IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 9, 2017, accepted November 13, 2017, date of publication November 29, 2017,
date of current version February 14, 2018.

*Digital Object Identifier 10.1109/ACCESS.2017.2778268*

# A Hybrid Feature Selection Method for Complex Diseases SNPs

**RAID ALZUBI[1], NAEEM RAMZAN[1], (Senior Member, IEEE), HADEEL ALZOUBI[1], AND ABBES AMIRA[2], (Senior Member, IEEE)**

[1]School of Engineering and Computing, University of the West of Scotland, Paisley PA1 2BE, U.K.
[2]Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

Corresponding author: Naeem Ramzan (naeem.ramzan@uws.ac.uk)

**ABSTRACT** Machine learning techniques have the potential to revolutionize medical diagnosis. Single Nucleotide Polymorphisms (SNPs) are one of the most important sources of human genome variability; thus, they have been implicated in several human diseases. To separate the affected samples from the normal ones, various techniques have been applied on SNPs. Achieving high classification accuracy in such a high-dimensional space is crucial for successful diagnosis and treatment. In this work, we propose an accurate hybrid feature selection method for detecting the most informative SNPs and selecting an optimal SNP subset. The proposed method is based on the fusion of a filter and a wrapper method, i.e., the Conditional Mutual Information Maximization (CMIM) method and the support vector machine-recursive feature elimination, respectively. The performance of the proposed method was evaluated against four state-of-the-art feature selection methods, minimum redundancy maximum relevancy, fast correlation-based feature selection, CMIM, and ReliefF, using four classifiers, support vector machine, naive Bayes, linear discriminant analysis, and *k* nearest neighbors on five different SNP data sets obtained from the National Center for Biotechnology Information gene expression omnibus genomics data repository. The experimental results demonstrate the efficiency of the adopted feature selection approach outperforming all of the compared feature selection algorithms and achieving up to 96% classification accuracy for the used data set. In general, from these results we conclude that SNPs of the whole genome can be efficiently employed to distinguish affected individuals with complex diseases from the healthy ones.

**INDEX TERMS** Single nucleotide polymorphism (SNP), feature selection, hybrid algorithms, complex diseases, machine learning.

## I. INTRODUCTION

The human genome is the whole set of Deoxyribonucleic acid (DNA) sequence for humans. It consists of approximately three billion base pairs, with more than 99% of nucleotides being exactly matched among the whole population, and less than 1% difference among persons. The majority of these genetic variations occur as Single Nucleotide Polymorphisms (SNPs). SNPs are the most important markers used for mapping diseases with genes. Although most of them are neutral, recent studies have shown that certain SNPs are functional and affect the phenotype, e.g height, skin colour, resistance, infection or responses to drugs, etc. The main advantage that makes SNPs preferable over microarray gene expressions are stability, high frequency and being easier and faster to collect [1]. In this context, many machine learning algorithms have been widely applied for SNP data classification. However, the "*curse of dimensionality*" is the main challenge encountered, in most studies, due to the number of samples (a few hundreds) being significantly smaller than the number of SNPs (up to one million) [1].

Building a model to classify samples as belonging to a healthy or affected individual is one of the main targets of SNP analysis [2]. However, the huge number of SNPs hinders the development of accurate prediction algorithms. Nevertheless the selection of a subset of descriptive and meaningful SNPs is crucial for reducing the time complexity and for increasing the accuracy. As a result, the initial stage of SNP data analysis should be the selection of the most discriminative and informative subset of SNPs, in order to enhance the performance of the classification algorithm and reduce the time requirements [3]. Multiple feature selection methods have been used for this purpose, but have been usually applied only to small numbers of selected genes associated with human disease [3], while only few works have applied the feature selection techniques to the whole genome [4]–[6].

The selection of a suitable feature selection method is crucial for the success of a machine learning-based system. Feature Selection (FS) is the process of significantly reducing the dimensionality of the feature space, while maintaining an accurate representation of the original data. It's main advantages are improved classification performance, reduced learning speeds, facilitating data interpretation, and improved generalization capability of the predictions. Nevertheless, FS algorithms suffer from increased computational complexity, as well as from the need for parameter tuning in order to select the best feature subsets [7].

FS methods are mainly categorised into two types: wrapper and filter. In recent years, two new techniques have also been proposed: the ensemble and the hybrid feature selection methods [7]. Wrapper approaches are classifier dependent, while the filter approach is classifier independent. Wrapper methods rely on a classification algorithm for selecting the optimal subset of features during the training phase. These approaches provide very competitive performance for the particular classifier used in the FS process. However, they are computationally expensive and prone to over-fitting. Applying wrapper methods to SNP data is usually inapplicable because of the high computational times needed, due to the high dimensionality of the data [8], [9]. In contrast to wrapper methods, filter methods do not depend on classifiers. They measure the discriminatory power of features using multiple criteria, such as fisher score, mutual information, and symmetrical uncertainty. The advantages of filter methods compared to other FS methods are that they are faster, more scalable, very efficient, and have high generalization ability. Nevertheless, filter methods usually under-perform compared to wrapper methods [10]. Recently, several hybrid feature selection algorithms have been developed which present the advantages of both filter and wrapper methods, usually by utilising filter approaches followed by wrapper ones [11].

Several algorithms have been developed to classify complex diseases based on SNP datasets. Evans [4] used two filter FS methods, difference sort and standard chi-squared statistical algorithms, along with Support Vector Machines (SVM) with multiple kernel functions as the classification method, reaching an accuracy of 73%. Batnyam *et al.* [5] combined various existing FS techniques: a three stages approach first, the selection of most informative SNPs using: R-value based Feature Selection (RFS) [12], Feature Selection based on Distance Discriminant (FSDD) [13], feature weight based ReliefF [14] and an algorithm based on Feature Clearness (CBFS) [15]. Second, generating an artificial feature from the selected SNPs using Feature Fusion Method (FFM), and third, the classification using an Artificial Gene Making (AGM), SVM and k- Nearest Neighbor (k-NN) classifier. The best accuracy for all tested datasets was achieved using the combination of CBFS and FFM, classified with SVM. The two aforementioned methods were applied on Mental Retardation (MR) and Autism Spectrum Disorder (ASD) datasets. Anekboon *et al.* [6] proposed three hybrid feature selection methods: The Correlation-based Feature Selection method (CBFS) was initially used as a filter and then the selected features were fed to a k-NN, Artificial Neural Network (ANN), and Ridge Regression (RR) classifiers in the wrapper phase. The algorithms were applied on simulated datasets and the CBFS and RR provided the most accurate feature subset.

High dimensionality forced researchers to avoid using wrapper techniques in whole SNP analysis. In the first scenario, researchers tried to solve this problem by increasing the Minimum Allele Frequency (MAF) value of the SNPs to decrease the feature space [16], [17]. That could put the selection algorithm in the risk of losing some valuable SNPs, since low allele frequencies SNPs could have a higher predictive value [18]. In the other scenario, researchers applied their algorithms on SNPs from selected genes in order to reduce the number of features [19]–[21]. This knowledge-driven data processing reduces the search area and limits it to only the SNPs inside the genes that have already been identified as having a high correlation with a given disease [18].

In this work, the authors try to address this problem, by employing the Conditional Mutual Information Maximization (CMIM) method in order to select a subset of the SNPs in the dataset that exhibits balance between SNP relevancy and redundancy. The selected subset is then provided as input to the Support Vector Machine - Recursive Feature Elimination (SVM-RFE) wrapper FS technique. This fusion reduces the redundancy among selected SNPs, leading to smaller subsets of SNPs and to improve prediction accuracy. The proposed method is evaluated on five different genomics datasets publicly available on the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO).

The rest of this paper is organised in three sections. The proposed feature selection and classification framework is described in Section II. Section III discusses the experimental evaluation and obtained results, while conclusions are drawn in section IV.

## II. METHODOLOGY

The proposed framework for analysing complex diseases is presented in FIGURE 1 and consists of three stages: a) a pre-processing stage consisting of data transformation and data refinement, b) a hybrid feature selection stage, and c) a classification stage.

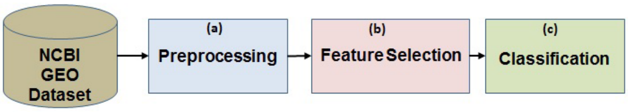
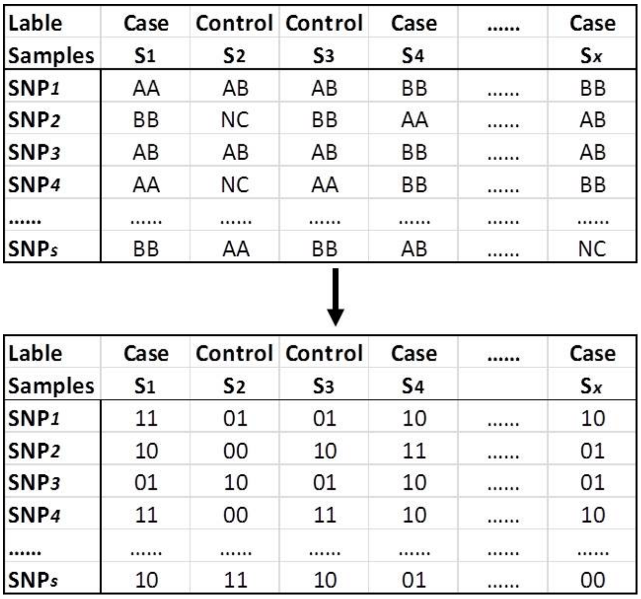


**FIGURE 1. Data flow diagram of the proposed framework.**

### A. PRE-PROCESSING

The pre-processing stage is divided into two sub-stages, i.e. data transformation and data refinement. SNP data are usually encoded as character strings and thus are not always suitable for analysis. As a result, they need to be transformed to numerical form. Transformation is a straightforward process that can be applied in various ways. The first sub-stage in the pre-processing stage of the proposed framework is the lossless transformation of the SNP data by directly converting the values (AA, BB, AB, and NC) to (11, 10, 01, and 00) respectively, as shown in FIGURE 2.



| Lable | Case | Control | Control | Case | ...... | Case |
|---|---|---|---|---|---|---|
| Samples | S1 | S2 | S3 | S4 | | Sx |
| SNP1 | AA | AB | AB | BB | ...... | BB |
| SNP2 | BB | NC | BB | AA | ...... | AB |
| SNP3 | AB | AB | AB | BB | ...... | AB |
| SNP4 | AA | NC | AA | BB | ...... | BB |
| ...... | ...... | ...... | ...... | ...... | ...... | ...... |
| SNPs | BB | AA | BB | AB | ...... | NC |

| Lable | Case | Control | Control | Case | ...... | Case |
|---|---|---|---|---|---|---|
| Samples | S1 | S2 | S3 | S4 | | Sx |
| SNP1 | 11 | 01 | 01 | 10 | ...... | 10 |
| SNP2 | 10 | 00 | 10 | 11 | ...... | 01 |
| SNP3 | 01 | 10 | 01 | 10 | ...... | 01 |
| SNP4 | 11 | 00 | 11 | 10 | ...... | 10 |
| ...... | ...... | ...... | ...... | ...... | ...... | ...... |
| SNPs | 10 | 11 | 10 | 01 | ...... | 00 |

**FIGURE 2.** **SNP data transformation, where *x* is the number of samples and *s* the number of features.**

Secondly, in the data refinement stage, the following steps are applied to the SNP data:

#### 1) REMOVAL OF REDUNDANT SNPs

Redundant SNPs are considered to be the SNPs that consist of the same values for all case and control samples. For example, a given SNP that has the value of "AA" in all case and control samples will not be helpful for training a machine learning model and as a result it is considered as redundant. For example ASD dataset contains less than 1% of redundant SNPs, out of 262338 SNPs only 1721 SNPs were redundant. The remaining SNPs after applying this preprocessing step was 260626 SNPs. In general removing these non-discriminative features leads to reduce data size, efficient usage of storage and reduced computational times for the applied algorithms.

#### 2) MISSING VALUES (NO CALL) REPLACEMENT

Missing values in the SNP datasets are usually marked as "NC". That means that the DNA sequencer was not able to determine the actual value of that allele. All SNPs that consist of more than 10% of "NC" values were discarded, otherwise the "NC" values are replaced by estimating the missing values using the attribute mode (most common value for a given feature in all samples). The aforementioned technique has been widely utilised in the literature [22], [23]. For example ASD dataset contains high amount of NC values more than 11%. The proposed framework consider 29192 SNPs to be removed in this stage. The remaining SNPs after applying this preprocessing step was 231434 SNPs.

### B. PROPOSED FEATURE SELECTION METHOD

After data pre-processing, a FS approach is required to select the most informative subset of features. In general, FS is a significant step in constructing a classification model. It works by limiting the number of input features in a classifier, aiming to have highly predictive and less computationally complex models [8]. CMIM and SVM-RFE are the feature selection algorithms that have been fused in this paper, where CMIM is used as a filter mutual information method and SVM-RFE is used as a wrapper selection method. We present a brief discussion about these methods below.

CMIM was introduced by Fleuret [24]. It is a very fast and efficient multivariate filter FS technique derived from Conditional Mutual Information (CMI). CMIM employs CMI to calculate the amount of relevancy and redundancy. It works by selecting features that maximize their mutual information with the class to predict, conditionally to the response of any feature already selected ($S$). This criterion chooses features different from ones that have already been picked, even if they are individually significant, as they do not carry more information about the class prediction. That will ensure a good trade-off between relevancy and redundancy [24]. A higher value of CMIM means feature $X_n$ is relevant to target $Y$and is highly complementary with another picked feature $X_j$ where $j \in S$ . The criterion is expressed in Eq. 1.

$$CMIM(X_n) = min_{j \in S} I(X_i; Y \mid X_j) \tag{1}$$

The CMIM method attempts to achieve balance between individual power and independence through the comparison of each new feature with the features that have already been selected. A feature $X_0$ will be considered as good only if $I(Y; X_0 \mid X)$ is large for every $X$ already selected, i.e. it is carrying information about $Y$ that has not been captured by any of the already selected $X$.

In this work we employ the fast implementation of CMIM algorithm. While the standard implementation of CMIM calculates CMI (number of samples*number of features) times, the fast implementation uses a feature score during the selection process, which calculates CMI only for the features that carry more information and are not redundant.

The fast CMIM stores a partial score $P_i$ for every feature $s_i$ which is the minimum out of the CMI which appears in the *min* in algorithm (1). Another vector $LU_i$ stores the index of the last picked feature based on the computation of $P_i$.

SVM-RFE is a wrapper method presented by Guyon *et al.* [25], which adopts backward feature elimination. SVM-RFE finds a subset of features that lead to the margin maximization of class separation. It begins with the whole set of features and eliminates the features that are

**Algorithm 1** Pseudocode for the proposed method

```
Input: Samples (X), Labels (Y), Initial feature set (S),
    MaxNumberOfFeatures, FinalNumberOfFeatures
Output: Final feature subset (M)
Set $SelectedFeatures = \emptyset$
for all features $s_i$ in $S$ do
    Calculate $MI_i$
    Set $P_i = MI_i$
    Set $LU_i = 0$
end for
for $k = 1$ to $MaxNumberOfFeatures$ do
    Set $score_k = 0$
    for all features $s_i$ in S do
        while $P_i > score_k$ AND $LU_i < k - 1$ do
            Set $LU_i = LU_i + 1$
            Compute $CMI_{ik}$ between $s_k$ and $s_i$
            Set $P_i = min(P_i, CMI_{ik})$
        end while
        if $P_i > score_k$ then
            Set $score_k = P_i$
            $SelectedFeatures = SelectedFeatures \cup \{s_i\}$
        end if
    end for
end for
Set $N = SelectedFeatures$
Set Ranked Feature Set $M = \emptyset$
while $N \neq \emptyset$ do
    Train a linear SVM: $a = SVMtrain(X, N)$
    Compute the weight vector for $N$: $w = \sum_i (y_i, x_i, \alpha_i)$
    Compute $J = w^2$
    Find the feature with the lowest ranking score:
        $f = \arg\min(J)$
    Set $M = M \cup \{f\}$
    Set $N = N - \{f\}$
end while
return $M$
```

*MI: Mutual Information, CMI: Conditional Mutual Information, LU: Last Used Index, P: Partial Score*

least important for the predictor recursively, in a backward elimination method. The weight vector $w$ was computed as shown in Eq. 2.

$$w = \sum_i (y_i, x_i, \alpha_i) \tag{2}$$

where $y_i$ belongs to the class label of the sample $x_i$ and the summation is taken over all the training samples. The maximum class separation margin is denoted by $\alpha_i$ [26].

SVM-RFE provides a ranked feature list from which a group of top-ranked features can be seletced in order to select the optimal features subset. SVM-RFE utilises a ranking criterion that is closely related to the general SVM classification algorithm. A linear SVM model is trained in each iteration of the feature selection algorithm and the feature with the smallest ranking criterion is removed from the feature set. This process is iterated until all features have been removed from the feature set. The final ranked list is created by sorting the features by the order of removal, with the latest removed features being considered as the most important. Considering that SVM has been succesfull in many Genome-wide Association Studies (GWAS) [27], it is expected to provide a ranking criterion with enhanced performance.

The ranking criterion for feature $k$ is the square of the $k^{th}$ element of $w$, as shown in Eq.3.

$$J(k) = w_k^2 \tag{3}$$

The proposed approach is a two-step hybrid feature selection method that combines the advantages of the two aforementioned filter and wrapper approaches. The first step (CMIM) consists of a pre-filtering process that is used for discarding irrelevant features and for choosing the relevant candidate SNPs. At the second step, the most informative features are selected out of the relevant ones using the wrapper method (SVM-RFE) in order to obtain the optimal SNPs subset. This hybridization is designed towards bridging the gap between the CMIM and SVM-RFE methods by addressing their disadvantages. Since SVM-RFE is not applicable on the whole SNPs dataset due to their huge feature space and the extremely high computational complexity, it is necessary to apply the CMIM pre-filtering step in order to reduce the feature space. Furthermore, SVM-RFE does not take into account the redundancy among SNPs, so irrelevant and redundant features need to be removed before its application. To this end, CMIM was selected due to its ability to carefully address the redundancy among features. More specifically, CMIM does not consider the informativeness of the features individually and does not select a feature unless it provides additional information when combined with other features. This property of CMIM makes it an ideal choice for the first step of the proposed approach. Nevertheless, using only CMIM does not result in high classification accuracy, hence we opted to combine it with the SVM-RFE wrapper method. The proposed approach is effective in removing uninformative and redundant features during the first step, which addresses the exponential computation problem of the second step, leading to the selection of an effective final feature subset.

In this work we follow the procedure outlined in Algorithm 1 and FIGURE 3 in order to select the most informative SNPs. Firstly, the top $N$ candidates which maximize the mutual information between them and the class to predict will be iteratively picked, conditionally to response of any feature already selected. After that the $N$ selected features will be injected into the wrapper SVM-RFE to choose a subset of $M$ features. Leave-One-Out (LOO) cross-validation was used for the evaluation of the feature selection process, with the ranking of the selected SNPs applied at each individual fold. $N$ and $M$ are predefined numbers and $M < N$. The proposed FS algorithm were implemented on Matlab 2016a [28], [29].

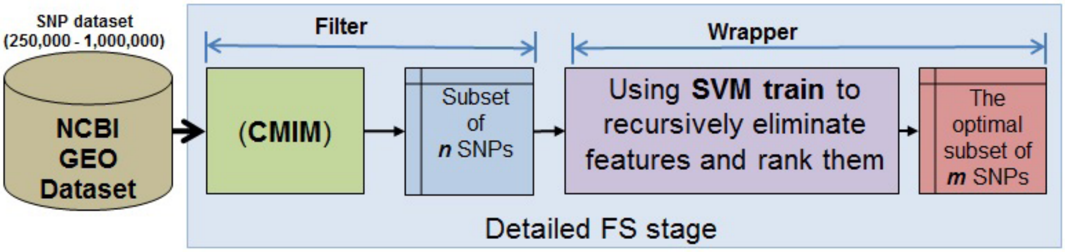


**FIGURE 3.** The proposed FS method.

### C. CLASSIFICATION

Machine learning algorithms aim to develop computer programs that are able to learn by themselves and detect patterns in data, and alter program actions according to new data. Various machine-learning techniques with different characteristics were used in this paper. The main purpose of using these classifiers is to measure the performance of our proposed hybrid feature selection method. The presented classifiers are binary systems trained to detect patterns of data to classify subjects as healthy or affected. The classifiers SVM, Naive Bayes (NB), Linear Discriminant Analysis (LDA) and k-NN were used in this research. In SVM a radial basis function was used as a kernel function. For k-NN, multiple values of $k$ were tested but results are reported only for $k = 6$ since it provided the best performance. Standard NB and LDA classifiers were used.

### D. SNP DATASETS

In this work we used five SNP microarray datasets which are publicly available from NCBI GEO. The GEO is a public repository that archives and freely distributes next-generation sequencing and other genomic data. Datasets consist of labelled samples as case for affected individuals or control for healthy ones, each of them has an identification number and a sequence of SNP alleles. SNP marker can be heterozygous or homozygous. The value AB is given to the SNP marker in the heterozygous case, while, it could be AA or BB in the homozygous case. When the sequencer failed to determined allele value, NC is given.

In this work the five SNP datasets shown in Table 1 were used. GSE67047 series includes SNP data from patients who have sporadic Medullary Thyroid Cancer (sMTC) and juvenile Papillary Thyroid Cancer (PTC). There are 225 individuals in the data set, 96 patients and 129 parents, and approximately 1,000,000 features [30]. GSE9222 series includes SNP data associated with ASD. The dataset consists of 567 individuals, 335 patients and 232 parents, and more than 250,000 features [31]. GSE34678 series includes SNP data from patients who have CC. This study utilizes 124 samples, 62 as a case and 62 as a control and approximately 250,000 features [32]. GSE13117 series includes SNP data associated with MR taken from affected patients and their healthy parents. The dataset consists of 360 individuals, 120 children and 240 parents, and nearly 250,000 features [33]. GSE16619 series includes SNP data related to BC with more than 500.000 SNPs. This study used 111 individuals, 69 as cases and 42 as controls [34].

Applying machine learning techniques on these datasets poses some challenges.

Small sample size is one of the most challenging problems in SNP dataset analysis. The small number of samples and huge number of SNPs greatly impacts an error estimation. It is worth mentioning that selecting the right validation method is essential to estimating the classification error.

For limited SNP values that could be one of four cases (AA,AB, BB or NC), most studies in the literature replace the NC value with attribute mode or AB, reducing the possible values for SNPs to three. The small number of possible SNPs will increase the probability of redundancy in the features, and make the FS method complicated [22].

## III. RESULTS

Supervised classification experiments were conducted in order to evaluate the performance of the examined FS approaches. The SVM, NB, k-NN, and LDA the classification schemes were applied along with Leave-One-Out cross-validation for obtaining the reported results. An ANN

**TABLE 1.** Used SNP datasets.

| **Dataset** | No. of SNP | No. of Samples | Case | Control | Information | Year | Ref |
|---|---|---|---|---|---|---|---|
| GSE67047 | 1,000,000 | 225 | 96 | 129 | Thyroid Cancer | 2016 | [30] |
| GSE9222 | 250,000 | 567 | 335 | 232 | Autism(ASD) | 2008 | [31] |
| GSE34678 | 250,000 | 124 | 62 | 62 | Colorectal Cancer | 2012 | [32] |
| GSE13117 | 250,000 | 360 | 120 | 240 | Mental Retardation | 2009 | [33] |
| GSE16619 | 500,000 | 111 | 69 | 42 | Breast Cancer | 2009 | [34] |

approach was also evaluated but underperformed compared to the other examined classification methods. The whole algorithms were implemented using Matlab 2016a. The prediction rate of the classifiers was evaluated using two measures, the average Accuracy (Acc) (Eq 5) and the F-measure ($F$) (Eq 4), with the $F$ value representing the ability of our proposed method to predict cases. Samples referring to cases of affected individuals are considered as the positive class for computing the following measures:

$$F = 2 \cdot \frac{Pre \cdot Re}{Pre + Re} \tag{4}$$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \cdot 100 \tag{5}$$

where

$$Re = \frac{TP}{TP + FN} \cdot 100 \tag{6}$$

$$Pre = \frac{TP}{TP + FP} \cdot 100 \tag{7}$$

and TP, FN, FP and TN refer to the number of true positive, false negative, false positive, and true negative, predictors respectively.

Table 2 shows the performance of different classifiers on five complex diseases SNP datasets. The results were obtained by conducting our experiments using 100 SNPs as the optimal subset. The proposed method results are compared with the *Acc* and $F$ achieved using the Minimum Redundancy Maximum Relevance (mRMR) algorithm [35], ReliefF [14], Fast Correlation Based Feature Selection (FCBF) [36] and CMIM. As can be clearly seen, the performance of the proposed FS was significantly better and outperformed the compared methods on the given datasets: ASD, MR, CC and TC and BC.

The obtained results show the superiority of our proposed method over the compared algorithms in all tested datasets. A notable increase in the average accuracy was achieved in all given datasets: for example, the accuracy achieved in the BC and CC dataset when using our method is up to 10 % better than that of the best competitors. The proposed algorithm showed consistent performance when it was applied on five different datasets by giving high *Acc* and $F$ value in all tested data.

The results establish that our hybrid model has strong, consistent performance over most classifiers. In BC and TC datasets, all given classifiers showed great performance when evaluating the proposed algorithm. In ASD and MR datasets, k-NN achieved the minimum accuracy out of the other classifiers. A similar performance was observed in the other feature selection methods. Moreover, in the CC dataset, when the most feature selection performed poorly over most classifiers except NB, our model gave excellent result with SVM classifier.

Based on the results shown in FIGURE 4, the dominance of proposed method is confirmed by being able to reach the best accuracy in all cases. The proposed method significantly outperforms mRMR, CMIM, FCBF and ReliefF.

**TABLE 2. Performance comparison between four FS methods ReliefF, FCBF, mRMR, CMIM and our proposed method. Different classifiers were used (SVM, NB, KNN and LDA) to evaluate the selection performance for five different datasets.**

| **MR** | | SVM | KNN | NB | LDA |
|---|---|---|---|---|---|
| FCBF | *Acc* | 70.83 | 69.11 | 71.80 | 67.89 |
| | *F* | 46.69 | 56.44 | 73.56 | 79.82 |
| ReliefF | *Acc* | 65.28 | 66.11 | 65.28 | 63.89 |
| | *F* | 16.49 | 6.90 | 63.16 | 49.80 |
| mRMR | *Acc* | 81.67 | 74.72 | 80.00 | 76.67 |
| | *F* | 76.72 | 69.76 | 81.93 | 73.92 |
| CMIM | *Acc* | 82.83 | 69.44 | 81.67 | 70.56 |
| | *F* | 81.88 | 66.31 | 82.99 | 70.45 |
| Proposed | *Acc* | **85.00** | **75.31** | **83.72** | **79.24** |
| | *F* | **85.21** | **70.45** | **83.42** | **77.21** |
| ASD | | | | | |
| FCBF | *Acc* | 74.28 | 76.11 | 65.28 | 73.89 |
| | *F* | 86.49 | 83.92 | 69.16 | 83.80 |
| ReliefF | *Acc* | 70.03 | 70.32 | 60.32 | 67.56 |
| | *F* | 68.18 | 74.08 | 66.82 | 76.36 |
| mRMR | *Acc* | 76.54 | 70.52 | 69.13 | 77.43 |
| | *F* | 77.81 | 78.12 | 71.65 | 81.10 |
| CMIM | *Acc* | 81.50 | 70.02 | 77.25 | 78.13 |
| | *F* | 83.39 | **80.20** | 76.20 | 80.64 |
| **Proposed** | *Acc* | **89.50** | **75.05** | **88.24** | **85.71** |
| | *F* | **89.63** | 76.14 | **86.69** | **83.15** |
| BC | | | | | |
| FCBF | *Acc* | 88.28 | 76.11 | 65.28 | 60.16 |
| | *F* | 92.49 | 76.92 | 60.46 | 49.86 |
| ReliefF | *Acc* | 88.18 | 70.32 | 56.68 | 48.66 |
| | *F* | 86.28 | 68.17 | 51.23 | 42.06 |
| mRMR | *Acc* | 89.09 | 85.49 | 89.09 | 76.68 |
| | *F* | 92.34 | 86.30 | 90.48 | 75.78 |
| CMIM | *Acc* | 89.09 | 86.36 | 89.09 | 70.08 |
| | *F* | 94.25 | 88.12 | 92.71 | 68.71 |
| **Proposed** | *Acc* | **96.39** | **90.09** | **94.14** | **88.18** |
| | *F* | **95.31** | **85.82** | **88.17** | **86.29** |
| CC | | | | | |
| FCBF | *Acc* | 82.29 | 63.11 | 55.28 | 49.78 |
| | *F* | 77.20 | 36.50 | 43.16 | 49.80 |
| ReliefF | *Acc* | 78.00 | 52.40 | 50.00 | 46.90 |
| | *F* | 69.55 | 7.99 | 46.30 | 51.90 |
| mRMR | *Acc* | 79.53 | 59.77 | **63.67** | **57.20** |
| | *F* | 73.17 | 29.29 | **53.86** | **56.24** |
| CMIM | *Acc* | 77.10 | **64.50** | 50.83 | 56.43 |
| | *F* | 70.35 | **38,70** | 41.68 | 55.56 |
| Proposed | *Acc* | **90.74** | 54.07 | 51.63 | 56.47 |
| | *F* | **92.13** | 13.29 | 44.78 | 52.20 |
| TC | | | | | |
| FCBF | *Acc* | 85.38 | 84.11 | 75.84 | 79.44 |
| | *F* | 86.59 | 86.20 | 68.76 | 77.84 |
| ReliefF | *Acc* | 83.78 | 82.89 | 67.56 | 78.89 |
| | *F* | 88.05 | 81.69 | 71.31 | 77.41 |
| mRMR | *Acc* | 83.33 | 76.22 | 76.22 | 78.44 |
| | *F* | 87.20 | 69.26 | 67.74 | 79.84 |
| CMIM | *Acc* | 86.44 | 85.56 | 86.00 | 79.78 |
| | *F* | 85.65 | 84.56 | 86.46 | 77.64 |
| **Proposed** | *Acc* | **90.37** | **89.56** | **90.05** | **88.76** |
| | *F* | **92.70** | **89.16** | **91.58** | **90.24** |

Note: $F$ value represents the ability of our proposed method to predict cases.

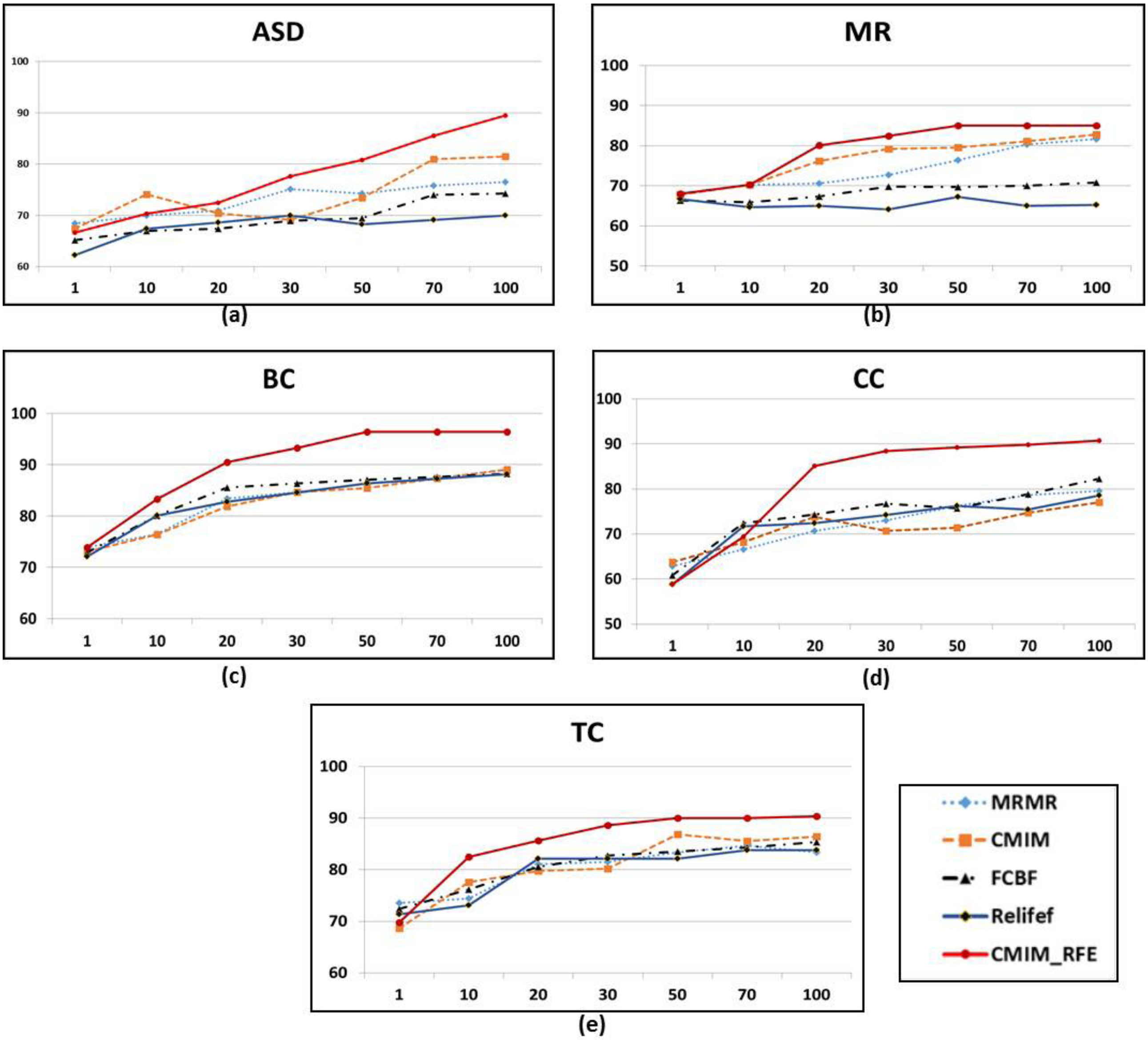


**FIGURE 4.** Performance comparison between four feature selection methods, namely ReliefF, MRMR, FCBF, CMIM and our proposed hybrid feature selection method CMIM-RVE. The results were obtained using SVM classifier over (1, 10, 20, 30, 50, 70 and 100) SNP, which was applied on five complex diseases datasets a) ASD; b) MR; c) BC; d) TC; e) CC, Where x and y axes indicate number of SNPs and accuracy respectively.

In Fig. 4 (b), (c) and (e) (MR, BC and TC respectively) the best *Acc* was reached with fewer than 50 SNPs. While the best *Acc* was achieved using 100 SNPs in (ASD, CC) in Fig.4 (a) and (d).

These observations demonstrate that the proposed method can select more informative features. Particularly, it can achieve up to 96 % prediction accuracy with 50 SNPs on BC dataset, which is very useful for medical diagnosis.

Most FS methods perform similarly when using less than 10 SNPs in the selection process, whereas, the accuracy varies after increasing the number of selected SNPs. The accuracy of our proposed method outperforms the other feature selection methods and the enhancements in accuracy is (4-10 %) more than the best algorithm. More precisely, the performance of CMIM and mRMR is often equivalent, and mRMR is slightly better in most cases. It is due to mutual information metrics that were used in both methods. On the other hand the performance of FCBF and ReliefF are almost similar, and FCBF is slightly better. It is due to coloration based metrics that were used in both methods.

The superiority of our proposed method is due to using two different metrics, mutual information and SVM recursive feature elimination, adopted to compute feature informativeness. Another interesting observation is that our algorithm performance keeps increasing simultaneously with the increase in the number of features, until it reaches the maximum accuracy, while the other methods present unstable increments.

One more interesting finding is that our proposed algorithm has the ability to perform better regardless of the dataset distribution, as presented in FIGURE 4 (c); the BC dataset has highly separated samples and all used FS methods achieve high performance including our proposed, While in FIGURE 4 (d) CC dataset has samples that are not very separated and all used FS methods perform poorly with no *Acc* more than 80%. In this dataset the proposed method outperforms the other methods and reaches up to 90% accuracy using 100 SNPs.

**TABLE 3. Comparison of classification accuracy for ASD dataset.**

| ASD | SNPs | Acc% | Ref |
|---|---|---|---|
| DS + SVM | 2 | 71 | [4] |
| Chi+ SVM | 98 | 64 | [4] |
| CBFS+ SVM | 10 | 64 | [5] |
| ReliefF+ SVM | 60 | 78 | [5] |
| RFS+ SVM | 10 | 64 | [5] |
| FSDD+ SVM | 100 | 64 | [5] |
| **Proposed(SVM)** | 100 | **89.50** | |

Table 3 shows the comparison between the proposed method and different frameworks that have been previously applied in the literature for the ASD dataset. The Chi+SVM and RFS+SVM algorithms had the worst performance, with their best accuracy reaching only 64% with approximately 100 SNPs. However, the proposed method reached the same accuracy with only one SNP. The ReliefF with SVM had the best accuracy 78%, obtained with 60 SNPs. We used a similar number of features to obtain the same accuracy. Moreover, the proposed system was able to reach an accuracy of 89.50% over 100 SNPs. The proposed technique outperformed all the other frameworks when applied on the ASD dataset.

For the MR dataset, the compared frameworks were able to classify the samples with high accuracy, as shown in Table 4. It illustrates the superiority of the proposed method over different frameworks. The Chi with SVM technique had the worst performance, where the best accuracy reached only 59% using two SNPs, the proposed method Accuracy was 68% using only one SNP. The CBFS with SVM achieved the highest accuracy up to 86% using 70 SNPs, while our method used 50 SNPs to reach similar accuracy. Our algorithm was able to reach that accuracy using less number of SNPs. Our proposed technique had competitive performance comparing to the other techniques for the MR dataset.

**TABLE 4. Comparison of classification accuracy for MR dataset.**

| MR | SNPs | Acc% | Ref |
|---|---|---|---|
| DS + SVM | 6 | 59 | [4] |
| Chi+ SVM | 2 | 59 | [4] |
| ReliefF+ SVM | 30 | 78 | [5] |
| RFS+ SVM | 10 | 73 | [5] |
| FSDD+ SVM | 20 | 79 | [5] |
| CBFS+ SVM | 70 | **86** | [5] |
| **Proposed(SVM)** | 50 | 85.00 | |

In the BC dataset, 94% accuracy was reached by CBFS and SVM using 60 SNPs. The proposed system provided better performance reaching an accuracy of 96.39% using 50 SNPs, as shown in Table 5. The proposed approach has the ability to compete with the other techniques in terms of the accuracy and selected number of features. Up to our knowledge there are no previous studies conducted on CC and TC datasets for feature selection and classification purpose.

**TABLE 5. Comparison of classification accuracy for BC dataset.**

| BC | SNPs | Acc% | Ref |
|---|---|---|---|
| ReliefF+ SVM | 10 | 57 | [5] |
| RFS+ SVM | 100 | 50 | [5] |
| FSDD+ SVM | 10 | 56 | [5] |
| CBFS+ SVM | 60 | 94 | [5] |
| **Proposed(SVM)** | 50 | **96.39** | |

The nonparametric *Kruskal-Wallis H* test [37] was used in order to examine the statistically significance of the results achieved using the proposed approaches compared to the other examined approaches for all the datasets used (ASD, MR, BC, TC, CC). Results presented in Table 6 demonstrate that the proposed method provides statistical significant results ( $p \leq 0.05$ ) for most of the examined scenarios. In this table, the presence of a letter denotes the statistical significance of the proposed approach (column) against an examined approach (row) for the respective dataset used. For example, the proposed method using the SVM classifier (column 1) provided statistically significant results for all datasets when compared to the ReliefF + SVM approach (row 1), but only for the ASD, CC and TC datasets when compared to the CMIM + Naive Bayes approach (row 10).

**TABLE 6. Statistical significance of the proposed scheme against the other examined schemes analysed by the non-parametric Kruskal-Wallis H test.**

| | | Proposed | | | |
|---|---|---|---|---|---|
| | | SVM | KNN | NB | LDA |
| SVM | FCBF | A B C T | A M C | A C T | A C T |
| | ReliefF | A M B C T | A M C | A M C T | A M C T |
| | CMIM | A C T | A M C | A C | A C |
| | mRMR | A B C T | M C | A C T | A C T |
| | Proposed | 1 | A M C T | C | A M B C |
| KNN | FCBF | A M B C T | A M | A M T | A M B |
| | ReliefF | A M B C T | A M B | A M B T | A M B T |
| | CMIM | A M B C T | A M C | A M B C T | A M |
| | mRMR | A M B C T | A T | A M B T | A T |
| | Proposed | A M C T | 1 | A M | A |
| NB | FCBF | A C T | A M C | A C | A C |
| | ReliefF | A M B C T | A M B T | A M B T | A M B T |
| | CMIM | A C T | A T | A T | A T |
| | mRMR | A B C T | A T | A C T | A T |
| | Proposed | C | A M | 1 | M |
| LDA | FCBF | A M B C T | A M C | A M B T | A B T |
| | ReliefF | A M B C T | A M B T | A M B T | A M B T |
| | CMIM | A M B C T | B T | A M B T | A M B T |
| | mRMR | A M B C T | B T | A M B T | A B T |
| | Proposed | A M B C | A | M | 1 |

(A: ASD, M: MR, B: BC, T: TC, C: CC) Each letter denotes a statistical significance with $p \leq 0.05$ for the respective dataset.

All the experiments were conducted using Matlab 2016a on a PC equiped with an Intel Core i7-5600

(4 cores@2.6 GHz) using a Windows 2010 64-b operating system. Table 7 presents the execution time in seconds for three different FS methods and our proposed, when extracting 100 SNPs as the optimal subset. Our proposed method has the best performance compared with ReliefF, mRMR and FCBF in term of complexity and increase in execution time over CMIM.

**TABLE 7. comparison of execution time in seconds for ReliefF, FCBF, mRMR, CMIM and the proposed method.**

| Dataset | ASD | MR | CC | TC | BC |
|---|---|---|---|---|---|
| ReliefF | 2660.03 | 1017.79 | 228.27 | 2001.47 | 601.70 |
| mRMR | 329.81 | 195.92 | 165.33 | 541.63 | 329.81 |
| FCBF | 226.18 | 145.47 | 113.97 | 305.03 | 198.51 |
| CMIM | 4.93 | 2.65 | 1.71 | 9.50 | 4.93 |
| Proposed | 115.97 | 91.84 | 60.77 | 70.31 | 68.36 |

## IV. CONCLUSION

The human genome sequence was a great achievement, and great progress in terms of diseases analysis was expected. We aimed in this work to participate in the understanding of complex diseases. The final goal was to build a framework able to analyze the SNP data and distinguish between healthy and affected samples. In this work we proposed a hybrid feature selection model to select the optimal subset of SNPs. We fused the CMIM filter method with the SVM-RFE wrapper method. The selected SNPs were injected into different classifiers to measure the performance of our proposed method. We conducted our experiment on five different datasets obtained from NCBI (GEO) TC, ASD, CC, MR and BC, and the proposed model outperformed the compared methods (mRMR, CMIM, FCBF, ReliefF). Accuracies up to 96 % were achieved in the tested datasets using different numbers of SNPs.

As presented previously, it is evident that the proposed technique provides noticeable improvements in the accuracy over all compared FS algorithms in the previous studies over three datasets [4], [5], for the BC dataset competitive results were obtained, with accuracy similar to the other frameworks. Nevertheless, the proposed method has the advantage of being more time efficient.

From the obtained results we conclude that SNPs of the whole genome can be efficiently employed to distinguish affected individuals with complex diseases from the healthy ones. A great deal of work remains to be done in order to understand the genetic basis of diseases and traits.

## REFERENCES

[1] M. Waddell, D. Page, and J. Shaughnessy, Jr., "Predicting cancer susceptibility from single-nucleotide polymorphism data: A case study in multiple myeloma," in *Proc. 5th Int. Workshop Bioinf.*, 2005, pp. 21–28.

[2] H. He, W. S. Oetting, M. J. Brott, and S. Basu, "Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study," *BMC Med. Genet.*, vol. 10, no. 1, p. 127, 2009.

[3] H. Schwender and K. Ickstadt, "Identification of SNP interactions using logic regression," *Biostatistics*, vol. 9, no. 1, pp. 187–198, 2008.

[4] D. T. Evans, "A SNP microarray analysis pipeline using machine learning techniques," M. S. thesis, School Elect. Eng. Comput. Sci., Russ College Eng. Technol.-Ohio Univ., Athens, OH, USA, 2010.

[5] N. Batnyam, A. Gantulga, and S. Oh, "An efficient classification for single nucleotide polymorphism (SNP) dataset," in *Computer and Information Science* (Studies in Computational Intelligence), vol. 493, R. Lee, Ed. Heidelberg, Germany: Springer, pp. 171–185.

[6] K. Anekboon, C. Lursinsap, S. Phimoltares, S. Fucharoen, and S. Tongsima, "Extracting predictive SNPs in Crohn's disease using a vacillating genetic algorithm and a neural classifier in case–control association studies," *Comput. Biol. Med.*, vol. 44, pp. 57–65, 2014, doi: 10.1016/j.compbiomed.2013.09.017.

[7] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[8] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci.*, vol. 282, pp. 111–135, Oct. 2014.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[10] C. Lazar *et al.*, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1106–1119, Jul. 2012.

[11] S. Uppu, A. Krishna, and R. Gopalan, "A review on methods for detecting SNP interactions in high-dimensional genomic data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2016.2635125.

[12] J. Lee, N. Batnyam, and S. Oh, "RFS: Efficient feature selection method based on *R*-value," *Comput. Biol. Med.*, vol. 43, no. 2, pp. 91–99, 2013.

[13] J. Liang, S. Yang, and A. Winstanley, "Invariant optimal feature selection: A distance discriminant and feature ranking based solution," *Pattern Recognit.*, vol. 41, no. 5, pp. 1429–1439, 2008.

[14] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, Oct. 2003.

[15] M. Seo and S. Oh, "CBFS: High performance feature selection algorithm based on feature clearness," *PLoS ONE*, vol. 7, no. 7, p. e40419, 2012.

[16] K.-A. Lê Cao, S. Boitard, and P. Besse, "Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems," *BMC Bioinf.*, vol. 12, no. 1, p. 253, 2011.

[17] K. Kim, M. Seo, H. Kang, S. Cho, H. Kim, and K.-S. Seo, "Application of logitboost classifier for traceability using SNP chip data," *PLoS ONE*, vol. 10, no. 10, p. e0139685, 2015.

[18] H.-Y. Yuan *et al.*, "FASTSNP: An always up-to-date and extendable service for SNP function analysis and prioritization," *Nucl. Acids Res.*, vol. 34, no. 2, pp. W635–W641, 2006.

[19] S. C. Shah and A. Kusiak, "Data mining and genetic algorithm based gene/SNP selection," *Artif. Intell. Med.*, vol. 31, no. 3, pp. 183–196, 2004.

[20] Z. Dawy, M. Sarkis, J. Hagenauer, and J. C. Mueller, "A novel gene mapping algorithm based on independent component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5. Mar. 2005, p. v-381.

[21] H.-W. Chang, L.-Y. Chuang, C.-H. Ho, P.-L. Chang, and C.-H. Yang, "Odds ratio-based genetic algorithms for generating SNP barcodes of genotypes to predict disease susceptibility," *OMICS A J. Integr. Biol.*, vol. 12, no. 1, pp. 71–81, 2008.

[22] T. Pahikkala, S. Okser, A. Airola, T. Salakoski, and T. Aittokallio, "Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations," *Algorithms Mol. Biol.*, vol. 7, no. 1, p. 11, 2012.

[23] Q. Wu, Y. Ye, Y. Liu, and M. K. Ng, "SNP selection and classification of genome-wide SNP data using stratified sampling random forests," *IEEE Trans. Nanobiosci.*, vol. 11, no. 3, pp. 216–227, Sep. 2012.

[24] F. François, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Nov. 2004.

[25] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[26] Z. Q. J. Lu, "The elements of statistical learning: Data mining, inference, and prediction," *J. Roy. Stat. Soc., A (Stat. Soc.)*, vol. 173, no. 3, pp. 693–694, 2010.

[27] X. Zhang *et al.*, "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC Bioinf.*, vol. 7, no. 1, p. 197, 2006.

[28] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sens. Actuators B, Chem.*, vol. 212, pp. 353–363, Jun. 2015.

[29] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 27–66, Jan. 2012.

[30] B. Luzón-Toro *et al.*, "Identification of epistatic interactions through genome-wide association studies in sporadic medullary and juvenile papillary thyroid carcinomas," *BMC Med. Genomics*, vol. 8, no. 1, p. 83, 2015.

[31] C. R. Marshall *et al.*, "Structural variation of chromosomes in autism spectrum disorder," *Amer. J. Hum. Genet.*, vol. 82, no. 2, pp. 477–488, 2008.

[32] F. Jasmine *et al.*, "A genome-wide study of cytogenetic changes in colorectal cancer using SNP microarrays: Opportunities for future personalized treatment," *PLoS ONE*, vol. 7, no. 2, p. e31968, 2012.

[33] D. J. McMullan *et al.*, "Molecular karyotyping of patients with unexplained mental retardation by SNP arrays: A multicenter study," *Hum. Mutation*, vol. 30, no. 7, pp. 1082–1092, 2009.

[34] M. Kadota *et al.*, "Identification of novel gene amplifications in breast cancer and coexistence of gene amplification with an activating mutation of PIK3CA," *Cancer Res.*, vol. 69, no. 18, pp. 7357–7365, 2009.

[35] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[36] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 856–863.

[37] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.

**RAID ALZUBI** received the bachelor's degree in computer science from Yarmouk University, Jordan, in 2003, and the master's degree in computer science from the Jordan University of Science and Technology, Jordan, in 2008. He is currently pursuing the Ph.D. degree in computer science with the University of the West of Scotland, U.K. His main research interests are bioinformatics, machine learning, and pattern recognition.

**NAEEM RAMZAN** (S'04–M'08–SM'13) received the M.Sc. degree in telecommunications from the University of Brest, France, in 2004, and the Ph.D. degree in electronics engineering from the Queen Mary University of London, London, U.K, in 2008. He is currently a Full Professor with the School of Engineering and Computing, University of the West of Scotland. He has authored or co-authored over 150 research publications, including journals, book chapters, and standardization contributions. He co-edited a book entitled *Social Media Retrieval* (Springer, 2013). He is a fellow of the Higher Education Academy. He has organized and co-chaired three ACM Multimedia Workshops. He served as the Session Chair/Co-Chair for a number of conferences. He is the Co-Chair of the Ultra HD Group of the Video Quality Experts Group (VQEG). He served as a Guest Editor for a number of special issues in technical journals. He is a Co-Editor-in-Chief of *VQEG E-Letter*.

**HADEEL ALZOUBI** received the bachelor's degree in computer science from the Jordan University of Science and Technology, Jordan, in 2008. She is currently pursuing the Ph.D. degree in computer science with the University of the West of Scotland, U.K. Her main research interests are natural language processing, machine learning, and pattern recognition.

**ABBES AMIRA** (S'99–M'01–SM'07) received the Ph.D. degree in computer engineering from Queen's University Belfast, Belfast, U.K., in 2001. He held many academic and consultancy positions in the U.K., Middle East, and Asia, including his recent positions as a Professor of computer engineering at Qatar University and a Professor of visual communications at the University of the West of Scotland, U.K. His research interests include reconfigurable computing, signal processing, and connected health. He is a fellow of the IET and a Senior Member of the ACM.