Original Article

# An Optimized Framework for Cancer Prediction Using Immunosignature


## Abstract

**Background:** Cancer is a complex disease which can engages the immune system of the patient. In this regard, determination of distinct immunosignatures for various cancers has received increasing interest recently. However, prediction accuracy and reproducibility of the computational methods are limited. In this article, we introduce a robust method for predicting eight types of cancers including astrocytoma, breast cancer, multiple myeloma, lung cancer, oligodendroglia, ovarian cancer, advanced pancreatic cancer, and Ewing sarcoma. **Methods:** In the proposed scheme, at first, the database is normalized with a dictionary of normalization methods that are combined with particle swarm optimization (PSO) for selecting the best normalization method for each feature. Then, statistical feature selection methods are used to separate discriminative features and they were further improved by PSO with appropriate weights as the inputs of the classification system. Finally, the support vector machines, decision tree, and multilayer perceptron neural network were used as classifiers. **Results:** The performance of the hybrid predictor was assessed using the holdout method. According to this method, the minimum sensitivity, specificity, precision, and accuracy of the proposed algorithm were 92.4 ± 1.1, 99.1 ± 1.1, 90.6 ± 2.1, and 98.3 ± 1.0, respectively, among the three types of classification that are used in our algorithm. **Conclusion:** The proposed algorithm considers all the circumstances and works with each feature in its special way. Thus, the proposed algorithm can be used as a promising framework for cancer prediction with immunosignature.


**Keywords:** *Cancer, feature selection, immunosignature, normalization*


**Fatemeh Safaei Firouzabadi[1], Alireza Vard[2], Mohammadreza Sehhati[2], Mohammadreza Mohebian[3]**

*[1]Student Research Committee, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, [2]Department of Bioelectrics and Biomedical Engineering, School of Advanced Technologies in Medicine and Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, [3]Department of Biomedical Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran*

A. Vard

ORCID ID

https://orcid.org/0000-0003-1768-910X


## Introduction

When antibodies circulate in the blood, they can connect to a large microarray of randomized sequence peptides.[1] An "immunosignature" is a pattern of random sequence peptides, which is obtained with a blood sample test.[2] Neoantigens are produced by cancer release native proteins and biomolecules that are not encountered by the immune system. Therefore, the change in the regulation of the gene expression and proteins in cells can be considered as a sign of cancer.[3] However, there is a slight overlap of the signatures among the cancers that resulted in a loss of specificity in distinguishing between cancers using immunosignatures. To resolve this, peptides were determined to be statistically significant in the cancer signatures using more stringent selection processes.

In the recent years, various methods have been introduced in the literature for predicting cancer with peptides and proteomic datasets. Zhang *et al*.[4] for classifying ten types of cancers used the protein expression profiles. They used the minimum redundancy maximum relevance and the incremental feature selection methods in order to select 23 out of 187 proteins as the inputs of the sequential minimal optimization classifier. Using 23 proteins, they classified with Matthews Correlation Coefficient (MCC) of 0.936 on an independent test set. Kaddi and Wang[5] used the proteomic and transcriptomic data to predict the early stage of cancer in head-and-neck squamous cell carcinoma. They proposed a filter and wrapper method for feature selection and employed the individual binary classification accompaniment with the ensemble classification methods. Stafford *et al*.[2] used ANOVA and *t*-test for approximately 10,000 peptide sequences. They proposed a novel feature selection method and the naive Bayes, linear discriminant analysis, and support vector machine (SVM) for classification in two libraries. An average accuracy of 98% and an average sensitivity of 89% were reported.[6]



**How to cite this article:** Firouzabadi FS, Vard A, Sehhati M, Mohebian M. An optimized framework for cancer prediction using immunosignature. J Med Sign Sens 2018;8:161-9.

***Address for correspondence:***
*Dr. Alireza Vard, Department of Biomedical Engineering, School of Advanced Technologies in Medicine and Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, Iran.*
*E-mail: vard@amt.mui.ac.ir, alivard@gmail.com*

**Website:** www.jmss.mui.ac.ir
**DOI:** 10.4103/jmss.JMSS_2_18

Nguyen *et al*.[7] used five filter-based feature selection methods including *t*-test, Wilcoxon, entropy, signal-to-noise ratio, and receiver operating characteristic curve. Moreover, they employed an analytic hierarchy process which is a multi-criteria decision analysis based on the type-2 fuzzy method for the classification of cancer microarray gene expression profiles. Based on the proposed classification method, the achieved accuracy was 95.24%, considering *t*-test as the feature selection method.

In the proposed framework, we combined the meta-heuristic population-based optimization with feature selection and normalization methods to improve the accuracy and efficiency of the classification algorithms for classifying 12 different cancer types. Briefly, the particle swarm optimization (PSO) method at first filters the significant peptides. Next, it selects the best method of normalization from the dictionary for one feature and chooses weights for features that are selected with the statistical feature selection process. Then, the selected features apply to classification. In our study, three types of classification, including SVM, multilayer perceptron (MLP), and decision tree (DT), are used and compared with each other.

The rest of the article is organized as follows: in the next section, information about the datasets and proposed methods which are used in this study are presented. The results of the proposed method and the discussion about them are provided in the Results and Discussion Sections and finally, the article is concluded in the Conclusion Section.

## Materials and Methods

### The dataset

In this study, we used a public immunosignature peptide microarray dataset (arrays of 10,371 peptides) which consists of 1516 patients for 12 different cancer types, 2 infectious diseases, and healthy controls.[7] The Gene Expression Omnibus series code of this dataset is GSE52581, which is publicly available, and other types of this kind of dataset are used in recent studies carried out on cancer.[2,8,9] The features of this dataset are not normalized and the dataset includes cancers such as astrocytoma, breast cancer, multiple myeloma, lung cancer, oligodendroglia, ovarian cancer, advanced pancreatic cancer, and Ewing sarcoma. We removed the data related to the infectious diseases and used 1292 subjects. Thus, the dataset consists of 1292 columns and 10,371 rows and, in this article, rows refer to mean peptides and columns refer to samples. More details about the used dataset are described in Table 1.

**Table 1: Basic information of patients per cancer in the utilized dataset**

| Cancer type | Number of patients (%) |
|---|---|
| Recurrence breast cancer | 61 (4.7) |
| Breast cancer stages II, III | 141 (10.9) |
| Breast cancer stage IV | 42 (3.2) |
| Aggressive-type astrocytoma | 27 (2.0) |
| Astrocytoma | 166 (12.8) |
| Lung cancer | 107 (8.2) |
| Multiple myeloma | 112 (8.6) |
| Oligodendroglia | 48 (3.7) |
| Oligoastrocytoma | 97 (7.5) |
| Ovarian cancer | 86 (6.6) |
| Pancreatic cancer | 136 (10.5) |
| Ewing sarcoma | 20 (1.5) |
| Healthy control | 249 (19.2) |

### The proposed algorithm

The structure of the proposed algorithm is depicted in Figure 1. As shown in this figure, at first, features are normalized with different methods that are selected by PSO from a dictionary of methods. Then, a statistical feature selection method is used to identify significant discriminative features. The selected features are assigned weights by PSO. The goodness of fit of PSO can be measured by F1-score of classifier on the training set (Eq. 1).

$$\text{Goodness of fit} = \text{FScore}_{\text{train}} \quad (1)$$

After selecting suitable features, classification methods are used to classify the types of cancers. In this study, we utilized three multi-class classification methods including multi-class SVM, DT, and multilayer perceptron. The weights of features are estimated using PSO during learning of classifiers. The algorithm stops if no remarkable improvement is seen in the objective function or the maximum number of iterations (set to 50 in our study) is reached.

In the following section, the methods and algorithms, employed in different parts of the proposed scheme, are described in detail.

#### *Normalization*

According to Liu *et al*.,[10] a major bioinformatics challenge is the normalization of the data. Normalization is a method that puts data in a similar domain when they are not in one domain. In other words, a data miner may encounter situations in which features in data include quantities in different domains. These features with large quantities may have higher impacts on cost function compared to features with low quantities. This issue is resolved by normalization of features such that their values are put into one domain.[11] Thus, if each feature normalizes properly, classification would be applied more effectively in feature space compared to using classifying without normalization.[12] In addition, it changes the characteristics of the underlying probability distributions.[13] A careful analysis of the geometry of feature space suggests a modification on normalization procedure that works on

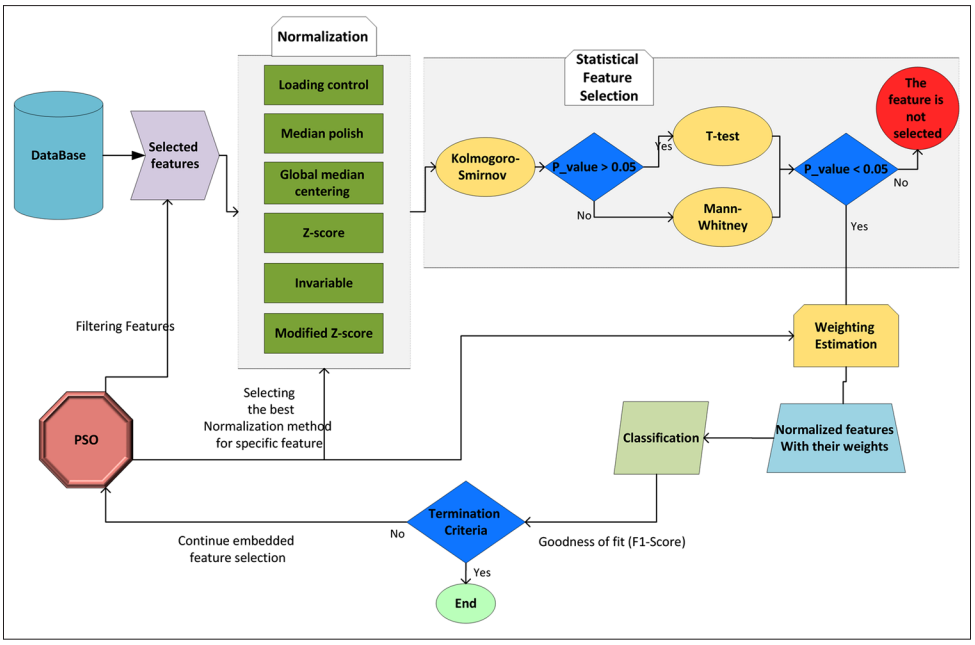


**Figure 1: The structure of the proposed algorithm**

each feature separately.[14] Therefore, in this study, a dictionary of normalization methods is used which gives the algorithm the ability to select the best normalization method for each feature irrelevant from other features. These normalization methods are selected from the previous studies used on peptides, proteomics, and other microarray datasets.[12,15-17] These methods are described in the following subsection.

### Global median centering

In this method, the median value of peptides is simply subtracted from each peptide. While this method needs only one sample and does not require other data column, it has bias when the number of peptides is low (<100).[12]

### Tukey (median polish)

This method is a nonparametric analysis of variance.[12] The column medians and row medians are subtracted until a defined criterion is satisfied. This criterion could be the number of iteration or a specific value of the proportional reduction in the sum of absolute residual.[18] If the number of samples and peptides is reasonably large, the approach will be robust.[11] In this study, the termination criterion of the algorithm is empirically defined as 400 iterations.

### Loading control method

This method uses the global median-centering approach at the first step for each peptide. Then, global median-centering approach is used for each sample rather than each peptide.[12]

### Robust Z-score

This method is robust like standard score and is formulated with Eq. 2.[19]

$$X'_{ij} = \frac{X_{ij} - Median_i}{MAD_i} \quad (2)$$

Where $X'_{ij}$ represents the $i^{th}$ peptide and the $j^{th}$ sample. In addition, $\text{MAD}_i$ and $\text{Median}_i$ denote median and median absolute deviation, respectively. According to the formulation, this method is not sensitive to outliers and therefore is appropriate for microarray data.[20]

### Invariable method

This method is well known for microarray data analysis in the literature[12,21,22] since it can eliminate systematic variation in microarray data. In this method, peptides are ranked, and then a peptide with the highest rank is discarded. This process is repeated until the remaining number of peptides reach a predetermined value (1000 in this article). Then, 25% of the highest and the lowest ranked data are removed and the average of each peptide creates a virtual reference. At the end, each sample is normalized to the virtual reference by the MA-plot approach.[12] In the MA-plot approach, the difference between each sample and the reference sample in the logarithm of base 2 is plotted against the mean of each sample and the reference sample. Then, the normalized values are generated using residuals of fit consequently.

### Modified Z-score method

In this method, $\log_2$ is performed on the whole data, firstly. Then, the standard score normalization is applied to each sample and each peptide subsequently. In the final step, the arctangent function is applied to the data.[23,24] This method has had a great performance on gene expression data comparatively.[23-26]

### *Statistical feature selection*

Feature selection approaches can be classified into three categories including wrapper, filter, and embedded methods.[27] The filter method is used in the preprocessing section and works independently from classifier.[28] On the other hand, other two methods are used during classification. The wrapper method evaluates the combination of features by formulating a problem and searches the problem space for the best features.[29] This method tests the entire feature subsets.[30] Finally, the embedded method evaluates the accuracy of the classifier for predicting the best features with searching that is guided by a learning classification process. This characteristic of the embedded feature selection method makes it robust against overfitting.[27]

According to the characteristics of the peptides' datasets that have only interval data types, the Kolmogorov–Smirnov (KS) test was appropriate.[31] The KS evaluates the maximum absolute difference in the overall distribution of the two groups (cancer or noncancer). Then, if a feature was significant, the independent-sample *t*-test was used to identify statistically discriminative normally distributed features.[32] Otherwise, the Mann–Whitney test is performed to check whether two independent samples are significantly different or not.[33] Finally, if numbers of selected features are >50, the algorithm selects the 50 highest ranks in *t*-test and then in Mann–Whitney test. Thus, the 50 discriminative features are selected consequently and prepared as the inputs of the classification system which is combined with PSO for weighing them as discussed in the following subsection.

### *Particle swarm optimization*

PSO is a meta-heuristic stochastic evolutionary population-based optimization algorithm that is inspired by birds.[34] PSO can solve a range of difficult optimization problems, but it has shown a faster convergence rate compared to other evolutionary algorithms.[35,36]

This algorithm is used in many different computational biology fields such as modeling in biology,[37] feature selection in gene expression data,[38,39] DNA sequence encoding,[40] and breast cancer recurrence prediction.[6] The PSO algorithm starts with random solutions that are called particle positions. Each particle has a velocity and a position that help it to search the whole problem space. The position of a particle is updated according to three parameters: its previous speed, the best position visited by particle, and the best position of the neighborhood (Eq. 4).

$$v_i^{n+1} = wv_i^n + c_1 r_1^n \left(p_i^n - x_i^n\right) + c_2 r_2^n \left(p_g^n - x_i^n\right) \quad (3)$$

$$x_i^{k+1} = x_i^k + v_i^{n+1} \quad (4)$$

where $n$ is the iteration number; $C_1$ and $C_2$ are the learning factor coefficients that usually set to 2; $i$ is the particle number, $r_1$, and $r_2$ are random numbers that are uniformly distributed in (0, 1); $w$ is the inertia weight, where a large number of it shows exploration while a small number of it denotes exploitation[41] Thus, the inertia coefficient was set to 1.00 at the first iteration and was linearly decreased with the damping coefficient of 0.99 at each iteration. Furthermore, the objective function is maximizing the goodness of fit that is mentioned in Eq. 1.

In our study, PSO algorithm has three main tasks. First, it selects probable peptides. For this purpose, 400 particles are considered in which each particle represents the index of one peptide and it can be an integer number between 1 and 10,371. Second, the PSO algorithm chooses the normalization method for each feature that was selected in the previous step. In these steps, the rounded value of the particle position is selected as an index of the normalization method. It is necessary to say, only one peptide which is normalized individually or with accompaniment of other peptides goes to the next step; for example, if the loading control method is selected for one peptide, the peptide will be normalized by this method with accompaniment of other peptides (because other peptides are used in normalization formulation). However, other peptides will be normalized with their own selected method and then move to the next step. Third, the selected features of the statistical feature selection method are weighed by the PSO algorithm that weighs a real number between 0 and 1.

In a nutshell, 400 particles for initial filtering of features, 400 particles in selecting normalization method, and 50 particles for weighing features were used. In total, 850 variables should be considered by PSO, which is an appropriate algorithm for solving high-dimensional problems.[42,43] After features are selected by statistical feature selection and weighed by PSO, they are used as inputs of the classifiers.

### *Classification*

In the proposed method, three classification methods such as MLP, DT, and SVM are used. These methods were used in previous similar studies.[6,44-46]

The SVM considers a set of the hyperplane in a high-dimensional space.[47] This algorithm has been widely utilized in peptides' datasets and other relevant fields.[6,48,49] The radial basis function (RBF) kernels were used in this study and it is paramount importance that the soft-margin parameter and the radius of the RBF kernel should be set appropriately since they do not cause poor classification results. We used the method proposed by Wu and Wang[50] for this purpose.

MLP is an improved version of the standard linear perceptron method and can be used for classification of nonlinearly separable data.[51] It is a famous machine learning approach in a variety of the computational biology field such as prediction of protein stability, prognosis DNA methylation biomarkers in ovarian cancer, and encoding aminoacids.[52-54] In this study, the MLP with one hidden layer, 20 neurons, and the sigmoid activation function are used because they seem suitable for prediction of cancer,

according to the literature.[55-57] The number of neurons is calculated empirically until increasing the number of neurons up has no effect on performance.

The DT classifier creates a structure of tree for modeling. We used C4.5 DT classifier in the proposed model, which is an entropy-based algorithm that can handles continuous attributes.[58,59] Furthermore, this method is widely proposed in predicting cancer, predicting specific target of peptides, and analyzing microarray dataset.[60-63]

### *Validation procedure*

The performance of the proposed algorithm was evaluated by the hold-out method in terms of sensitivity, specificity, precision, accuracy, F1-score, and MCC, which are defined in Table 2.

The sensitivity and specificity are highly dependent on the prevalence of the diseases. On the other hand, a reliable diagnostic system has sensitivity and specificity more than 80% (the minimum statistical power of 80%) and 95% (the maximum Type I error of 0.05), respectively.[64,65] Thus, a conservative method should satisfy both parameters.

## Results and Discussion

Leave-one-out cross-validation often underestimates error and leads to overfitting.[2] Therefore, in this study, we used hold-out validation and the results of the proposed method on the test set are shown in Tables 3-5. In this study, the hold-out method is used with 70% data as a training set and 30% data as a test set.[6] The 70% and 30% of data were chosen randomly for training and test sets in each time of running the whole algorithm. The complete procedure of data processing has been done 7 times and the results that are shown in Tables 3-5 are mean of 7 times of running. These results indicate that the procedure with the proposed methods is robust. The final classification system is a multiclass system in which its parameters are calculated based on systematic analysis of multiclass classification approach[66] which are shown in Tables 3-5.

The F-score on the training set and the test set against PSO iteration is depicted in Figure 2.

The average time measured for validating 454 patients were 0.33 ± 0.07, 0.30 ± 0.05, and 0.34 ± 0.08 second, respectively, for SVM, DT, and MLP. All results were obtained on a computer of Intel Core-i7, 2 GHz CPU with 8 GB of RAM.

Due to data imbalance, it seems that overall accuracy is not a suitable fitness measure.[67] As a matter of fact, selecting inappropriate objective function instead of Eq. 1 creates bias toward majority prevalence. The average sensitivity and specificity of the algorithm based on SVM classifier are estimated as 99.8 ± 0.7 and 99.9 ± 0.6 [Table 4], respectively, as the best classifier in this approach. Furthermore, the maximum error type I ($\alpha$) and II ($\beta$) are 0.009 and 0.076, subsequently. Thus, in the proposed method, the average of type I and II errors showed consistency in the results of the algorithm and this method has the capability to be used in clinical applications.

The proposed method could be called a general framework since it uses a dictionary of normalization's method and optimized feature selection. As a comparison with others' works which just use *t*-test or Wilcoxon feature selection with a fixed type normalization method, this method could search more space of the solutions and present a comprehensive answer. In other words, methods which only use one normalization method and feature selection procedure are one of the solutions in the search space of the proposed framework. As the purpose of illustration, Stafford *et al*.[2] worked on the same dataset and used *t*-test method as feature selection and global median centering as normalization that

**Table 2: Validation parameters**

| Parameters | Definition |
|---|---|
| Accuracy | $\frac{TP + TN}{TP + TN + FP + FN}$ |
| Sensitivity | $\frac{TP}{TP + FN}$ |
| Specificity | $\frac{TN}{TN + FP}$ |
| Precision | $\frac{TP}{TP + FP}$ |
| F1-score | $\frac{2 \times Pr \times Se}{Pr + Se}$ |
| MCC | $\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$ |

MCC – Matthews correlation coefficient

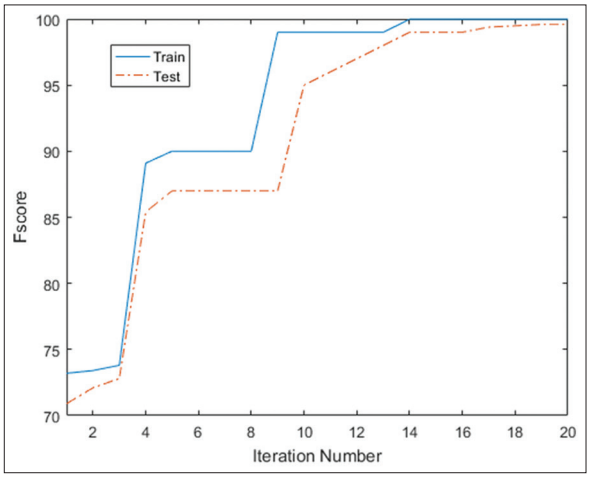


**Figure 2: The value of the fitness function. F-score on the training set is the solid line and the F-score on the test set is the dash-dot line during optimization procedure. The termination criterion was only the maximum number of iterations (i.e., 20) in this plot**

**Table 3: The average of five holdout performance estimates for multilayer perceptron method**

| Proposed method: MLP | | | | | | |
|---|---|---|---|---|---|---|
| **Cancer type** | **Sensitivity (%)** | **Specificity (%)** | **Accuracy (%)** | **Precision (%)** | ***F* (%)** | **MCC (%)** |
| Recurrence breast cancer | 98.1±1.3 | 97.9±0.5 | 98.9±0.4 | 86.3±1.6 | 93.0±1.1 | 91.8±1.2 |
| Astrocytoma | 99.4±0.8 | 99.9±0.3 | 99.9±0.1 | 99.5±0.6 | 98.3±1.4 | 99.3±0.9 |
| Breast cancer stages II, III | 96.5±2.1 | 99.9±0.6 | 96.5±1.2 | 74.4±1.4 | 84.5±1.3 | 84.3±1.5 |
| Breast cancer stage IV | 99.0±1.1 | 99.5±0.8 | 99.9±0.4 | 99.0±1.5 | 99.2±1.1 | 98.9±0.9 |
| Ovarian cancer | 54.8±4.2 | 99.9±2.2 | 96.3±2.3 | 99.8±1.0 | 70.4±0.5 | 71.5±1.9 |
| Lung cancer | 99.0±1.0 | 99.7±1.0 | 99.7±1.0 | 98.3±2.3 | 99.0±1.4 | 98.7±2.1 |
| Multiple myeloma | 99.1±2.0 | 99.6±0.7 | 99.9±0.1 | 93.6±1.1 | 97.5±0.1 | 99.2±1.8 |
| Aggressive-type astrocytoma | 98.5±0.4 | 99.8±0.5 | 99.9±0.1 | 99.0±0.4 | 98.1±0.7 | 97.3±0.4 |
| Oligodendroglia | 96.6±0.3 | 99.9±0.7 | 99.3±0.4 | 95.1±0.9 | 97.3±0.8 | 98.0±0.7 |
| Oligoastrocytoma | 99.9±0.1 | 99.8±0.2 | 98.2±1.3 | 91.6±0.9 | 96.4±0.6 | 95.6±1.5 |
| Pancreatic cancer | 99.9±0.4 | 97.7±0.6 | 98.5±1.0 | 86.4±1.6 | 93.3±1.8 | 92.4±1.1 |
| Ewing sarcoma | 66.9±2.4 | 99.3±0.4 | 99.5±0.8 | 73.4±1.5 | 71.9±2.2 | 70.2±1.9 |
| Overall | 92.4±1.1 | 99.5±0.8 | 98.3±1.0 | 91.2±0.9 | 91.4±1.1 | 91.5±1.5 |

MCC – Matthews correlation coefficient; MLP – Multilayer perceptron

**Table 4: The average of five holdout performance estimates for support vector machine method**

| Proposed method: SVM | | | | | | |
|---|---|---|---|---|---|---|
| **Cancer type** | **Sensitivity (%)** | **Specificity (%)** | **Accuracy (%)** | **Precision (%)** | ***F* (%)** | **MCC (%)** |
| Recurrence breast cancer | 99.1±1.1 | 99.9±0.1 | 99.9±0.1 | 97.3±1.5 | 99.9±0.1 | 99.5±0.6 |
| Astrocytoma | 99.3±0.4 | 99.2±0.6 | 99.0±0.3 | 98.2±0.4 | 99.6±0.2 | 99.5±0.5 |
| Breast cancer stages II, III | 99.1±0.4 | 99.9±0.2 | 99.8±0.3 | 99.0±0.6 | 99.1±0.3 | 99.7±0.4 |
| Breast cancer stage IV | 99.8±1.0 | 99.2±0.9 | 99.9±0.3 | 99.3±1.3 | 99.4±1.0 | 99.4±0.7 |
| Ovarian cancer | 99.1±0.2 | 99.9±0.4 | 99.9±0.3 | 96.9±1.1 | 98.1±1.6 | 99.4±0.6 |
| Lung cancer | 99.9±0.2 | 99.4±0.3 | 99.9±0.1 | 99.2±0.6 | 99.4±0.4 | 99.4±0.5 |
| Multiple myeloma | 99.1±1.0 | 99.9±0.3 | 99.9±0.2 | 99.9±0.7 | 99.1±1.1 | 99.6±0.4 |
| Aggressive-type astrocytoma | 99.2±0.3 | 99.9±0.5 | 99.9±0.1 | 99.9±0.8 | 99.0±0.2 | 99.7±0.6 |
| Oligodendroglia | 99.0±1.3 | 99.6±0.4 | 99.1±0.4 | 99.0±1.1 | 98.3±0.7 | 99.2±0.5 |
| Oligoastrocytoma | 98.9±1.0 | 99.9±0.2 | 99.9±0.2 | 98.1±0.9 | 99.0±0.8 | 99.5±0.6 |
| Pancreatic cancer | 98.6±0.9 | 99.9±0.1 | 99.9±0.4 | 98.4±0.3 | 98.7±0.4 | 99.3±0.7 |
| Ewing sarcoma | 99.8±1.2 | 97.0±2.4 | 98±1.1 | 86.7±3.1 | 93.3±1.2 | 99.4±0.5 |
| Overall | 99.8±0.7 | 99.9±0.6 | 99.9±0.5 | 97.6±1.6 | 98.9±1.6 | 99.5±0.8 |

MCC – Matthews correlation coefficient; SVM – Support vector machine

**Table 5: The average of five holdout performance estimates for decision tree method**

| Proposed method: DT | | | | | | |
|---|---|---|---|---|---|---|
| **Cancer type** | **Sensitivity (%)** | **Specificity (%)** | **Accuracy (%)** | **Precision (%)** | ***F* (%)** | **MCC (%)** |
| Recurrence breast cancer | 81.5±1.5 | 99.1±0.9 | 97.0±1.1 | 87.5±2.1 | 84.3±1.9 | 81.7±2.6 |
| Astrocytoma | 99.1±1.4 | 99.0±1.1 | 99.9±0.4 | 99.2±1.2 | 98.2±1.3 | 99.1±0.7 |
| Breast cancer stages II, III | 99.9±0.8 | 99.2±1.3 | 99.3±0.4 | 71.3±2.6 | 83.5±1.4 | 85.7±2.4 |
| Breast cancer stage IV | 99.0±1.1 | 99.9±0.2 | 99.3±0.6 | 99.1±0.6 | 99.2±1.1 | 99.2±0.9 |
| Ovarian cancer | 98.9±0.6 | 98.5±0.8 | 98.5±1.0 | 89.2±1.5 | 93.6±0.8 | 94.1±2.3 |
| Lung cancer | 99.3±0.8 | 99.9±0.4 | 99.9±0.2 | 99.3±0.8 | 98.6±1.0 | 98.9±0.9 |
| Multiple myeloma | 99.2±0.6 | 99.9±0.2 | 99.9±0.1 | 99.1±0.9 | 98.4±1.2 | 98.8±0.7 |
| Aggressive-type astrocytoma | 98.7±1.0 | 99.2±0.6 | 99.5±0.6 | 91.6.±2.0 | 95.3±1.3 | 95.1±0.8 |
| Oligodendroglia | 99.8±0.5 | 97.4±3.0 | 98.5±1.1 | 86.9±3.0 | 93.7±3.0 | 91.3±2.4 |
| Oligoastrocytoma | 99.9±0.9 | 97.8±3.2 | 98.5±1.0 | 88.3±1.5 | 93.5±1.8 | 92.1±2.0 |
| Pancreatic cancer | 99.2±1.3 | 99.9±0.1 | 99.9±0.1 | 99.3±0.9 | 98.1±0.9 | 99.1±1.0 |
| Ewing sarcoma | 67.4±1.5 | 99.4±1.3 | 98.2±1.9 | 75.9±3.1 | 71.3±2.4 | 72.0±1.5 |
| Overall | 95.4±1.5 | 99.1±1.1 | 99.3±1.4 | 90.6±2.1 | 92.2±1.3 | 92.5±1.1 |

MCC – Matthews correlation coefficient; DT – Decision tree

gained an average accuracy of 98% and an average sensitivity of 89%, while in the proposed method, the average accuracy was 99.16% and the average sensitivity was 95.87% which revealed the advantages of the proposed method.

## Conclusion

This article provides a novel method for predicting cancer with immunosignature. In the proposed method, the PSO algorithm was first used to filter some features. The selected features were refined by the statistical feature selection methods and estimated weights by PSO. The overall feature selection process is performed as part of a learning procedure. Instead of PSO, other meta-heuristic population-based stochastic optimization methods that could deal with discrete (feature and normalization selection) and continuous (feature weights) problems could be used. The performance of the algorithm was dependent on the PSO initialization and classifier tuning. As an instance, choosing proper numbers of hidden layers and neurons in each layer in MLP, kernel's parameters in SVM, and search space for initialing weight on feature selection by PSO procedure could have the paramount effect on results.

In the proposed method, the normalization dictionary is independently used for each feature because each normalization method was tested on the statistical selected features; however, we cannot confidently select an appropriate normalization method to reach high performance; therefore, an optimization framework was designed.

In a nutshell, the modified Z-score normalization method is selected more than other methods by PSO optimization. To shed light on it, it seems that modified Z-score normalization could map features to the suitable new space, in which discrimination between classes would be made more effective than other subspaces.

In the proposed study, different algorithms were analyzed on immunosignature data to give a clear insight into the classification and identification of biological markers for the diagnosis of diseases. It can help to adopt useful approaches to early diagnoses and treatments. More specifically, this study proposes a comprehensive algorithm by presenting different methods of normalization and feature selection using PSO which can help to attain optimal results.

The proposed algorithm is promising and can be utilized as a new offline tool in clinical applications. The developed program is available to interested readers upon request.

One of the limitations of the current study is that results might have been biased because scant of enough available samples. The sample size must be increased to improve the statistical power in our diagnosis system.[68] Another limitation of the proposed algorithm is that output of the classification system is not fuzzy.[69] It will be useful to report the risk of having a cancer type, which is the focus of our future activity.

### Financial support and sponsorship

Nil.

### Conflicts of interest

There are no conflicts of interest.

## References

1. Angenendt P. Progress in protein and antibody microarray technology. Drug Discov Today 2005;10:503-11.
2. Stafford P, Cichacz Z, Woodbury NW, Johnston SA. Immunosignature system for diagnosis of cancer. Proc Natl Acad Sci U S A 2014;111:E3072-80.
3. Otto T, Sicinski P. Cell cycle proteins as promising targets in cancer therapy. Nat Rev Cancer 2017;17:93-115.
4. Zhang PW, Chen L, Huang T, Zhang N, Kong XY, Cai YD, *et al.* Classifying ten types of major cancers based on reverse phase protein array profiles. PLoS One 2015;10:e0123147.
5. Kaddi CD, Wang MD. Models for predicting stage in head and neck squamous cell carcinoma using proteomic and transcriptomic data. IEEE J Biomed Health Inform 2017;21:246-53.
6. Mohebian MR, Marateb HR, Mansourian M, Mañanas MA, Mokarian F. A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. Comput Struct Biotechnol J 2017;15:75-85.
7. Nguyen T, Nahavandi S. Modified AHP for gene selection and cancer classification using type-2 fuzzy logic. IEEE Trans Fuzzy Syst 2016;24:273-87.
8. Figueiredo A, Monteiro F, Sebastiana M. Subtilisin-like proteases in plant-pathogen recognition and immune priming: A perspective. Front Plant Sci 2014;5:739.
9. Xu H, Tian Y, Yuan X, Liu Y, Wu H, Liu Q, *et al.* Enrichment of CD44 in basal-type breast cancer correlates with EMT, cancer stem cell gene profile, and prognosis. Onco Targets Ther 2016;9:431-44.
10. Liu W, Ju Z, Lu Y, Mills GB, Akbani R. A comprehensive comparison of normalization methods for loading control and variance stabilization of reverse-phase protein array data. Cancer Inform 2014;13:109-17.
11. Giorgi FM, Bolger AM, Lohse M, Usadel B. Algorithm-driven artifacts in median polish summarization of microarray data. BMC Bioinformatics 2010;11:553.
12. Graf AA, Smola AJ, Borer S. Classification in a normalized feature space using support vector machines. IEEE Trans Neural Netw 2003;14:597-605.
13. Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughead JW, *et al.* Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. Neuroimage 2005;28:663-8.
14. Xing EP, Karp RM. CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. Bioinformatics 2001;17 Suppl 1:S306-15.
15. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, *et al.* Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30:e15.
16. Rudnick PA, Wang X, Yan X, Sedransk N, Stein SE. Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data. Mol Cell Proteomics 2014;13:1341-51.
17. Scholma J, Fuhler GM, Joore J, Hulsman M, Schivo S, List AF, *et al.* Improved intra-array and interarray normalization of peptide microarray phosphorylation for phosphorylome and kinome profiling by rational selection of relevant spots. Sci Rep 2016;6:26695.
18. Bolstad BM. Comparing the effects of background, normalization and summarization on gene expression estimates. 2002. Available from: http://stat-www.berkeley.edu/users/bolstad/.
19. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, *et al.* Statistical methods for analysis of high-throughput RNA interference screens. Nat Methods 2009;6:569-75.

20. Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. Pattern Recognit 2005;38:2270-85.
21. Pelz CR, Kulesz-Martin M, Bagby G, Sears RC. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. BMC Bioinformatics 2008;9:520.
22. Chua SW, Vijayakumar P, Nissom PM, Yam CY, Wong VV, Yang H, *et al*. A novel normalization method for effective removal of systematic variation in microarray data. Nucleic Acids Res 2006;34:e38.
23. Sehhati MR, Dehnavi AM, Rabbani H, Javanmard SH. Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence. J Med Signals Sens 2013;3:87-93.
24. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389-422.
25. Berger JA, Hautaniemi S, Mitra SK, Astola J. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. IEEE/ACM Trans Comput Biol Bioinform 2006;3:2-16.
26. Gharibi A, Sehhati MR, Vard A, Mohebian MR. Identification of gene signatures for classifying of breast cancer subtypes using protein interaction database and support vector machines. In: Computer and Knowledge Engineering (ICCKE), 2015, 5th International Conference on. Iran: Mashhad; IEEE; 2015.
27. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507-17.
28. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 2004;31:91-103.
29. Maldonado S, Weber R. A wrapper method for feature selection using support vector machines. Inf Sci 2009;179:2208-17.
30. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273-324.
31. Destercke S, Strauss O. Kolmogorov–Smirnov test for interval data. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Switzerland: Springer; 2014.
32. Heeren T, D'Agostino R. Robustness of the two independent samples *t*-test when applied to ordinal scaled data. Stat Med 1987;6:79-90.
33. Birnbaum ZW. On a use of the Mann-Whitney statistic. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics. Vol. 1. California: Berkeley; The Regents of the University of California; 1956. p. 13-17.
34. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: Micro Machine and Human Science, MHS'95, Proceedings of the Sixth International Symposium on. Japan : Nagoya; IEEE; 1995. p. 39-43.
35. Eberhart RC, Shi Y, Kennedy JF. Swarm Intelligence (The Morgan Kaufmann Series in Evolutionary Computation). 2001. p. 81-86.
36. Sahu A, Panigrahi SK, Pattnaik S. Fast convergence particle swarm optimization for functions optimization. Procedia Technol 2012;4:319-24.
37. Zhou X, Li Z, Dai Z, Zou X. QSAR modeling of peptide biological activity by coupling support vector machine with particle swarm optimization algorithm and genetic algorithm. J Mol Graph Model 2010;29:188-96.
38. Chinnaswamy A, Srinivasan R. Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data, In: Snášel V, Abraham A, Krömer P, Pant M, Muda A, editors. Innovations in Bio-Inspired Computing and Applications. Switzerland: Springer, Cham; 2016. p. 229-39.
39. Jain I, Jain VK, Jain R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl Soft Comput 2018;62:203-15.
40. Liu Y, Zheng X, Wang B, Zhou Sh, Zhou Ch. The optimization of DNA encoding based on chaotic optimization particle swarm algorithm. J Comput Theor Nanosci 2016;13:443-9.
41. Panda A, Ghoshal S, Konar A, Banerjee B, Nagar AK. Static learning particle swarm optimization with enhanced exploration and exploitation using adaptive swarm size. In: IEEE Congress on Evolutionary Computation (CEC 2016), Canada: Vancouver; 2016. p. 1869-76.
42. Chu Y, Mi H, Liao H, Ji Z, Wu QH. A fast bacterial swarming algorithm for high-dimensional function optimization. In: IEEE Congress on Evolutionary Computation, CEC 2008.(IEEE World Congress on Computational Intelligence), Hong Kong: IEEE Service Center; 2008. p. 3134-39.
43. Tran B, Xue B, Zhang M. Improved PSO for feature selection on high-dimensional datasets. In: Asia-Pacific Conference on Simulated Evolution and Learning. Lecture Notes in Computer Science ((LNCS, volume 8886), Cham, Switzerland: Springer; 2014. p. 503-15.
44. Kuksa PP, Min MR, Dugar R, Gerstein M. High-order neural networks and kernel methods for peptide-MHC binding prediction. Bioinformatics 2015;31:3600-7.
45. Kazemian HB, Yusuf SA, White K. Signal peptide discrimination and cleavage site identification using SVM and NN. Comput Biol Med 2014;45:98-110.
46. Lira F, Perez PS, Baranauskas JA, Nozawa SR. Prediction of antimicrobial activity of synthetic peptides by a decision tree model. Appl Environ Microbiol 2013;79:3156-9.
47. Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Appl 1998;13:18-28.
48. Zhang GL, Petrovsky N, Kwoh CK, August JT, Brusic V. PRED(TAP): A system for prediction of peptide binding to the human transporter associated with antigen processing. Immunome Res 2006;2:3.
49. Bhasin M, Raghava GP. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics 2004;20:421-3.
50. Wu KP, Wang SD. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Pattern Recognit 2009;42:710-7.
51. Raudys A, Long J. MLP based linear feature extraction for nonlinearly separable data. Pattern Anal Appl 2001;4:227-34.
52. Wei SH, Balch C, Paik HH, Kim YS, Baldwin RL, Liyanarachchi S, *et al*. Prognostic DNA methylation biomarkers in ovarian cancer. Clin Cancer Res 2006;12:2788-94.
53. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M, *et al*. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 2009;25:2537-43.
54. Maetschke S, Towsey MW, Boden M. BLOMAP: An encoding of amino acids which improves signal peptide cleavage site prediction. 3rd Asia Pacific Bioinformatics Conference, Singapore; 2005. p. 141-50.
55. Goryński K, Safian I, Grądzki W, MarszałłJerzy MP, Krysiński J, Goryński S, *et al*. Artificial neural networks approach to early lung cancer detection. Central European Journal of Medicine 2014;9:632-41.

56. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl 2011;38:9573-9.
57. Abd El-Rehim DM, Ball G, Pinder SE, Rakha E, Paish C, Robertson JF, *et al.* High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. Int J Cancer 2005;116:340-50.
58. Quinlan JR. Bagging, Boosting, and C4. 5. In: AAAI/IAAI. Vol. 1. California: Menlo Park; 1996. p. 725-30.
59. Salzberg S. Book Review: C4. 5: Programs for machine learning. Machine Learning 1993;16:235-40.
60. Vlahou A, Schorge JO, Gregory BW, Coleman RL. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. J Biomed Biotechnol 2003;2003:308-14.
61. Su Y, Shen J, Qian H, Ma H, Ji J, Ma H, *et al.* Diagnosis of gastric cancer using decision tree classification of mass spectral data. Cancer Sci 2007;98:37-43.
62. Mousavizadegan M, Mohabatkar H. An evaluation on different machine learning algorithms for classification and prediction of antifungal peptides. Med Chem 2016;12:795-800.
63. Tsai MH, Wang HC, Lee GW, Lin YC, Chiu SH. A decision tree based classifier to analyze human ovarian cancer cDNA microarray datasets. J Med Syst 2016;40:21.
64. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J 2009;18:127-31.
65. Ellis PD. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge, UK: Cambridge University Press; 2010.
66. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009;45:427-37.
67. Chawla NV. Data mining for imbalanced datasets: An overview. In: Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer; 2009. p. 875-86.
68. Rubin A. Statistics for Evidence-Based Practice and Evaluation. 3rd Edition, Boston, MA: Cengage Learning; 2012.
69. Suryanarayanan S, Reddy NP, Canilang EP. A fuzzy logic diagnosis system for classification of pharyngeal dysphagia. Int J Biomed Comput 1995;38:207-15.


## BIOGRAPHIES

**Fatemeh Safaei Firouzabadi** was born in Tehran, Iran, in 1989. She received her undergraduate degree in Physics Engineering from Islamic Azad University, Science and Research Branch, Tehran, Iran in 2013. In 2018, she succeeded to complete a master program in Biomedical Engineering from Isfahan University of Medical Sciences, Isfahan, Iran. She focused her master thesis on "Finding an appropriate method for extraction of protein biomarkers to classify different cancers using protein microarray". Her research interests are Bioinformatics, Computational Biology, and Biological Modelling.

**Email:** fatemeh.safaei.f@gmail.com

**Alireza Vard** received his B.Sc. degree in software engineering from University of Isfahan, Isfahan, Iran in 2004, and he got his M.Sc. and Ph.D. degrees in computer engineering from University of Isfahan, in 2007 and 2012 respectively. He is currently an assistant professor at Department of Biomedical Engineering, Isfahan University of Medical Sciences, Isfahan, Iran. His research interests include medical image processing, pattern recognition, design and development of medical data processing software.

**Email:** vard@amt.mui.ac.ir

**Mohammadreza Sehhati** is an assistant professor at the Isfahan University of Medical Sciences, in the Biomedical Engineering Department. He received the BS and MS degrees in biomedical engineering from Shahed University and University of Tehran, Tehran, Iran, respectively. He obtained his PhD at the Isfahan University of Medical Sciences in 2015. His research interests include bioinformatics, machine learning, data mining, and image processing.

**Email:** mr.sehhati@amt.mui.ac.ir

**Mohammadreza Mohebian** studied biomedical engineering at the University of Isfahan. In his B.Sc. project, he looked at "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning". He is graduated with the master degree in biomedical engineering from the University of Isfahan too. In his M.Sc. he worked on "Non-invasive Decoding of the Motoneurons: A Guided Source Separation method based on Convolution Kernel Compensation with Clustered Initial Points". He ranked in the top 10 percent of the B.Sc. students with the same entrance date and was the first rank in M.Sc. degree in university. Moreover, he developed the first website for using machine learning and data mining in clinical data in Iran. Presently, he is studying on biomedical engineering at electrical and computer department of Saskatchewan University.

**Email:** van_mohebian@yahoo.com