]. The main aim of this study is to select a subset of informative and relevant genes that accept the findings of related biomedical research. At the same time, eliminating irrelevant or redundant genes and improve the classification performance of high dimension microarray data.

This research paper presents a hybrid feature selection model called (SVM-mRMRRe) that combines different methods ensemble minimum redundancy–maximum relevancy (mRMR) feature selection [6] and support vector machine technique (SVM) as an embedded method. For evaluating the proposed (SVM-mRMRRe) model, eight of the most frequently used microarray datasets for various types of cancer are used. The (SVM-mRMRRe) model is evaluated using four different classifiers: SVM [7], random forest (RF) [8], k-nearest neighbor (KNN) [9], and multilayer perceptron (MLP) [10]. According to the experimental results, the proposed method outperforms the existing standard algorithms regarding classification accuracy and execution time. Furthermore, the genes selected using the brain cancer dataset are biologically interpreted, matching the results of related biomedical studies.

This paper’s main contributions are as follows:
- The most informative and relevant genes are subjected to the proposed SVM-mRMRRe model (features).
- The proposed model is compared to the current SVM-RFE method. The findings show that with SVM-RFE, feature selection takes a long time. However, our proposed model solves this problem by incorporating the following stages:
  - In the first stage, the linear SVM is used as a features (genes) selector, considering feature interaction. The SVM output subset features are then fed into a support vector machine that performs recursive feature elimination and cross-validation (SVM_RBF_CV) in the second stage. As a result, a preliminary list of informative features is generated.
  - The ensemble mRMRRe selects non-redundant and relevant genes to the biological context, leading to more detailed biological interpretations. Later, the output of the gene’s subset is combined with SVM_RBF_CV, and then a voting process is applied to get the unique, informative genes with high relevance and minimum redundancy.
  - The selected subset features of (brain cancer) are biologically interpreted, and it agrees with the outcome of relevant biomedical studies.
- We also present a comprehensive review of various filter and classification methods related to working, particularly for cancer microarray data analysis.

The following is how the paper is organized. The second section is a book review. The procedures used are described in detail in Section 3. The proposed model is presented in Section 4. Section 5 delves into the experimental findings based on publicly available cancer microarray datasets. The conclusion is found in Section 6.

II. RELATED WORK

In recent years, many significant research efforts have been produced to study the cancer microarray data classification using different feature selection techniques, with feature selection playing an important role in cancer classification. As a context for the research discussed in the paper, we provide an overview of this work. Table 1 summarizes some of the previous research methods for microarray cancer classification.

Cancer classification accuracy is considered in all these previous studies without disclosing biological information on the cancer classification process. The SVM-mRMRRe model aims to close the gap between the classification and biological interpretation of cancer by improving accuracy and selecting significant genes that agree with pertinent biomedical studies.

III. MATERIALS AND METHODS

A. FEATURE SUBSET SELECTION

Feature subset, also known as gene subset collection, excludes no longer relevant or redundant features. In certain instances, this is an NP-hard problem (nondeterministic polynomial time hard [19].) The subset of features chosen should obey Occam’s razor theory and have the best value in terms of any objective function. There are three different kinds of feature selection algorithms [20]:

a) Filters extract features from data without prior information, and Filters function without considering the classifier. Therefore, they are highly effective in terms of computation. They are split into two categories: multivariate and univariate processes. Relationships between features can be discovered using multivariate techniques. A multivariate approach is (mRMR).

b) Wrappers evaluate which features are useful using machine learning techniques. Wrappers are best at feature selection because they practice and measure the feature space, considering the model hypothesis. The wrappers’ main drawback is computational inefficiency as a result of this.

c) Embedded approaches incorporate the steps of feature selection and classifier development. In terms of computational efficiency, embedded approaches outperform wrappers, but they allow classifier-specific judgments that do not fit with any other classifier. The SVM method of recursive feature elimination RFE) is embedded.

B. SUPPORT VECTOR MACHINES

SVM is a classification algorithm based on mathematical learning theory [21], [22]. SVM (Support Vector Machine) has long been praised for its superior classification efficiency
### TABLE 1. Review of previous studies on the cancer microarray data classification.

| RefNO | Author                        | Method           | Remark                                                                                                                                                                                                 |
|-------|-------------------------------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [11]  | Raj, D. D., & Mohanasundaram, R. (2020) | (BMR’s)          | Proposed a new feature weighting scheme to address the RELIEF family’s common drawbacks. The main benefit of the new variation is that the margin is more resistant to noise and outliers than previous works. As a result, the feature weights can more accurately characterize the local structure. On most test datasets, the proposed weighting scheme outperformed others in terms of classification error. |
| [12]  | Shukla, A. K. (2020)           | (EMPAGA)         | Developed a new hybrid technique for gene selection known as the ensemble multi-population adaptive genetic algorithm (EMPAGA), which can ignore irrelevant genes while accurately classifying cancer. The experimental results demonstrated that the efficacy of SVM is more suitable as a fitness function in the proposed method when performing tumour diagnosis and confirmed that the proposed method successfully selects relevant genes that are highly significant to sample classes and outperforms state-of-the-art methods in terms of classification accuracy with an optimal gene subset. |
| [13]  | Hengpraspoohm, S., & Jungjit, S. (2020) | (EnSNR)          | EnSNR, an ensemble feature selection approach, was proposed for breast cancer microarray data classification gene selection. The ensemble-based feature selection trend is followed by this method. The ‘entropy’ and ‘SNR’ evaluation functions are used to find the relevant, informative features. Following that, the selected feature subset is fed into the GA classifier. To assess the classification model’s prediction accuracy, the well-known 10-fold cross-validation procedure is used. The experiments show that the Ensemble, EnSNR, approach selects fewer features than the Entropy and SNR approaches. Furthermore, improved performance in terms of prediction accuracy is obtained. |
| [14]  | Meftali, S., Dabba, A., & Tari, A. (2020) | (MIM-mMFA)       | The proposed strategy to solve gene selection in microarray data classification was developed, the Mutual Information Maximization - modified Moth Flame Algorithm (MIM-mMFA). (MIM-mMFA) merged mMFA and Mutual Information Maximization (MIM). The MIM-based pre-filtering technique is used to assess gene relevance and redundancy, while the mMFA is used to evolve gene subsets. The experiments show that the proposed algorithm is very efficient and provides better solutions than other algorithms such as NSGAII, Binary DE, BGA, BPSO, and IBPSO. |
| Reference | Authors | Model | Description |
|-----------|---------|-------|-------------|
| [15]      | Mazumder, D. H., & Veilumuthu, R. (2019). | (JNMIF) | This study proposed a feature selection filter algorithm enhancement based on Joe's normalized mutual information and its application to gene selection. The proposed algorithm is tested and implemented on seven benchmark microarray cancer datasets. The proposed algorithm was implemented, and its performance was evaluated in terms of classification accuracy, AUC, and time on seven high dimensional benchmark microarray gene expression datasets using five classification algorithms. The proposed method improved classification accuracy and AUC values while decreasing classification time in all five classifier cases across all seven datasets tested. The proposed method improved classification accuracy and AUC values while decreasing classification time in all five classes. |
| [16]      | Santhakumar, D., & Logeswari, S. (2020). | (ALO) | Created an integrated method for detecting breast cancer in its early stages. The new feature selection method includes three types of algorithms: support vector machine based on recursive feature elimination and grid search (SVM-RFE-GS), support vector machine based on recursive feature elimination and particle swarm optimization (SVM-RFE-PSO), and support vector machine based on recursive feature elimination and genetic algorithm (SVM-RFE-GS) (SVM-RFE-GA). |
| [17]      | Hameed, S. S., Hassan, R(2021) | HDG-select | HDG-select is a novel stand-alone application based on the graphical user interface (GUI) that performs the full functionality of gene selection and classification in high-dimensional datasets. The proposed HDG-select tool employs an efficient combined filter-GBPSSO-SVM algorithm, and it outperforms the other tools in terms of overall performance, accessibility, and functionality. |
| [18]      | Albashish, D., Hammouri, A. I.(2021) | BBO-SVM-RFE | An efficient Binary Biogeography Based Optimization (BBO) based optimizer with Recursive Feature Elimination for Feature Selection (SVM-RFE) method was proposed. BBO-SVM-RFE used the SVM-RFE to improve the quality of the habitats by selecting the most relevant features in each habitat, which are the most relevant features for the classification task and omitting the irrelevant ones. By balancing exploration and exploitation ability, BBO-SVM-RFE can solve FS tasks. As a result, the results demonstrate the BBO's ability to be hybridized with other embedded methods to improve classification task performance. |
and intrinsic feature selection ability. SVMs may be used to pick features as well as classify them. In each round, features that do not lead to classification are omitted until no further change in classification is feasible [23]. In our model paper, we use linear SVM as a simple gene (feature) selector due to the high dimensionality of microarray data.

**Linear kernel**
\[ k(x, y) = \langle x, y \rangle \]  

For a linear kernel SVM, where \( x \) and \( y \) are points in a \( d \)-dimensional Euclidian space, the margin width can be determined using (2)-(3):

\[ \omega = \sum_{i=1}^{N_s} \alpha_i y_i x_i \]  
\[ \text{margin width} = 2/ \| \omega \| \]

where \( N_s \) denotes the number of support vectors, which are the training samples of \( \theta < \alpha_i \leq c \).

**C. SUPPORT VECTOR MACHINE RECURSIVE FEATURE ELIMINATION AND CROSS-VALIDATION (SVM-RFE-CV)**

Guyon *et al.* proposed SVM-RFE for ranking genes from gene expression data for cancer classification [24]. The SVM-RFE algorithm produces a ranking coefficient dependent on the SVM’s weight vector during preparation, eliminating the signature attribute with the smallest ranking coefficient in each iteration until all signature attributes decrease order. Small variations in the training set may cause the feature exclusion process to fail; features extracted from the training set which not perform well in an independent testing set. Zhang *et al.* [25] used a leave-one-out cross-validation approach to enhance the reliability and robustness of SVM-RFE. The following is a summary of the SVM Recursive Feature Elimination approach based on Cross-Validation (SVM-RFE-CV):

Enter the training samples \( \{x_i, y_i\}, y_i \in \{-1, +1\} \). The \( R \) feature ordering set is the output feature ordering set.

1) The initialization processes. D. the function ordered set \( R = 0 \), the initial feature set \( S = 1, 2, \ldots \)
2) Repeat the process until \( R \) equals 0.
   a) Get the applicant feature set and the instruction set.
   b) To get \( \omega \), train the SVM classifier.
   c) Determine the ranking of the classification criteria:

\[ c_k = \omega_2, k, \quad k = 1, 2, \ldots, |S| \]  

\[ \text{arg min } c_k \]

\[ \text{e) Update feature set } R = \text{PUR}. \]
\[ \text{f) If you’re looking for a special route, Delete this function from } S, \text{ rendering } S = S/P. \]

**D. ENSEMBLE MINIMUM Redundancy–MAXIMUM RELEVANCE (mRMR) FEATURE SELECTION**

The mRMR is a filter-type feature selection approach that maximizes the correlation between features and categorized variables while minimizing the correlation between features to obtain the best feature collection. The issue is that, like all feature selection algorithms in a low sample-to-dimensionality ratio environment, mRMR produces difficult-to-interpret results. The Ensemble (mRMR) feature selection is a variation of mRMR that creates various feature sets rather than a single feature list. Also, the package provides a function for calculating a mutual information matrix (MIM) using the necessary estimators for each variable type. Small variations in sample data frequently result in radically different sets of chosen features, so the effects are highly unpredictable. Paraphrase formalized by the mRMR methodology [6], as applied in the mRMR classic function, allows rapid detection of important and non-redundant features [26]. In the set \( S \), the most important and least redundant gene \( l \) is:

\[ \text{arg max } X_l \in S \frac{R_S}{Q_{S,l}} \]  

Two ensemble methods were used in the Ensemble feature selection (mRMR): exhaustive and bootstrap ensemble mRMR. The exhaustive mRMR heuristic extends the mRMR heuristic by beginning several feature selection procedures with the \( k > 1 \) most important feature. Then, \( k \) mRMR solutions are created in parallel, with the first feature guaranteed to be different. The bootstrap variant resamples the original dataset (with replacement) to produce \( k \) bootstraps and then performs classical mRMR feature selection for each bootstrapped dataset in parallel, yielding \( k \) mRMR solutions. The proposed model applies SVM-RFE-cv to the SVM output subset genes, mRMR to the original data set, then shuffles the output subset genes of both algorithms, creates a voting process for the resulted subset gene (features) and obtains the final informative subset feature with high relevance, high importance, and minimal redundancy.

**IV. THE PROPOSED MODEL (SVM-mRMR)**

The proposed model SVM-mRMR is defined in this section (shown in Fig. 1) for Identifying Informative genes from high dimensional microarray data. The proposed SVM-mRMR model has two stages. In the first stage (Inner election), the data were partitioned using the k-fold cross-validation technique (\( k = 8 \)) to avoid overfitting problems. The training folds are used for training the SVM classifier, and the testing part is used to evaluate the final model. In this stage, the SVM classifier is used as a feature selector through the following steps:

A. SVM generates coefficients (weights) for each gene during training; coefficients (weights) will be used for the prediction of class targets for unseen (test) data.

B. We have eight different weight vectors considering that the weight for the same gene will be different from
one fold to another. To overcome the Cutpoint Partition Problem, which represents the threshold value, we calculate the means of coefficients(weights) vector for each fold separately and use it as a threshold (Cutpoint). The genes with a coefficient less than the mean in each iteration are removed. The genes with coefficients(weights) bigger than means are considered important ones. The genes are filtered according to (means), then the new means are calculated for the filtered genes in each fold. The genes with coefficients(weights) bigger than new means are selected (importance genes).

C. From process two we have eight different weight vectors, that has a redundancy gene in 8 folds, in this process we make the inner election in two-stage for important genes first: merge all genes in the eightfold...
TABLE 2. Gene microarray dataset.

| Dataset | Disease                  | No. Samples | No. Features |
|---------|--------------------------|-------------|--------------|
| D1 [27] | Prostate                 | 102         | 12600        |
| D2 [28] | Lung                     | 181         | 12533        |
| D3 [29] | Brain                    | 28          | 1070         |
| D4 [30] | Central nervous system (CNS) | 62          | 2000         |
| D5 [31] | Breast                   | 60          | 7129         |
| D6 [32] | Leukemia (ALL, AML)     | 72          | 7129         |
| D7 [2]  | Ovarian                  | 253         | 15155        |

and select unique genes, second: each gene in the list of unique genes are ranked according to the voting process in which for example the gene A in fold 0 will take one vote and if its exit in fold one it will take two-vote and so on for all genes.

D. From process three, we get a vector with the gene name and its rank. We suggest that the new threshold check the condition that gene _ importance > 7(represented in all N Folds) is the final_svm_genes. One of the advantages of the SVM is considering the interaction between genes in the training process, but not irrelevant or redundant. To overcome that, we use the wrapper method mRMRe in the second stage.

E. in the second stage, mRMRe is applied on the original microarray data, choosing threshold to mRMRe is considered a challenge, so we suggest that the threshold = final_svm_genes from the previse process as in Figure 1.

F. From processes 4 and 5, we have two gene lists. To select the most relevant genes to the target class, we merge the two lists of genes (final_svm_genes, mRMRe-genes); we suggest an arbitration process that has two stages:

1. Merge the two lists (final_svm_genes, mRMRe-genes) and select unique genes.
2. A voting process, in which the gene, A, takes one vote if it exists in the first list (final_svm_genes), and if it exists in fold one, it will take two votes if it exists in found in the second list (mRMRe-genes). We suggest that the new threshold will be the genes that have a voting value >= 1.

G. The SVM-RFE-CV is applied to select the final subset genes with high performance. The final subset genes are genes with high importance and informative, high relevance, and minimum redundancy, the detailed description of SVM-mRMRe implementation is shown in Algorithm 1.

V. SIMULATIONS AND PERFORMANCE EVALUATION

The experimental validation of the model proposed in this paper is the focus of this section. A PC with the following specifications was used to test the SVM-mRMRe model: Intel(R) Core (TM) i5-7500 CPU with 32-bit operating system 4 GB RAM and Windows 7 operating system, as well as frameworks such as NumPy, SciPy, Keras, Matplotlib and Pandas, and the programming language Python 2.7. Many of the tests use stratified 8-fold cross-validation, which should be illustrated. The stratified cross-validation approach means that the proportions of instances belonging to two groups in both the training and test sets are equivalent, so we prefer 8-fold cross-validation. An average ± standard deviation represents the obtained results.

A. DATASET

Table 2 displays eight benchmark microarray datasets of large dimensionality, limited sample size, and binary classification. The databases are related to global cancer analysis, including Colon, Breast, Leukemia, Prostate, Ovarian, Central nervous system, Brain, and Lung Cancer.

B. EVALUATION METHOD

To assess the efficiency of the proposed model in this section, firstly SVM (first stage) used as a features(genes) selector the gene ranked according to its weight(ω) and each feature(gene) having importance value, the feature(gene) less than mean is removed then the unique output features(genes) with importance feed to SVM-RBF-CV (second stage) is well suited to analyzing noisy high-throughput microarray data; it outperforms SVM-RFE in terms of noise robustness and ability to recover informative features, and it can boost prediction efficiency (Area Under Curve) in the testing data set, Using ensemble mRMRe, the optimal output features of SVM-RBF-CV were shuffled with the output features(genes) added in the original results, outperforming the traditional mRMR method in terms of prediction accuracy. They can contribute to richer biological explanations by recognizing genes that are more important to the biological context. The final optimal list of features (genes) in each data set is evaluated output with the four classifiers SVM, KNN, where k is 3, RF, and MLP checked with eight times cross-validation after implementing the voting method in the shuffled list of features (genes). The final optimal list of features (genes) in each data set is high value, descriptive, minimum redundancy, and highest relevance.

TABLE 3. Total number of features (gene) selected and runtime of SVM-mRMRe model.

| Dataset | Original features(genes) | Number of features(genes) selected | Runtime (in sec.) |
|---------|--------------------------|-----------------------------------|-------------------|
| D1      | 12600                    | 65                                | 248               |
| D2      | 12533                    | 292                               | 557               |
| D3      | 1070                     | 16                                | 1                 |
| D4      | 2000                     | 40                                | 2                 |
| D5      | 7129                     | 33                                | 29                |
| D6      | 7129                     | 64                                | 33                |
| D7      | 7129                     | 40                                | 73                |
| D8      | 15155                    | 6                                 | 863               |
Algorithm 1 Pseudo-Code of SVM-mRmRe

**Input:** microarray GeneSet \( \in \{ \text{Brain_Cancer, ALL-AML, CNS, Colon, Gordon, Ovarian, Singh, West}\} \)

**Output:** Optimal Gene_subset

For \( D \in \) Datasets
   Initialize \( S = \) set of \( D \) genes \( \{X_1, \ldots, X_n\} \)
   Ranked set of genes, \( E = \{\} \)
   Stage_genes, \( P = \{\} \)
   Fold_mean \( M = \{\} \)
   Set \( F = \) SplitKfold crossvalidation (\( S, \) folds_num = 8, shuffle = true)
   For \( X_{\text{train}}, X_{\text{test}} \in F \)
   Preprocessing (Xtrain, Xtest).
   Train SVM on Xtrain
   Set \( w = \) SVM_coef (Xtrain)
   Compute w mean \( me_i \)
   Update \( M = M \cup \{me_i\} \)
   Update \( E = E \cup \{w\} \)

End

For \( R \in E \)
   Select the genes \( X^*_i \) with \( w \geq \text{max} (M) \), Stage_one_genes
   Compute Stage_one_genes new mean, new_mean
   Select the genes \( X^*_i \) with \( w \geq \text{new_mean} \), Stage_two_genes
   Update \( P = P \cup \{X^*_i\} \)

End

Merge \( P \) and select unique gene \( X^*_i \) Index, candidate_genes.

For \( g \in \) candidate_genes
   For \( i \in E \)
      If \( g \) in \( i \) then
         \( gvoted++ \)
   End

Select the genes \( g \) with \( gvoted \geq 7 \), inner_election_genes
Compute number of genes in inner_election_genes, \( N \)
Set \( \text{mRmR\_genes} = \text{Ensample\_mRmR} (D, N) \)
Merge \( \text{mRmR\_genes} \) and inner_election_genes and select only unique gene index, candidate_genes

For \( g \in \) candidate_genes
   For \( i \in \{\text{mRmR\_genes, inner_election\_genes}\} \)
      If \( g \) in \( i \) then
         \( gvoted++ \)
   Select the genes \( g \) with \( gvoted \geq 1 \), arbitration_genes
Set \( \text{Optimal\_Genes\_subset} = \text{RFECV} (\text{linearSVM, D, arbitration\_genes}) \)
Evaluate_Optimal_sub_set (Optimal_Genes_subset, SVM, RF, KNN, MLP)

---

Our evaluation included the following four measurements:

1. **Accuracy (ACC)**: is the most commonly used evaluation standard for the proportion of correctly predicted pairs, but using it alone is usually insufficient.
   
2. **Sensitivity (also known as recall)**: is the proportion of true positive pairs correctly defined. (3) **Specificity**, or the proportion of correctly defined negative pairs; (4) **The region under the ROC curve (AUC)**, which is a probability value for correctly classifying one sample; the larger the AUC, the better.

   Accuracy (ACC) = \( \frac{TP + TN}{TP + TN + FP + FN} \)
   
   Sensitivity = \( TP/(TP + FN) \)
   
   Specificity = \( TP/(TP + FP) \)

   \[ AUC = \frac{1 + TPR - FPR}{2} \]

   TP denotes true positive, FP is false positive, TN is a true negative, and FN is a false negative. Based on the confusion matrix, we evaluated the performance of the proposed method and rival gene selection.

   Two statistical testing methods are also used to evaluate the performance of our model. ANOVA [34], which stands for analysis of variance, the goal of the test is to determine whether two or more means are equal. The Friedman test [35] is applied to data with three or more correlated or repeated outcomes with non-normal distribution. The null hypothesis states that the distribution remains constant across repeated measurements.
TABLE 4. Comparison of SVM-mRMRe performance by four classification algorithms.

| Data Set       | Methods          | Classifier | AUC   | Sen  | Spec | Accuracy  |
|----------------|------------------|------------|-------|------|------|-----------|
| D1, Brain      | Embedded         | SVM        | 0.91 ± 0.17 | 0.88 ± 0.22 | 0.94 ± 0.17 | 0.91 ± 0.17 |
| (28,1070)      | SVM-mRMRe        | SVM        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | RF         | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | KNN        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | MLP        | 1.00 ± 0.00 | 0.88 ± 0.22 | 0.94 ± 0.17 | 0.90 ± 0.14 |
| D2, Colon      | Embedded         | SVM        | 0.81 ± 0.10 | 0.77 ± 0.25 | 0.85 ± 0.17 | 0.82 ± 0.09 |
| (62,2000)      | SVM-mRMRe        | SVM        | 0.99 ± 0.03 | 1.00 ± 0.00 | 0.97 ± 0.07 | 0.98 ± 0.05 |
|                |                  | RF         | 0.92 ± 0.07 | 0.92 ± 0.14 | 0.93 ± 0.10 | 0.92 ± 0.06 |
|                |                  | KNN        | 0.91 ± 0.07 | 0.88 ± 0.16 | 0.95 ± 0.09 | 0.92 ± 0.06 |
|                |                  | MLP        | 1.00 ± 0.00 | 0.96 ± 0.11 | 0.97 ± 0.07 | 0.97 ± 0.06 |
| D3, Breast     | Embedded         | SVM        | 0.61 ± 0.15 | 0.69 ± 0.13 | 0.54 ± 0.29 | 0.62 ± 0.15 |
| (49,7129)      | SVM-mRMRe        | SVM        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | RF         | 0.90 ± 0.17 | 0.92 ± 0.14 | 0.88 ± 0.23 | 0.90 ± 0.17 |
|                |                  | KNN        | 0.98 ± 0.06 | 1.00 ± 0.00 | 0.96 ± 0.11 | 0.98 ± 0.06 |
|                |                  | MLP        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| D4, CNS *      | Embedded         | SVM        | 0.62 ± 0.17 | 0.82 ± 0.12 | 0.42 ± 0.26 | 0.67 ± 0.15 |
| (60,7129)      | SVM-mRMRe        | SVM        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | RF         | 0.82 ± 0.15 | 0.97 ± 0.08 | 0.67 ± 0.29 | 0.87 ± 0.10 |
|                |                  | KNN        | 0.99 ± 0.03 | 0.97 ± 0.07 | 1.00 ± 0.00 | 0.98 ± 0.04 |
|                |                  | MLP        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| D5, Prostate   | Embedded         | SVM        | 0.94 ± 0.05 | 0.96 ± 0.07 | 0.92 ± 0.11 | 0.94 ± 0.05 |
| (102,12600)    | SVM-mRMRe        | SVM        | 0.99 ± 0.02 | 1.00 ± 0.00 | 0.98 ± 0.05 | 0.99 ± 0.03 |
|                |                  | RF         | 0.96 ± 0.04 | 0.98 ± 0.05 | 0.94 ± 0.08 | 0.96 ± 0.04 |
|                |                  | KNN        | 0.99 ± 0.02 | 1.00 ± 0.00 | 0.98 ± 0.05 | 0.99 ± 0.03 |
|                |                  | MLP        | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.96 ± 0.07 | 0.98 ± 0.03 |
| D6, Leukemia   | Embedded         | SVM        | 0.94 ± 0.09 | 0.98 ± 0.06 | 0.90 ± 0.19 | 0.94 ± 0.08 |
| ALL-AML        | SVM-mRMRe        | SVM        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| (72,7130)      |                  | RF         | 0.98 ± 0.06 | 1.00 ± 0.00 | 0.96 ± 0.11 | 0.99 ± 0.04 |
|                |                  | KNN        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | MLP        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| D7, Lung       | Embedded         | SVM        | 0.98 ± 0.04 | 1.00 ± 0.00 | 0.97 ± 0.08 | 0.99 ± 0.01 |
| (181,12533)    | SVM-mRMRe        | SVM        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | RF         | 0.97 ± 0.08 | 1.00 ± 0.00 | 0.94 ± 0.17 | 0.99 ± 0.03 |
|                |                  | KNN        | 0.98 ± 0.04 | 1.00 ± 0.00 | 0.97 ± 0.08 | 0.99 ± 0.01 |
|                |                  | MLP        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| D8, Ovarian    | Embedded         | SVM        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| (253,15155)    | SVM-mRMRe        | SVM        | 1.00 ± 0.00 | 0.99 ± 0.02 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | RF         | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
|                |                  | KNN        | 0.99 ± 0.01 | 1.00 ± 0.00 | 0.99 ± 0.03 | 1.00 ± 0.00 |
|                |                  | MLP        | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |

P-value

- ANOVA: 0.009
- Friedman chi-square: 0.002

FIGURE 2. SVM-mRMRe time consumption for all datasets.

C. RESULT AND DISCUSSION

The results of the experiments are presented in this section to evaluate the proposed model (SVM-mRMRe); the experimental findings for the chosen number of features and runtime are summarized in Table 3, Figs. 2 and 3, and the evaluation of the proposed method using RF, KNN, MLP, and SVM is summarized in Table 4, Figs. 4, 5, 6, and 7. Four performance metrics were chosen for result estimation: ACC,
TABLE 5. Present gene accession number and gene description of the selected genes of brain cancer by the proposed model.

| Probe Set | Gene ID | Gene name |
|-----------|---------|-----------|
| 1786_at   | 1786    | MER proto-oncogene, tyrosine kinase (MERTK) |
| 32174_at  | 32174   | SLC9A3 regulator 1 (SLC9A3R1) |
| 33813_at  | 33813   | TNF receptor superfamily member 1B (TNFRSF1B) |
| 35297_at  | 35297   | NDUFAB1 – NADH:ubiquinone oxidoreductase subunit AB1 |
| 34768_at  | 34768   | thioredoxin related transmembrane protein 1(TMX1) |
| 38752_r_at| 38752   | H+ transporting, ATP synthase, mitochondrial Fo complex subunit E(ATP5I) |
| 38774_at  | 38774   | syntaxin 7(STX7) |
| 39721_at  | 39721   | ephrin B1(EFNB1) |
| 37360_at  | 37360   | lymphocyte antigen 6 complex, locus E (LY6E) |
| 1988_at   | 1988    | platelet-derived growth factor receptor alpha (PDGFRα) |
| 31990_at  | 31990   | kinesin family member 17(KIF17) |
| 32269_at  | 32269   | membrane-associated guanylate kinase, WW, and PDZ domain containing 1(MAGI1) |
| 40090_at  | 40090   | Williams-Beuren syndrome chromosome region 22(WBSCR22) |
| 41549_s_at| 41549_s| adaptor related protein complex one sigma two subunits (AP1S2) |
| 41624_r_at| 41624_r| fizzy/cell division cycle 20 related 1(FZR1) |
| 31386_at  | 31386   | immunoglobulin kappa variable 1/OR2-118 (IGKV1OR2-118) |

AUC, sensitivity, and specificity, we performed a statistical p-value test to determine the significance of the results.

To reduce the computational complexity of the problem at hand and select the most informative genes, we ran SVM-mRMRe against each dataset. We obtained the number of optimal selected features, as shown in Table 3 and Figure 3. To reduce the computational complexity of the problem at hand and select the most informative genes, as described before, SVM-mRMRe used more than one stage (Embedded SVM, SVM-RBF-CV, SVM-RBF-CV-mRMRe) to select optimal features, we ran SVM-mRMRe against each dataset and obtained the number of optimal selected features, as shown in Table 3 and Figure 3. The number of genes chosen
by SVM-mRMRe for each microarray gene dataset is shown in Table 3. It should be noted that SVM-mRMRe provides an ordered list of the genes (features) according to the optimal genes with importance, relevant and informative; it is obvious that SVM-mRMRe achieves the highest level of dimensionality reduction by selecting the fewest number of informative genes, the highest dimensional dataset is ovarian cancer with 15155 features (genes) and 253 samples. The optimal subset selected gene by the SVM-mRMRe are six features (genes) from 15155; these six genes are the ranked genes with the highest importance, informative, and relevance.

It is also observed that the SVM-mRMRe model consumes less computational cost in experiments in all data sets, as shown in Figure 2. The lowest runtime is 1 (sec)

| Dataset      | SVM       |        |        |        |        | KNN       |        |        |        |        |
|--------------|-----------|--------|--------|--------|--------|-----------|--------|--------|--------|--------|
|              | precision | recall | fl-score | support |        | precision | recall | fl-score | support |        |
| Breast cancer| 0.0       | 1.00   | 1.00    | 1.00    | 25     | 0.89      | 1.00   | 0.94    | 25     |        |
|              | 1.0       | 1.00   | 1.00    | 1.00    | 24     | 1.00      | 0.88   | 0.93    | 24     |        |
|              | accuracy  |        |         | 1.00    | 49     |           |        | 0.94    | 49     |        |
|              | Macro avg.| 1.00   | 1.00    | 1.00    | 49     | 0.95      | 0.94   | 0.94    | 49     |        |
|              | Weighted avg.| 1.00 | 1.00    | 1.00    | 49     | 0.95      | 0.94   | 0.94    | 49     |        |
| Brain cancer | SVM       |        |         |         |        | KNN       |        |        |        |        |
|              | 0.0       | 1.00   | 1.00    | 1.00    | 14     | 1.00      | 1.00   | 1.00    | 14     |        |
|              | 1.0       | 1.00   | 1.00    | 1.00    | 14     | 1.00      | 1.00   | 1.00    | 14     |        |
|              | accuracy  |        |         | 1.00    | 28     |           |        | 1.00    | 28     |        |
|              | Macro avg.| 1.00   | 1.00    | 1.00    | 28     | 1.00      | 1.00   | 1.00    | 28     |        |
|              | Weighted avg.| 1.00 | 1.00    | 1.00    | 28     | 1.00      | 1.00   | 1.00    | 28     |        |
| CNS          | SVM       |        |         |         |        | KNN       |        |        |        |        |
|              | 0.0       | 1.00   | 1.00    | 1.00    | 39     | 1.00      | 0.87   | 0.93    | 39     |        |
|              | 1.0       | 1.00   | 1.00    | 1.00    | 31     | 0.81      | 1.00   | 0.89    | 21     |        |
|              | accuracy  |        |         | 1.00    | 60     |           |        | 0.92    | 60     |        |
|              | Macro avg.| 1.00   | 1.00    | 1.00    | 60     | 0.92      | 0.94   | 0.91    | 60     |        |
|              | Weighted avg.| 1.00 | 1.00    | 1.00    | 60     | 0.93      | 0.92   | 0.92    | 60     |        |
| Prostate     | SVM       |        |         |         |        | KNN       |        |        |        |        |
|              | 0.0       | 1.00   | 1.00    | 1.00    | 50     | 0.98      | 1.00   | 0.99    | 50     |        |
|              | 1.0       | 1.00   | 1.00    | 1.00    | 52     | 1.00      | 0.98   | 0.99    | 52     |        |
|              | accuracy  |        |         | 1.00    | 102    |           |        | 0.99    | 102    |        |
|              | Macro avg.| 1.00   | 1.00    | 1.00    | 102    | 0.99      | 0.99   | 0.99    | 102    |        |
|              | Weighted avg.| 1.00 | 1.00    | 1.00    | 102    | 0.99      | 0.99   | 0.99    | 102    |        |
| Datasets          | SVM Confusion Matrix | KNN Confusion Matrix | RF Confusion Matrix |
|-------------------|----------------------|----------------------|---------------------|
| Prostate          | ![Prostate SVM](image) | ![Prostate KNN](image) | ![Prostate RF](image) |
| Lung              | ![Lung SVM](image)    | ![Lung KNN](image)    | ![Lung RF](image)    |
| Brain Colon       | ![Brain Colon SVM](image) | ![Brain Colon KNN](image) | ![Brain Colon RF](image) |
| The central nervous system (CNS) | ![CNS SVM](image) | ![CNS KNN](image) | ![CNS RF](image) |
| Breast            | ![Breast SVM](image)  | ![Breast KNN](image)  | ![Breast RF](image)  |

**TABLE 7.** Confusion matrix of SVM, KNN and RF.
TABLE 7. (Continued.) Confusion matrix of SVM, KNN and RF.

|                | SVM          | KNN          | RF           | MLP           |
|----------------|--------------|--------------|--------------|---------------|
| Leukemia (ALL, AML) |              |              |              |               |
|                |              |              |              |               |
|                |              |              |              |               |
|                |              |              |              |               |
| Ovarian        |              |              |              |               |
|                |              |              |              |               |
|                |              |              |              |               |
|                |              |              |              |               |

with 1070 features and 16 genes selected in the brain cancer dataset, while the highest runtime is 863 (sec) for the Ovarian dataset with the highest dimensional. Supplementary number C displays heat maps of the genes chosen in the SVM-mRMRe model.

The classic learning algorithms SVM, KNN, RF, and MLP are used to evaluate the gene classification accuracy of selected optimal genes by the SVM-mRMRe model. The learning algorithms are applied to the newly collected dataset, which only includes the best genes, and the overall accuracy is calculated.

Table 4 and Figure 4 outline the learning accuracy of four classifiers on various feature sets. SVM-mRMRe increases the accuracy of SVM, KNN, RF, and MLP classifiers in most datasets while the accuracy is weighted overall data sets; on the other hand, SVM achieves the highest classification accuracy. However, as previously stated, a single classifier such as SVM is not accurate enough when applied to the problem of gene microarray classification, which typically faces several challenges such as the curse of dimensionality, small sample size datasets, and a large amount of noise and uncertainty. The accuracy of SVM as embedded methods on the original CNS datasets is 0.67 0.15, as shown in Table 4. As we know, accuracy alone is insufficient for model evaluation, so we use three other evaluation matrices: specificity, sensitivity, and AUC, as shown in Table 4 and Figures 5, 6, and 7.

From Fig. 5, it is observed that the SVM-mRMRe improves the sensitivity of SVM, KNN, RF, and MLP classifiers as seen the sensitivity of SVM is the best in most datasets then MLP, SVM has a sensitivity of 1.00 ± 0.00 in breast dataset and 0.69 ± 0.13 in the same original dataset.

From Fig. 6, it is observed that the SVM-mRMRe improves the specificity for most classifiers; SVM has a specificity of 1.00 ± 0.00 in the CNS dataset and 0.42 ± 0.26 in the same original dataset.

From Fig. 7, it is observed that the SVM-mRMRe improves the AUC; AUC is best with both SVM and MLP in most datasets.

Supplementary A and B show the confusion matrix and the recall, precision, f1-score, and support of SVM, KNN, RF, and MLP for all datasets in detail.

1) BIOLOGICAL INTERPRETATION OF BRAIN CANCER

The leading cause of cancer mortality in children is brain cancer, which is also the second leading cause of cancer death in general [36]. According to studies, brain tumors are highly heterogeneous, which poses the main challenge for brain tumor classification and segmentation, and thus diagnosis and prognosis [37]. A subset of genes (features) from the brain cancer data set is biologically interpreted to demonstrate the efficacy of the proposed model in improving both critical items such as classification accuracy and
TABLE 8. Heat maps of the genes selected in the (SVM-mRMRe) model.
selecting genes with important biological backgrounds. Just a few classes of important genes derived from microarray technologies are used for the diagnosis and prognostic purposes of brain cancer after we used the biological portrait (SVM-mRMRe). The aim of (SVM-mRMRe) is to know crucial gene subsets with the maximum outcome feedback accuracy to treat a brain cancer patient. In this segment, the selected group of probe sets could be studied by using the web tool DAVID (Database for Annotation, Integrated Discovery, and Visualization) https://david.ncifcrf.gov/list.jsp \[38\], \[39\]. Table (5) shows the gene name and gene ID from the Entrez probe set. GO Research Tools: Ncbi.nlm.nih.gov/geoprofiles and https://david.ncifcrf.gov/list.jsp are generally considered the most inclusive and fastest-growing public repository for grouping functionally related genes. Following that, it can be shown that the proposed approach is the most effective way to pick a large group of genes for brain cancer pathway detection and prognosis.

VI. CONCLUSION AND FUTURE WORK

Limited sample size, high dimensionality, and high complexity are the key characteristics of microarray data, as well as the main obstacles for researchers performing microarray data analysis. To address this issue, this paper proposes SVM-mRMRe, an efficient SVM-based feature selection model for identifying informative features from high dimensional microarray data. SVM-mRMRe combines a filter, an embedded method, and an ensemble method to select the most informative genes with the least redundancy and the highest relevance. When evaluating the proposed method with three different classifiers, experimental results on eight microarray datasets validated our findings. On most test datasets, the proposed model outperformed others in terms of classification error. Extensive testing revealed that the proposed model has four distinguishing features: (1) high classification accuracy, (2) successful time complexity resolution, and (3) effective informative gene selection, with the biological interpretation
of the selected genes for brain cancer dataset agreeing with the results of relevant biomedical studies. In the future, the bioinformatics Gene networks analysis will be shown many functionally to our studying genes to predict cancer prognosis. Also, this may indicate a new relationship between our genes and other regulated genes to foresee possible functional interactions among them to cancer disease pathways. A comparison between the proposed approach and a hybrid technique depending upon GA and PSO will be investigated.

**APPENDIX A**
See Table 6.

**APPENDIX B**
See Table 7.

**APPENDIX C**
See Table 8.

**REFERENCES**

[1] B. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics,” CA Cancer J. Clinicians, vol. 66, no. 1, pp. 7–30, 2020.

[2] R. T. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,” Science, vol. 286, no. 5439, pp. 531–537, 1999.

[3] N. Abhnegren and H. A. Elshamlian, “A survey on hybrid feature selection methods in microarray gene expression data for cancer classification,” IEEE Access, vol. 7, pp. 78533–78548, 2019.

[4] H. Pang, S. L. George, K. Hui, and T. Tong, “Gene selection using iterative feature elimination random forests for survival outcomes,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 9, no. 5, pp. 1422–1431, Sep. 2012.

[5] P. A. Mundra and J. C. Rajapakse, “SVM-RFE with MRMR filter for gene selection,” IEEE Trans. Nanobiosci., vol. 9, no. 1, pp. 31–37, Mar. 2009.

[6] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. E. Hachem, G. Bonempi, and B. Haibe-Kains, “MRMR: An R package for parallelized mRMR ensemble feature selection,” Bioinformatics, vol. 29, no. 18, pp. 2365–2368, Sep. 2013.

[7] C. S. R. Annavarapu and S. Dara, “Cluster-based hybrid feature selection approach for high dimension microarray data,” Chemometrics Intell. Lab. Syst., vol. 213, Jun. 2021, Art. no. 104305.

[8] G. Dagniew and B. H. Shekar, “Ensemble learning-based classification of microarray cancer data on tree-based features,” Cognit. Comput. Syst., vol. 3, no. 1, pp. 48–60, Mar. 2021.

[9] L. E. Peterson, “K-nearest neighbor,” Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[10] A. M. Abed, S. A. Giraffi, and A. H. Issa, “Quadrate support vector machine and K-nearest neighbor based robust sensor fault detection and isolation,” Eng. Technol. J., vol. 39, no. 5A, pp. 859–869, May 2021.

[11] D. M. D. Raj and R. Mohanasundaram, “An efficient filter-based feature selection model to identify significant features from high-dimensional microarray data,” Arabian J. Sci. Eng., vol. 45, no. 4, pp. 1–12, 2020.

[12] A. K. Shukla, “Identification of cancerous gene groups from microarray data by employing adaptive genetic and support vector machine technique,” Comput. Intell., vol. 36, no. 1, pp. 102–131, Feb. 2020.

[13] S. Hengpraprohm and S. Jungjit, “Gene selection and classification of microarray data method based on mutual information and moth flame algorithm,” Expert Syst. Appl., vol. 166, Mar. 2021, Art. no. 114012.

[14] D. H. Mazumder and R. Veitumuthu, “An enhanced feature selection filter for classification of microarray cancer data,” ETRI J., vol. 41, no. 3, pp. 358–370, Jun. 2019.

[15] D. Santhakumar and S. Logeswari, “Efficient attribute selection technique for leukemia prediction using microarray gene data,” Soft Comput., vol. 24, pp. 1–10, 2020.

[16] S. S. Hameed, R. Hassan, W. H. Hassan, F. F. Muhammadsharif, and L. A. Latiff, “HDG-select: A novel GUI based application for gene selection and classification in high dimensional datasets,” PloS ONE, vol. 16, no. 1, Jan. 2021, Art. no. e0246039.

[17] D. Albashish and A. I. Hammouri, M. Braik, J. Atwan, and S. Sahran, “Binary biogeography-based optimization based SVM-RFE for feature selection,” Appl. Soft Comput., vol. 101, Mar. 2021, Art. no. 107026.

[18] Z. M. Hira and D. F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data,” Adv. Bioinf., vol. 2015, pp. 1–13, Jun. 2015.

[19] D. Santhakumar and S. Logeswari, “An efficient filter-based feature selection technique for leukaemia prediction using microarray gene data,” Expert Syst. Appl., vol. 46, no. 1, pp. 358–370, Jun. 2019.

[20] D. Santhakumar and S. Logeswari, “An efficient filter-based feature selection technique for leukaemia prediction using microarray gene data,” Soft Comput., vol. 24, pp. 1–10, 2020.
PASSENT EL KAFRAWY (Senior Member, IEEE) received the bachelor’s degree from the Computer Science and Engineering Department, The American University in Cairo, Cairo, the master’s degree from the Faculty of Science, Menoufia University, and the Ph.D. degree in computer science and engineering from the University of Connecticut, USA, in 2006, with a focus on computational geometry and artificial intelligence. She joined the Information Technology and Computer Science School, Nile University, in 2019. She has been a Professor, since 2018. Then, she taught at the Eastern the State University of Connecticut, for one year. In 2007, she has worked as an Assistant Professor with the Mathematics and Computer Science Department, Faculty of Science, Menoufia University. In 2011, she joined the Computer Science and Engineering Department, American University, as an Adjunct Professor. She has appointed as an Associate Professor, in 2013. She has been supervising several research studies between Ph.D. and M.Sc. degrees in the field of natural language processing, semantic knowledge, bioinformatics, big data analytics, and knowledge mining and acquisition. She has joined the IBRO School for neurodegenerative physician training organized by ENND and personalized medicine workshop organized by AUC. She has over 45 publications and an editor in three books. Her research interests include software engineering, bioinformatics, big data analytics, machine learning, and cloud computing. She is a member of the Egyptian Society of Language Engineering and the Editor-in-Chief of The Egyptian Journal of Language Engineering.

HANAA FATHI received the B.S. and M.S. degrees in computer science from Menoufia University, in 2006 and 2015, respectively. She attended the 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, in December 2019, and the Fourth Edition of the International Conference on Intelligent Systems and Computer Vision (ISCV 2020). She has five publications in the fields of software engineering and machine learning.

MOHAMMED QARAAD received the B.Sc. degree from the Department of Computer Science, Mutah University, Jordan, in 2006, and the M.Sc. degree from the Department of Science, University Abdelmalek Essaadi, in 2018, where he is currently pursuing the Ph.D. degree. He attended the 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, in December 2019, and the Fourth Edition of the International Conference on Intelligent Systems and Computer Vision (ISCV 2020). His research interest includes optimizing features selection techniques for classification high dimension data.

AYDA K. KELANY received the B.Sc., M.S., and Ph.D. degrees from the Faculty of Science in Genetics and Molecular Biology, Faculty of Science, Cairo University, Cairo, Egypt, in 2006, 2013, and 2016, respectively. She joined Kasar El-Aini Hospital, Faculty of Medicine, Cairo University. She has attended the Seventh Annual Conference of Clinical and Chemical Pathology Department, in 2017. From December 2018 to January 2019, she has supervised the following Modules in Bioinformatics Winter School (Module I: Basic Bioinformatics, Module II: Programming and NGS, Module III: Gene Expression and Function Annotation, and Module IV: Population Genomics) with the Faculty of Medicine, Cairo University. Her research interests include cancer, molecular genetics, gene expression, gene mutations, and prognosis and diagnosis.

XUMIN CHEN received the master’s and Ph.D. degrees from Sun Yat-sen University, Guangzhou, China, in 2011 and 2014, respectively. She is currently an Attending Doctor with the Department of Nephrology, The First Affiliated Hospital of Wenzhou Medical University. Her research interests include data mining, optimization, and machine learning in bio-medical field. She has written one scientific publication in this area.

HANAA FATHI

MOHAMMED QARAAD

AYDA K. KELANY

XUMIN CHEN