Amino Acids (2014) 46:977–993
DOI 10.1007/s00726-013-1659-x

ORIGINAL ARTICLE

IDM-PhyChm-Ens: Intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids

Safdar Ali · Abdul Majid · Asifullah Khan

Received: 14 October 2013 / Accepted: 20 December 2013 / Published online: 4 January 2014

Abstract Development of an accurate and reliable intelligent decision-making method for the construction of cancer diagnosis system is one of the fast growing research areas of health sciences. Such decision-making system can provide adequate information for cancer diagnosis and drug discovery. Descriptors derived from physicochemical properties of protein sequences are very useful for classifying cancerous proteins. Recently, several interesting research studies have been reported on breast cancer classification. To this end, we propose the exploitation of the physicochemical properties of amino acids in protein primary sequences such as hydrophobicity (Hd) and hydrophilicity (Hb) for breast cancer classification. Hd and Hb properties of amino acids, in recent literature, are reported to be quite effective in characterizing the constituent amino acids and are used to study protein foldings, interactions, structures, and sequence-order effects. Especially, using these physicochemical properties, we observed that proline, serine, tyrosine, cysteine, arginine, and asparagine amino acids offer high discrimination between cancerous and healthy proteins. In addition, unlike traditional ensemble classification approaches, the proposed '*IDM-PhyChm-Ens*' method was developed by combining the decision spaces of a specific classifier trained on different feature spaces. The different feature spaces used were amino acid composition, split amino acid composition, and pseudo amino acid composition. Consequently, we have exploited different feature spaces using Hd and Hb properties of amino acids to develop an accurate method for classification of cancerous protein sequences. We developed ensemble classifiers using diverse learning algorithms such as random forest (RF), support vector machines (SVM), and K-nearest neighbor (KNN) trained on different feature spaces. We observed that ensemble-RF, in case of cancer classification, performed better than ensemble-SVM and ensemble-KNN. Our analysis demonstrates that ensemble-RF, ensemble-SVM and ensemble-KNN are more effective than their individual counterparts. The proposed '*IDM-PhyChm-Ens*' method has shown improved performance compared to existing techniques.

Keywords Cancer classification · Ensemble classifier · Random forest · Breast cancer · Protein primary sequences · Amino acid · Physicochemical properties · Hydrophobicity and hydrophilicity

Matlab based codes developed for this study can be provided to academicians on request.

S. Ali · A. Majid (✉) · A. Khan
Department of Computer and Information Sciences, Pakistan Institute of Engineering, and Applied Sciences, Nilore, Islamabad 45650, Pakistan
e-mail: abdulmajiid@pieas.edu.pk

S. Ali
e-mail: safdarali@pieas.edu.pk

A. Khan
e-mail: asif@pieas.edu.pk

Introduction

Identification of cancerous tissues in early stages provides higher chances of survival and is helpful in avoiding unnecessary toxicity (Milenković et al. 2013). Consequently, development of an efficient and reliable intelligent decision-making method for cancer classification system is highly desirable. Such system has the capability to provide adequate information for cancer diagnosis and drug discovery. Cancer is one of the rapidly growing diseases in the world (Caroline et al. 2012) and it is the second largest

cause of death after heart-related diseases (William 2010). Breast cancer, colorectal cancer, lung cancer, prostate cancer, and pancreatic cancer are the major types of cancers. The American Cancer Society (2013) estimated that, in 2013, approximately 226,870 women in the US would diagnose with breast cancer and about 39,510 women would die due to breast cancer. In western countries, due to early detection and treatment of breast cancer, approximately 89 % of women diagnosed with breast cancer remain still alive 5 years after their diagnosis. Every year, about 1.5 million cases of women breast cancer are registered worldwide. Approximately 90,000 out of 1.5 million cases are reported in Pakistan (Bray et al. 2004). Every year, approximately 40,000 women die with breast cancer in Pakistan; it is the highest figure in Asia (Nasim et al. 2012). Only 10 % of the breast cancer cases (i.e., 2,250 out of 90,000) are diagnosed and treated, while 75 % of the patients do not have treatment facility and thus die within 5 years after diagnosis.

Cancer development is a multistep process in which normal cells of the body go through stages that ultimately change them to abnormal cells without any control. The protein signals, coded by a very small proportion of the total genes in each cell, regulate the cell growth and division. These regulatory genes include two groups of genes called proto-oncogenes and tumor suppressor genes. A sequence of mutations in the DNA of any group of these growth-controlling genes can eventually lead to cancer.

The rapid development in proteome and genome resulted in the generation of a large number of gene and protein sequences. The use of protein sequencing is ever increasing and consequently, size of protein databases is exponentially growing with time. The rapid increase in size of protein sequence databases is challenge for extracting useful information for the diagnosis of diseases and drug discovery. Protein may help for prediction of breast cancer prognosis. It has been identified that the levels of expression of some 1,200 genes that are directly controlled by the enzyme, EZH2, correlates with the aggressiveness of breast cancer cases (Jene-Sanz et al. 2013). Protein sequences have been used for the classification of ovarian cancer (Ji-Yeon et al. 2013), lung cancer (Ramani and Jacob 2013a), colon cancer, and breast cancer (Munteanu et al. 2009).

Role of feature extraction is very crucial in the development of a reliable and effective method for the construction of diagnosis system (Xin et al. 2012). Extracted feature vectors are used by such system to establish the classification model. An improvement in performance of support system is obtained via integration of the decision(s) of base classifier(s). A good feature set is supposed to be highly correlated within a class and uncorrelated with other classes. In literature, several feature extraction strategies and classification systems have been used for breast cancer diagnosis (Şahan et al. 2007; Einipour 2011; Eshlaghy et al. 2013; Bing-Yu et al. 2011; Xu et al. 2007; Saima et al. 2013; Emmanuel et al. 2010; Ramani and Jacob 2013b, c). A detail review of different feature extraction and selection techniques, in particular for protein sequence analysis, was provided in (Yvan et al. 2007). In another study, multiple sonographic and textural features based methods were developed for classification of benign and malignant breast tumors (Liao et al. 2010). Recurrence modeling using incidence and survival based population feature extraction strategy have shown good performance (Eshlaghy et al. 2013). Similarly, for classification of breast and colon cancers, useful features of topological indices (TIs) were extracted using graph theory (Munteanu et al. 2009). However, these features were used to develop multi-target quantitative proteome–disease relationship (QPDR) models. These models were based on simple multiple linear regression technique. Due to linear nature of QPDR models, it is not easy to accurately model the nonlinearity in the features to corresponding target labels. Therefore, performance of QPDR models was limited to 91.8 %. Clinical features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass were used for development of optimized-learning vector quantization (LVQ), Big-LVQ, and artificial immune recognition system (AIRS) (Goodman et al. 2002). On the other hand, in a recent study, SVM combined with evolutionary algorithms (EAs) using clinical features from a digitized image of a FNA of a breast mass has reported an accuracy up to 97.07 % (Ruxandra and Stoean 2013).

Proteins are major constituents of all cells and composed of a sequence of amino acids (Pierrick et al. 2013). Amino acids have different physicochemical properties, such as charge, mass, hydrophobicity (Hd) and hydrophilicity (Hb), etc. To the best of our knowledge, in previous studies, the physicochemical properties of amino acids in protein primary sequences such as Hd and Hb have not been exploited for breast cancer classification. These properties of amino acids are quite effective in characterizing the constituent amino acids and are thus helpful in the study of protein foldings, interactions, structures, and sequence-order effects (Chou 2005). In our opinion, these properties of amino acids would be helpful for breast cancer classification. Additionally, ensemble classification approach that combine the decision spaces of a specific classifier trained on different feature spaces has not been explored in the context of breast cancer classification. In this scenario, unlike traditional ensemble classification approaches, we developed *IDM-PhyChm-Ens* cancer classification methodology by combining the decision spaces of a specific classifier trained on different feature spaces.

In our proposed methodology, primary protein sequences were transformed in: (1) amino acid composition

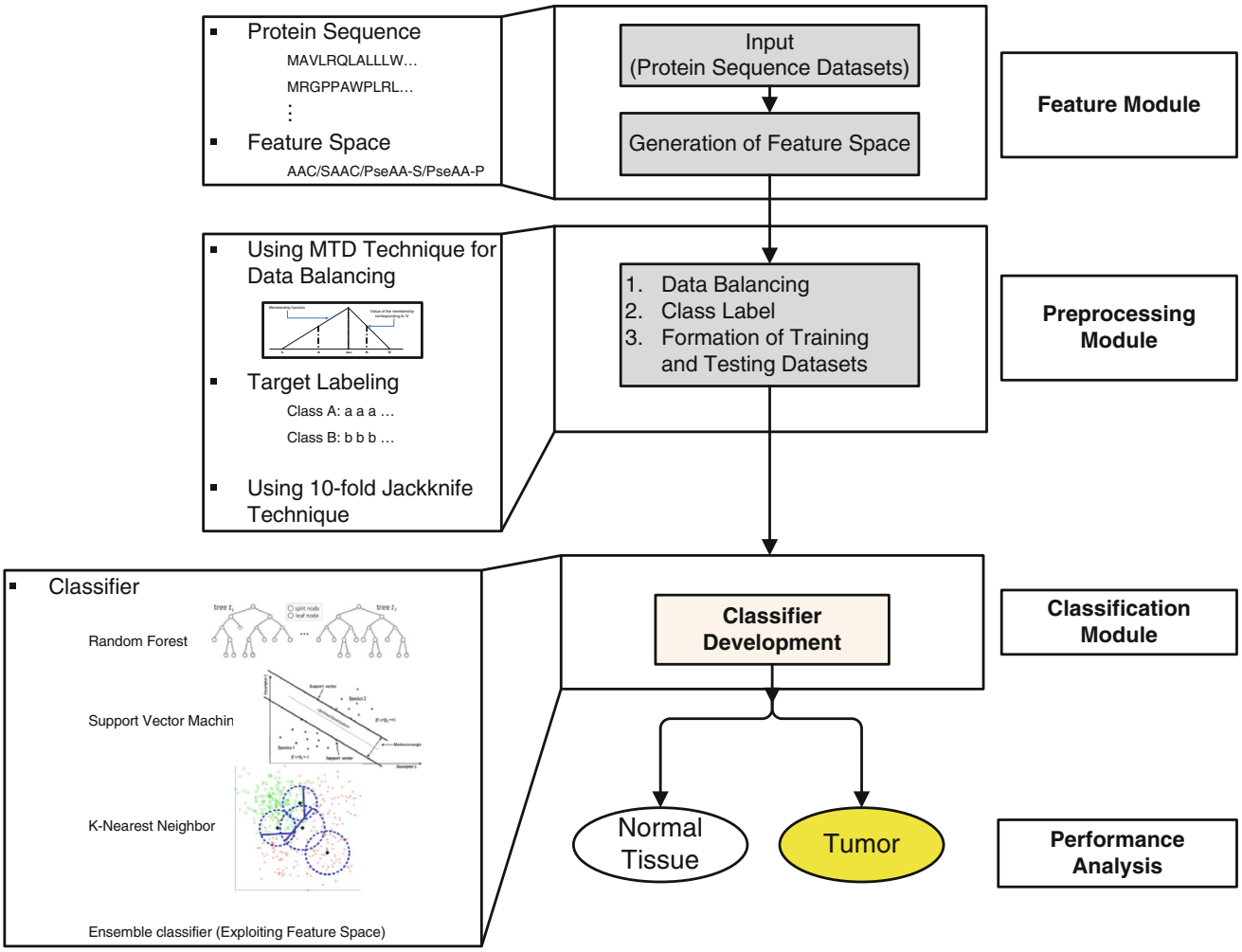

Fig. 1 Basic block diagram of the proposed *IDM-PhyChm-Ens* classification method

(AAC), (2) split amino acid composition (SAAC), (3) pseudo amino acid composition-series (PseAAC-S), and (4) pseudo amino acid composition-parallel (PseAAC-P) feature spaces. For data balancing, we have used mega-trend-diffusion (MTD) function to create diffuse samples for the minority class (Li et al. 2010). This function oversamples the minority class in the feature space. We have developed ensemble classifiers using diverse learning mechanisms of RF, SVM, and KNN trained on individual feature spaces.

Experiments were conducted under different feature spaces. The classification performance was reported using tenfold cross-validation data resampling technique. Our ensemble-RF and ensemble-SVM have achieved the highest accuracies of 99.48 and 97.63 %, respectively. For the breast/non-breast cancer, our analysis showed that performance of ensemble-RF, ensemble-SVM, and ensemble-KNN was enhanced by 7.43, 8.96, and 8.55 %, respectively, than their individual counterparts, using combined feature spaces of PseAAC-S and PseAAC-P. Further, we have observed that proline, serine, tyrosine, cysteine, arginine, and asparagine amino acids offer high discrimination for cancer using physicochemical properties.

The main contributions of this work are:

1. To develop an effective cancer classification methodology "*IDM-PhyChm-Ens*" by exploiting Hd and Hb physicochemical properties of protein primary sequences in different feature spaces,
2. To develop a high-performance ensemble such as ensemble-RF, ensemble-SVM, and ensemble-KNN for classification of healthy and cancerous protein sequences, and
3. To explore which amino acids are more affected in terms of its physicochemical properties, when a normal cell converts to cancerous one.

Proposed *IDM-PhyChm-Ens* cancer classification methodology

Figure 1 shows the basic block diagram of the proposed intelligent decision-making methodology for classification of cancerous and healthy protein sequences. The proposed cancer classification method *IDM-PhyChm-Ens* combines the decision spaces of a specific classifier trained on different feature spaces. *IDM-PhyChm-Ens* consists of three main modules: feature space generation, preprocessing, and classifier development modules. In preprocessing module, data balancing is performed using mega-trend-diffusion (MTD) function to create diffuse samples for the minority class (Li et al. 2010). This function oversamples the minority class in feature space.

Feature space generation module

In feature module, primary protein sequences using Hd and Hb properties of amino acids are transformed into AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces.

Protein sequence datasets

In this work, the performance of the proposed *IDM-PhyChm-Ens* method has been evaluated using two real datasets of human protein primary sequences for cancer/non-cancer and breast/non-breast cancers (Sjoblom et al. 2006; Dobson et al. 2004). They extracted cancer-related proteins after the experimental analysis of 13,023 genes in 11 breast and 11 colorectal cancers. The first dataset has 865 non-cancer protein sequences. The second dataset contains a total of 191 cancer protein sequences, in which 122 protein sequences are related to breast cancer sequences.

Generally, a protein sequence is represented by a series of amino acid codes in the form of English letters. From one-dimensional point of view, a protein sequence contains characters from the 20-letter native amino acid alphabet A = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. Following are two examples of non-cancer (Seq. 1) and breast cancer (Seq. 2) protein primary sequences:

Seq. 1: MSWQSYVDDHLMCDVEGNHLTAAAILGQ DGSVWAQSAKFPQLKPQEI…

Seq. 2: MAARPLPVSPARALLLALAGALLAPCEAR GVSLWNQGRADEVVSASV…

where English letters A, C, E,…, represent the standard single-letter codes of different native amino acids in protein sequence.

Variation in amino acids composition of cancer proteins

We have analyzed the variation of amino acids composition in cancer- and breast cancer-related protein sequences with reference to non-cancer proteins for discrimination of cancer and healthy proteins. We observed from experiments that this analysis of variation of amino acid composition is quite prospective for cancer classification.

Proteins perform a wide variety of functions in the body and serve as essential components of body tissues, enzymes, and immune cells. They also act as hormonal messengers and enzyme catalysts. Proteins are made of polymers of amino acids. Amino acids help in proper growth and sustenance of the body. Traditionally in nutrition textbooks, amino acids have been categorized into two types: essential amino acids and non-essential amino acids. Essential amino acids are obtained from our diet. They include: isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine. On the other hand, body on its own manufactures non-essential amino acids. These amino acids are not necessarily obtained from diet. Non-essential amino acids consist of glutamate, alanine, aspartate, glutamine, arginine, proline, serine, tyrosine, cysteine, taurine, and glycine.

All of the native amino acids have similar general structure, except proline. The general structure of native amino acids holds a central α-carbon to an amino group, a carboxylate group, a hydrogen atom, and an R side chain group. Several side chains are hydrophobic, while others are hydrophilic. On the contrary, in proline, amino group is secondary and is formed by ring closure between the R group and the amino nitrogen. In proline, rotation about carbon is impossible because of its rigidity to the peptide chain. This structural characteristic of proline is important in the structure and function of proteins with high proline content. Studies have revealed that proline plays a special role in cancer metabolism (Phang and Liu 2012). The proline biosynthetic pathway is considered a key mediator of breast tumor cell metastasis (Richardson 2011).

On the other hand, amino acids such as serine, threonine, tyrosine, asparagine, and glutamine have contained polar hydroxyl group and thus are called “hydrophilic” or “water-loving”. Polar hydroxyl group enables serine, threonine, tyrosine, asparagine, and glutamine to participate in hydrogen bonding, which is an important factor in protein structure. The hydroxyl groups serve other functions in proteins. Hydrogen-bonding capability of asparagine and glutamine has a significant effect on protein stability. Each protein has its own unique sequence of amino acids and native conformation. Any change in the sequence of amino acids, even one amino acid change, can potentially change the capability of protein to function. Consequently, study of variation of amino acids composition in cancer proteins with reference to non-cancer proteins is important. Because the Hd and Hb properties of amino acids are affected by variation of amino acids in cancer proteins. Hence, such Hd and Hb

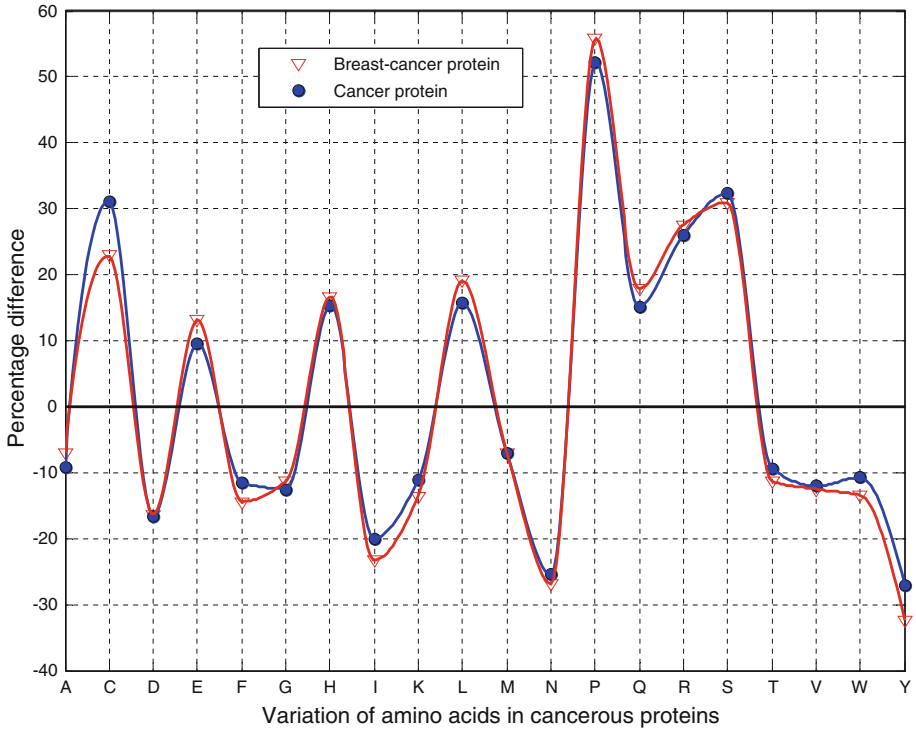

Fig. 2 Variation of amino acids compounds in general cancer and breast cancer protein sequences with reference to non-cancer proteins

properties of amino acids are quite useful in the classification of cancer.

In Fig. 2, we observed the change in the concentration of amino acid compounds in cancer proteins for general cancer and breast cancer proteins sequences with reference to non-cancer proteins. In Fig. 2, curves of cancer and breast cancer proteins show almost similar behavior. The absolute percentage differences in all of the amino acids are relatively higher in breast cancer proteins as compared to the general cancer proteins, except C and S amino acids.

In Fig. 2, the values above zero-line indicate the percentage increase of various amino acids such as C, E, H, L, P, Q, R, and S in the cancer-related proteins with reference to non-cancer proteins. Similarly, the values below zero-line represent the percentage decrease of various amino acids such as A, D, F, G, I, K, M, N, T, V, W, and Y in the cancer-related proteins with reference to non-cancer proteins. The maximum percentage increase of 52.27 and 55.91 % is observed for 'Proline' amino acid in both general cancer and breast cancer protein sequences, respectively. On the other hand, the maximum percentage decrease of 27.05 and 32.43 % is observed for 'Tyrosine' amino acid in both general cancer and breast cancer protein sequences, respectively. We have noticed that higher change in the values for P, S, Y, C, R, and N offer high discrimination between breast cancerous and healthy proteins.

From Fig. 2, it is evident that the composition of all of the amino acids is altered in cancerous proteins. Based upon these results, we expect that the variation composition of amino acid compounds in cancer proteins may help in the treatment of cancer and drug targeting. However, in the current work, the focus is to understand the role of such distribution of amino acid compounds in protein primary sequences using physicochemical properties in early stages of cancer development.

Feature generation strategy

An important issue in applying classifier to protein sequence classification is how to encode protein sequences, i.e., how to represent the protein sequences as the input of the classifier. Indeed, sequences may not be the best representation at all. Good input representations make it easier for the classifier to recognize underlying regularities. Thus, good input representations are crucial to the success of classifier learning. To generate effective sequence-based features for protein system, one of the fundamental steps is to formulate the protein sequences into mathematical expression that could be capable to reflect the inherent correlation with corresponding target labels.

For protein sequences, different feature spaces exhibited different characteristics and yielded different classification results. Therefore, instead of using conventional feature selection (wrappers, filters) and transform based extraction strategies, we used sequence-order and correlation-based feature generation strategies of AAC, SAAC, PseAAC-S, and PseAAC-P. Protein sequence data contain intrinsic dependencies between their constituent elements. Given a

protein sequence over the amino acid alphabets, the dependencies between neighboring elements can be modeled by generating all the contiguous present in sequence data.

Thus, exploiting dependencies in the data increases the richness of the representation. The feature spaces used in this study try to exploit discriminating information about protein sequence order, which is useful for classification. These feature spaces have been employed in nearly all fields related to protein attribute, such as predicting enzyme family class (Qiu et al. 2010), protein subcellular localization (Khan et al. 2010), outer membrane proteins (Lin 2008), and protein structural class (Sahu and Panda 2010). Further, in this study our aim is to determine which classifier has more discriminant exploitation of certain types of feature spaces. In this way, we find the effective and best combination of "classifier+feature-spaces" for cancer classification.

Features derived from physicochemical properties of protein sequences are quite helpful in classifying cancerous proteins. All amino acids have different physicochemical properties owing to their differences in side chains. Some side chains are hydrophobic, while others are hydrophilic. Generally, amino acids based on physicochemical properties are grouped into three main classes: hydrophobic, hydrophilic, and charged residues. Hd and Hb properties of protein are effectively used to convert amino acid codes into numerical values for the classification of cancer disease. The values of Hd and Hb of amino acids are taken from (Tanford 1962; Hopp and Woods 1981) to generate four feature spaces. We explain these feature spaces below:

AAC is formed using 20 discrete numbers, in which each number represents the occurrence frequency of 20 native amino acids in the protein sequence (Chou and David 1999; Khan et al. 2010). The protein primary sequence can be expressed as:

$$\mathbf{x} = [p_1,\ p_2, \ldots, p_{20}]^T \tag{1}$$

where, $p_1, p_2, \ldots, p_{20}$ are the native composition of 20 amino acids of a protein $\mathbf{x}$.

For the generation of SAAC feature space, the protein primary sequence is divided into three distinct parts, called the N-terminal, internal segments and C-terminal (Chen et al. 2006; Maqsood et al. 2012). Figure 3 shows the construction of N-terminal, internal segments and C-terminal of the SAAC feature space. The amino acid compositions at N-terminal and C-terminal segments are calculated separately. The length of each segment, N-terminal, internal and C-terminal is set equal to 20. Thus, the length of each feature vector in SAAC model is 60D (dimension).

Generally, in simple AAC feature space, important position-specific or hidden information in protein sequences would be entirely lost. Thus, the concept of PseAAC was

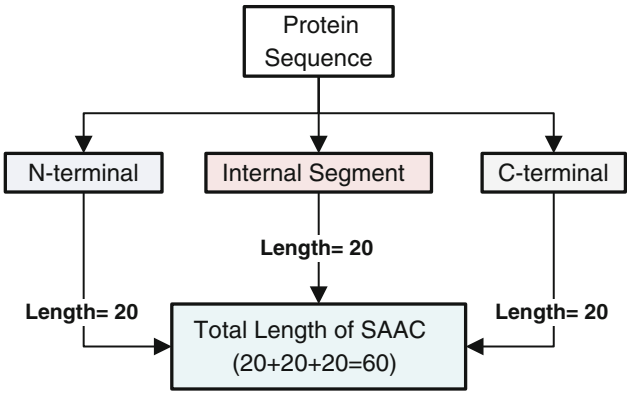

Fig. 3 Construction of *SAAC* features

introduced to avoid completely losing the sequence-order information for representing the protein sample. The PseAAC feature generation strategy reflects better sequence order and length of a protein sequence sample (Chou 2005).

In this study, we have used two types of series and parallel PseAAC correlations. In series correlation (PseAAC-S), a protein feature vector is generated by $20 + i \times \lambda$ discrete components. The number of amino acid attributes selected is represented by 'i'. We used $\lambda = 20$ and $i = 2$ for hydrophobic and hydrophilic values to form 60D feature vector. The protein vector in series correlation is represented as:

$$\mathbf{x} = [p_1 \ldots p_{20} p_{21} \ldots p_{20+\lambda} p_{20+\lambda+1} \ldots p_{20+2\lambda}]^T \quad (\lambda < L) \tag{2}$$

with

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & \text{for } 1 < \mu < 20 \\ \dfrac{w\tau_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & \text{for } 21 < \mu < 20 + 2\lambda \end{cases} \tag{3}$$

where, L is the amino acid residues, $f_i (i = 1, 2, \ldots, 20)$ are the normalized occurrence frequencies of 20 native amino acids in the protein $\mathbf{x}$ and τ_j the jth-tier sequence-correlation factor computed according to $\tau_{2\lambda-1} = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \mathrm{Hd}_{i,i+\lambda}$ and $\tau_{2\lambda} = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \mathrm{Hb}_{i,i+\lambda}$. The constant factor w is the weighting factor. Here, empirically, we set $w = 0.05$.

In parallel correlation (PseAAC-P), a protein feature vector is represented by $20 + \lambda$ discrete components. Parallel correlation factor or sequence order-correlated factor is given as:

$$\theta_\lambda = \frac{1}{L - \lambda} \sum_{i=1}^{L-\lambda} \Theta(R_i, R_{i+\lambda}) \tag{4}$$

where, θ_λ represents λ-tier correlation factor, which reflects the sequence-order correlation between all the λ most contiguous residues along a protein chain. We used $\lambda = 20$, thus, generated a feature set consisting of 40D. The value of $\Theta(R_i, R_j)$ is calculated as follow:

$$\Theta(R_i,R_j)=\frac{1}{3}\left\{\left[\mathrm{Hd}(R_j)-\mathrm{Hd}(R_i)\right]^2+\left[\mathrm{Hb}(R_j)-\mathrm{Hb}(R_i)\right]^2\right\} \quad (5)$$

where, Hd and Hb are hydrophobicity, and hydrophilicity of ith amino acid R_i and jth amino acid R_j, respectively.

Preprocessing module

Generally, the size of medical data for positive examples is small as compared to negative examples. Thus, during classifier development stage, the decision boundary established by classifier is biased toward the majority class and the performance related to the minority class is affected. In the literature, this type of problem is managed to create diffuse samples for the minority class by applying different data balancing approaches (Chawla et al. 2002; Li et al. 2007; Muhammad et al. 2013). In this module, we have employed mega-trend-diffusion (MTD) function for oversampling the minority class in feature space (Li et al. 2010).

The classification capability is reported using various data resampling techniques. However, jackknife technique is considered the most objective and rigorous one and produces a unique outcome. Investigators have widely adopted jackknife technique to examine the quality of various classifiers (Khan et al. 2011; Mohabatkar 2010; Hayat and Khan 2011). Thus, we have employed tenfold jackknife technique to evaluate and compare the performance of models. In tenfold Jackknife technique, dataset is divided into ten parts. The 9/10th parts of the dataset are used to train the model and the remaining 1/10th part of the dataset was utilized to test the model. This step is repeated ten times using different training/testing data and average performance of the model is reported.

Classification module

This section explains the development of proposed *IDM-PhyChm-Ens* ensemble and individual classifiers. RF, SVM, and KNN are used as single learning algorithm for the generation of ensemble-RF, ensemble-SVM, and ensemble-KNN. The ensemble classifiers are constructed by combining the feature spaces at decision level. For performance comparison, RF, SVM, and KNN are also used for the development of individual classifiers. Now, we describe the ensemble and individual classifiers as follows:

Development of ensemble classifiers

Ensemble classifier may be formed using several diverse learning algorithms. Alternatively, ensemble classifier can be created with single learning algorithm. In case of several learners, ensemble is generated using the same training dataset to train each learner. Then their decisions are combined by simple voting (Khan et al. 2010; Hayat and Khan 2011), or more sophisticated methods like consensus theory (Benediktsson and Swain 1992; Džeroski and Ženko 2004). This type of ensemble that achieves diversity with multiple algorithms is called heterogeneous ensemble. On the other hand, homogeneous ensemble uses a single algorithm and achieves diversity on the same dataset or injecting randomness into the parameters of the learning algorithm.

Homogeneous ensemble was developed with a specific base learner, say, L using different feature spaces from the set $S=\{\mathrm{Fs}_1,\mathrm{Fs}_2,\ldots,\mathrm{Fs}_m\}$, where, $\mathrm{Fs}_1,\mathrm{Fs}_2,\ldots,\mathrm{Fs}_m$ are m different feature spaces. A feature vector $\mathbf{v}_j^i$ of ith feature space, s_i, is formed form a protein data vector $\mathrm{x_j}$ as: $\mathbf{v}_j^i=\mathrm{Fs}_i(\mathbf{x}_j)$, $i=1,2,\ldots,m$. For a base learner L, the predictions $\hat{y}_i^j$ are extracted as: $\hat{y}_i^j=L\left(\mathbf{v}_j^i\right)$. Each prediction $\hat{y}_i^j$ belongs to one of the k classes, i.e., $\{\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_m\}\in\{c^1,c^2,\ldots,c^k\}$. This assigned jth label to one of the class from the set $\{c^1,c^2,\ldots,c^k\}$. The value of ensemble Z_{ens}^j is computed using majority-voting mechanism for all m feature spaces as:

$$Z_{\mathrm{ens}}^j=\sum_i^m \Delta(\hat{y}_i^j,c^j),\quad j=1,2,\ldots,k \quad (6)$$

where, for our binary problem $k=2$ and $\Delta(\hat{y}_i^j,c^j)=\begin{cases}1 & \text{if } \hat{y}_i^j\in c^j\\ 0 & \text{otherwise}\end{cases}$. The data point $\mathbf{x}_j$ will be assigned to the class, which has maximum voting.

In this work, we have constructed homogeneous ensemble by varying feature spaces of different dimensions. Figure 4 demonstrates the various stages of the proposed ensemble method. Three algorithms RF, SVM, and KNN are selected as single learning algorithm for the construction of homogeneous ensemble-RF, ensemble-SVM, and ensemble-KNN by exploiting different feature spaces. Each algorithm displays a different inductive bias and learning hypotheses such as instance-based, trees and statistics. Thus, each algorithm provides potentially more independent and diverse predictions.

Development of individual classifiers

Detail information about RF, SVM, and KNN learning algorithms is available in the literature of computational science. RF and KNN algorithms are implemented in Matlab 2010 environment. Libsvm software is used for the implementation of SVM (Chang and Lin 2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. In future, we intend to make web server, which will be

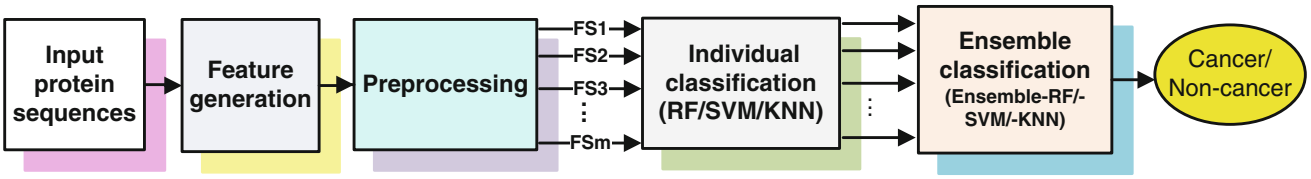

Fig. 4 Framework of the proposed *IDM-PhyChm-Ens* ensemble method

publicly available for readers and practitioners. Here, we explain implementation details of the learning algorithms.

Random forest classifier

RF algorithm uses an ensemble of classification trees (Balmain et al. 2003). It is being used to solve different types of problems in bioinformatics. It has performed well with small data size but complex data structures. It is a successful approach for combining unstable learners (Hastie et al. 2001), and random variable selection for tree building.

Working procedure of the RF algorithm is shown in Fig. 5. It uses two random mechanisms (1) random feature selection and (2) bootstrap data sampling. RF utilizes random sampling and ensemble properties that make it possible to attain accurate prediction and better generalization. For n features, a number $mtry < n$ is specified such that at each node, $mtry$ out of n features are randomly selected. The value of $mtry$ is selected to be $\sqrt{n}$. During forest growing, the value of $mtry$ remains constant. The best split on these $mtry$ features is used to split the node. Gini measure is used as a splitting criterion that selects the split of each node with the lowest impurity. Gini measure of impurity is given by the following Eq.

$$\text{Gini}(t, \mathbf{x}) = \frac{1}{N} \sum_{i=1}^{j} a_i I(t_{x_i}) \tag{7}$$

where, j is the number of children at node t, N is the number of samples, $I(t_{x_i})$ is Gini impurity, which gives the class label distribution in the node. $I(t_{x_i})$ for the variable $\mathbf{x}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j\}$ at node t is given by:

$$I(t_{x_i}) = 1 - \sum_{c=0}^{c} \left(\frac{n_{c_i}}{a_i} \right)^2 \tag{8}$$

where, n_{c_i} is the number of samples with value $\mathbf{x}_i$ belonging to class c, a_i is the number of samples with value $\mathbf{x}_i$ at node t. Consider a dataset D of N protein sequences for training: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i$, $i = 1, 2, 3, \ldots, n$, is ith protein vector corresponding to y_i target label. The output predicted values, for a given $\mathbf{x} \in D$, are given by $\{\hat{y}_1(\mathbf{x}), \hat{y}_2(\mathbf{x}), \ldots, \hat{y}_T(\mathbf{x})\}$, where $\hat{y}_t(\mathbf{x})$ is the predicted value of protein sequence ($\mathbf{x}$) by the tth tree. Outputs of all predicted values are aggregated using majority vote for final decision.

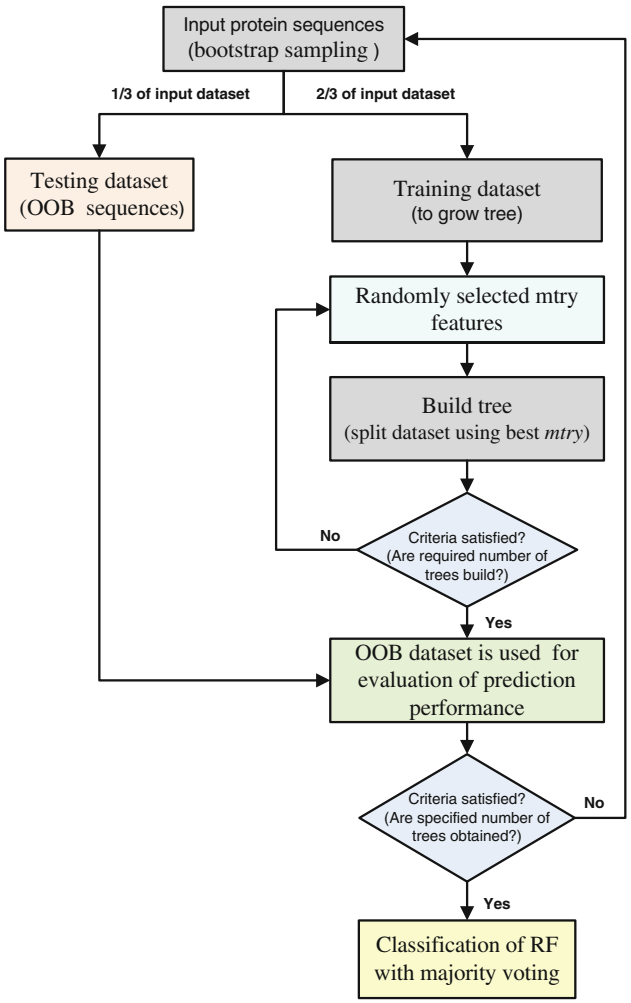

Fig. 5 Block diagram of the random forest classification

SVM classifier

SVM is extensively used to perform prediction and classification tasks (Vapnik 1995). It finds a decision surface that has maximum distance to the closest points in the training set. It generates a decision surface by mapping the training samples, support vectors (Ns), into a higher dimensional space. For a linearly separable data of N

training pairs $\langle \mathbf{x}_i, y_i \rangle$, the function of SVM decision surface is expressed as:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T . \mathbf{x} + \text{bias} \quad (9)$$

where, the coefficient $\alpha_i > 0$ is the Lagrange multiplier in an optimization problem. The function $f(\mathbf{x})$ is independent of the dimension of the feature space. To find an optimal hyperplane surface for non-separable patterns, solution of the following optimization problem is sought:

$$\Psi(\mathbf{w}, \zeta) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \zeta_i \quad (10)$$

subject to the condition $y_i(\mathbf{w}^T \Psi(\mathbf{x}_i) + \text{bias}) \geq 1 - \zeta_i, \zeta_i \geq 0$ where, C is the penalty parameter of the error term $\sum_{i=1}^{N} \zeta_i$. It represents the cost of constraint violation of those data point, which occurs on the wrong side of the decision boundary. The weight vector $\mathbf{w}$ minimizes the cost function term $\mathbf{w}^T\mathbf{w}$. For *SVM* decision function, we used radial basis function, $\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{(2\sigma)^2}\right)$, where σ shows the width of Gaussian function. During optimization, the optimal values of C and σ are obtained to 100 and 0.05, respectively.

KNN classifier

KNN approach is a simple but effectively used in pattern recognition problems. The proximity based KNN classifies the objects based on closest training examples in the feature space. Here, we consider Euclidean distance as a metric to measure the proximity. For a protein sequence $\mathbf{x}$ under consideration, according to the nearest neighbor principle, we find distance d between $\mathbf{x}$ and $\mathbf{x}_i$ as:

$$d(\mathbf{x}, \mathbf{x}_i) = 1 - \frac{\mathbf{x}.\mathbf{x}_i}{\|\mathbf{x}\|\|\mathbf{x}_i\|} \ (i = 1, 2, 3, \ldots, N) \quad (11)$$

where, $\mathbf{x}.\mathbf{x}_i$ is the dot product of vectors, $\mathbf{x}$ and $\mathbf{x}_i$; and $\|\mathbf{x}\|$ and $\|\mathbf{x}_i\|$ are, respectively, their modulus. Then the minimum of the distances is computed as:

$$d(\mathrm{x}, \mathrm{x}_k) = \text{Min}\{d(\mathrm{x}, \mathrm{x}_1), d(\mathrm{x}, \mathrm{x}_2), \ldots, d(\mathrm{x}, \mathrm{x}_N)\} \quad (12)$$

The query protein sequence $\mathbf{x}$ is assigned the category corresponding to the training protein $\mathbf{x}_k$. During the implementation, we have tried different k values 1, 3, 5 and 7. However, we found the best results for $k = 1$.

Results and discussions

Several experiments were conducted to explore the effectiveness of various classification algorithms in different feature spaces. We analyzed experimental results of the proposed *IDM-PhyChm-Ens* method using: (1) cancer/non-cancer data, and (2) breast/non-breast cancer data. We adopted tenfold cross-validation protocol to evaluate the performance of proposed scheme. Performance of the proposed method is assessed in individual and combined feature spaces using different measures such as accuracy (Acc), sensitivity (Sn), specificity (Sp), G-mean, F-score, and Mathews correlation coefficient (MCC).

Performance analysis of individual and ensemble-RF classifier

In this section, we discuss our findings regarding individual and ensemble-RF classifiers using different feature spaces. Figure 6 indicates the performance accuracy of RF classifier with increasing number of trees (ntree) using different feature spaces. We computed the accuracy of RF classifier by varying ntree from 1 to 200 with equal interval of 20 trees. This figure shows that initially the performance of RF classifier is considerably improved with the increase of number of trees. Beyond a certain limit (20–80), there is no appreciable change in overall accuracy. We observed that SAAC feature space has provided the best overall accuracy of 96.76 % for 90 trees, followed by PseAAC-S feature space with 95.84 % of accuracy for 40 trees. However, AAC and PseAAC-P feature spaces have given the overall accuracies of 95.78 and 95.61 % for 60 and 160 trees, respectively.

Figure 7 depicts overall accuracies against number of trees in various feature spaces for breast/non-breast cancer dataset. It is observed that PseAAC-S feature space has provided the best overall accuracy of 95.34 % for 40 trees. However, PseAAC-P feature space has yielded an accuracy of 95.13 % for 30 trees. We have used optimal numbers of trees to compute the results of individual and ensemble classifiers.

Table 1 highlights the performance of RF-based individual and ensemble classifiers using different feature spaces of AAC, SAAC, PseAAC-S, and PseAAC-P. It is observed that, for cancer/non-cancer dataset, individual-RF has given the highest values of G-mean of 96.47 % and F-score of 96.64 % for SAAC feature space. However, for other feature spaces, individual-RF has provided values of G-mean and F-score near to 95.37 and 95.63 %, respectively. However, ensemble-RF has yielded the highest value of 99.48 % for Acc, G-mean, and F-score using the combined feature space of SAAC+PseAAC-S. It is inferred that when the predications of RF classifier using PseAAC-S feature space are combined with SAAC feature space, the performance of ensemble-RF is enhanced up to 3.01 % in terms of G-mean. Therefore, for cancer/non-cancer dataset, we observed that SAAC feature space provides sufficient information for classification.

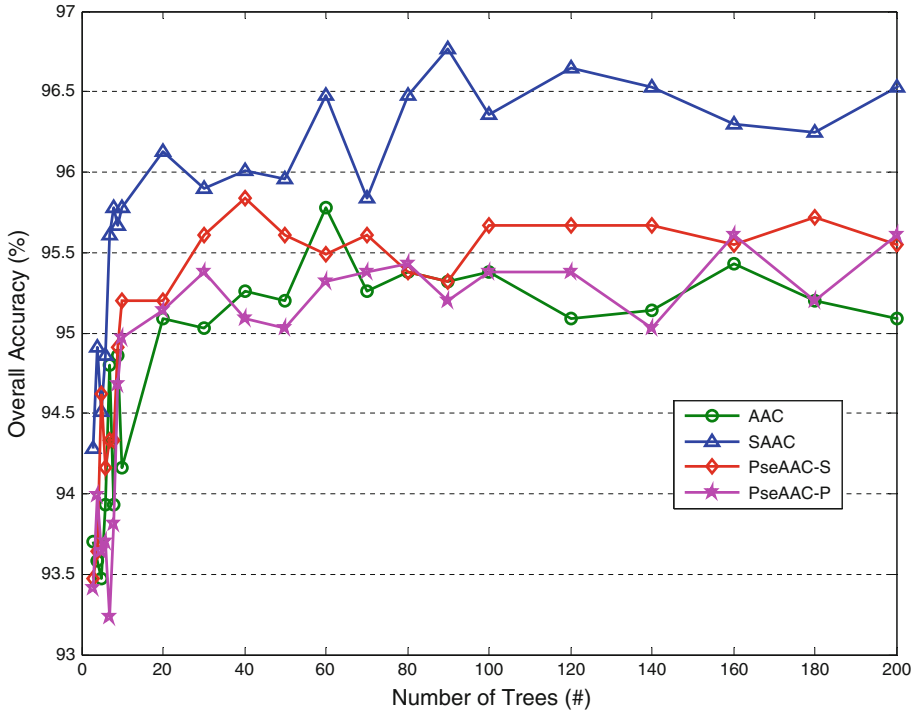

Fig. 6 Classification accuracies of RF vs. number of trees for cancer/non-cancer using different feature spaces of AAC, SAAC, PseAAC-S and, PseAAC-P

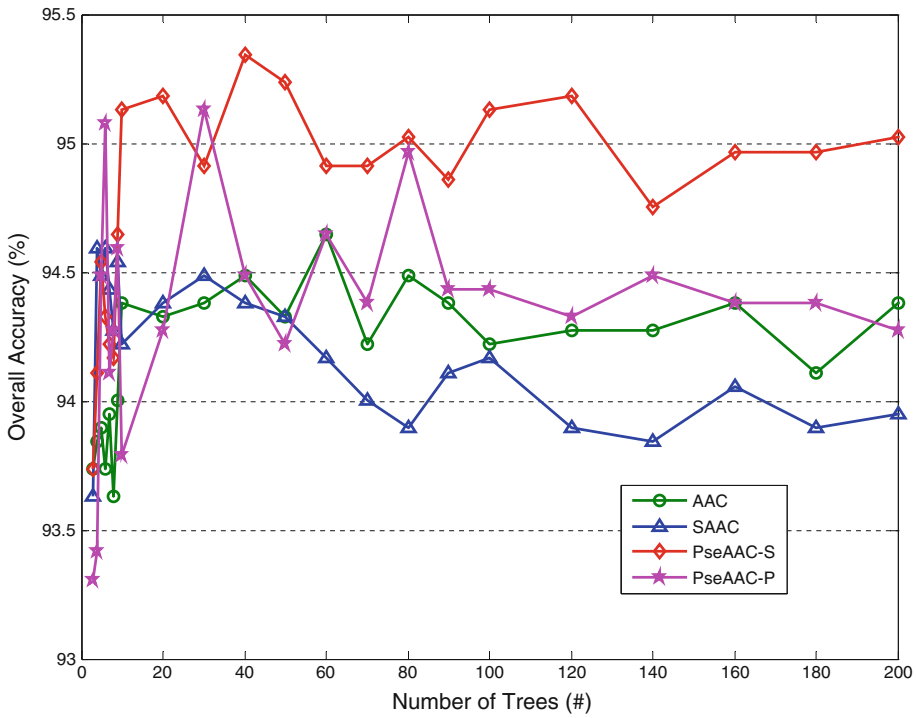

Fig. 7 Classification accuracies of RF vs. number of trees for breast/non-breast cancer using different feature spaces of AAC, SAAC, PseAAC-S and, PseAAC-P

Using breast/non-breast cancer dataset, for PseAAC-S feature space, Table 1 indicates that individual-RF provided the highest values of Acc (95.24 %), Sn (93.90 %), Sp (96.57 %), G-mean (95.23 %), and F-score (95.17 %). However, the same feature space has the highest MCC value of 63.84 %. Using combined feature space (AAC+SAAC), ensemble-RF classifier has given the highest values of Acc (97.91 %), Sp (96.79 %), G-mean (97.91 %), and F-score (97.94 %) for breast cancer dataset. It is observed that when predictions of RF classifier using PseAAC-S feature space are combined with the predictions using PseAAC-P feature space, the performance of ensemble-RF classifier is enhanced by 7.43 % for MCC measure. We inferred that PseAAC-S feature space provides better information with

Table 1 Performance of RF-based individual and ensemble classifiers using different feature spaces

Method	Feature space	Cancer/non-cancer						Breast/non-breast cancer					
		Acc	Sn	Sp	G_{mean}	F-score	MCC	Acc	Sn	Sp	G_{mean}	F-score	MCC
Individual-RF	AAC	95.78	99.19	91.56	95.30	95.55	72.01	94.65	91.01	98.07	94.47	94.34	60.51
	SAAC	**96.76**	**99.88**	**93.18**	**96.47**	**96.64**	72.00	94.49	89.29	**98.72**	93.89	93.71	58.66
	PseAAC-S	95.85	99.65	91.56	95.52	95.78	72.02	**95.13**	**93.90**	96.57	**95.23**	**95.17**	**63.84**
	PseAAC-P	95.61	99.54	91.21	95.28	95.56	72.00	94.49	91.11	97.86	94.43	94.29	60.63
Ensemble-RF	AAC+SAAC	99.08	98.84	99.31	99.07	99.07	70.22	**97.91**	99.04	**96.79**	**97.91**	**97.94**	70.05
	AAC+PseAAC-S	98.90	98.61	99.19	98.90	98.90	70.12	97.16	99.36	94.97	97.14	97.22	70.77
	AAC+PseAAC-P	98.90	98.50	99.31	98.90	98.90	70.04	97.81	**99.68**	95.93	97.79	97.85	71.02
	SAAC+PseAAC-S	**99.48**	**99.54**	**99.42**	**99.48**	**99.48**	70.57	97.43	99.57	95.29	97.41	97.48	71.00
	SAAC+PseAAC-P	99.25	99.42	99.08	99.25	99.25	70.56	97.75	98.93	96.57	97.74	97.78	69.95
	PseAAC-S+PseAAC-P	99.25	99.19	99.31	99.25	99.25	70.40	97.11	**99.68**	94.54	97.08	97.18	**71.27**

Bold values indicate the maximum value of the methods

the use of both individual-SVM and ensemble-SVM classifiers. Therefore, it can be concluded that, for breast/non-breast cancer classification, PseAAC-S space using Hd and Hb properties of amino acids carries the most discriminant information from input data.

Performance analysis of individual and ensemble-SVM classifier

In the training phase of SVM, several experiments were performed by adding diffuse samples in minority class datasets. In our case, dataset of breast cancer (BC) proteins represents the minority class. BC dataset was increased by a step size of 122. The majority class, i.e., non-breast cancer (NBC) proteins was fixed to 934. Figure 8 shows number of SVs produced during training of SVM with varying dataset size of minority class. From this figure, it can be observed that number of SVs is decreasing with increasing the number of proteins (diffuse points) for minority class. We found that number of SVs greater than 10 % of the original dataset would result in overfitting. To avoid overfitting, we have selected the size of dataset for minority class that produced approximately 10 % number of SVs.

To evaluate the performance of SVM classifier, several experiments were performed by varying the number of samples of minority class. Diffuse samples are added in the minority class datasets because it helps in avoiding overfitting of SVM. Figure 9 illustrates the performance of SVM classifier for breast cancer dataset. From Fig. 9, it is observed that Acc and Sp are not dependent on the size of the dataset. However, the performance of SVM, in terms of Sp, G-mean, F-score, and MCC is enhanced up to a certain limit.

Table 2 reports the performance of SVM based on individual and ensemble classifiers using different feature spaces. For cancer/non-cancer dataset, individual-SVM has

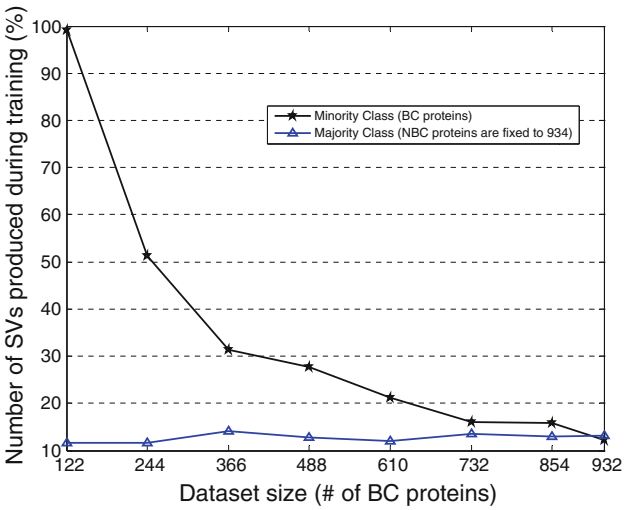

Fig. 8 Number of SVs produced during SVM training against the size of dataset for minority class

provided the highest Acc (96.71 %), G-mean (96.70 %), and F-score (96.73 %) using PseAAC-S feature space. However, PseAAC-P feature space has provided maximum values of Sn (97.69 %) and MCC (68.58 %). In this table, using the combined feature space of PseAAC-S+PseAAC-P, ensemble-SVM has given the highest Acc (97.63 %), Sn (95.38 %), Sp (99.88 %), G-mean (97.60 %), F-score (97.58 %), and MCC (68.47 %). However, when PseAAC-S feature space is combined at decision level with PseAAC-P, the performance of ensemble-SVM is enhanced up to 0.90 % (G-mean).

For breast/non-breast cancer, Table 2 highlights that individual-SVM classifier has given the highest values of Acc (95.18 %), Sn (93.04 %), G-mean (95.16 %), F-score (95.08 %), and MCC (62.80 %) for PseAAC-S feature space. Using combined feature spaces of PseAAC-S and

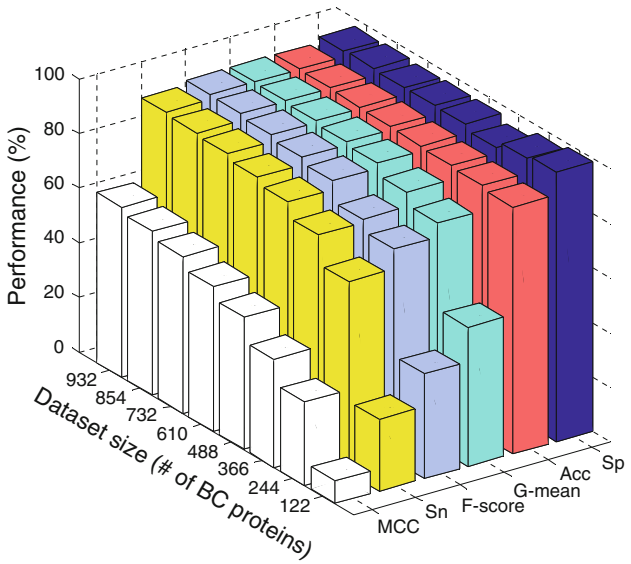

Fig. 9 Performance of individual-SVM against varying the number of samples of minority class

PseAAC-P, Ensemble-SVM has provided the highest values of Sp (99.79 %), and MCC (71.76 %). However, ensemble-SVM has given the values of Acc (96.95 %), Sp (94.22 %), G-mean (96.91 %) and F-score (97.03 %). It is observed that when predicated value of SVM classifier using PseAAC-S feature space is combined with predicated value of PseAAC-P feature space, at decision space, the performance of ensemble-SVM classifier is enhanced by 8.96 % in term of MCC. We inferred that combined feature spaces of PseAAC-S and PseAAC-P have provided better information with the use of both individual-SVM and ensemble-SVM classifiers. We thus concluded that these feature spaces possess the most discriminant information using Hd and Hb properties of amino acids for the classification of breast/non-breast cancer.

Performance analysis of individual and ensemble-KNN classifier

Table 3 shows the performance of KNN-based individual and ensemble classifiers using different feature spaces. For cancer/non-cancer dataset, it is observed that individual-KNN has provided the highest Acc (96.01 %), Sn (95.14), G-mean (96.01 %), and F-score (95.98 %) using PseAAC-S feature space. However, individual-KNN, for other feature spaces, has given Acc, Sn, G-mean and F-score nearly 94.74, 93.83, 94.73 and 94.69 %, respectively. From this table, ensemble-KNN using combined feature spaces AAC and SAAC has provided the highest values of 98.84, 99.07, 99.07 and 70.22 % for Sn, G-mean, F-score, and MCC, respectively.

Using breast/non-breast cancer dataset, Table 3 indicates that individual-KNN classifier has yielded the highest Acc (94.59 %), Sp (94.86 %), G-mean (94.59) %, and F-score (94.58 %) for PseAAC-P feature space. However, for PseAAC-S feature space, Sn and MCC measures have the highest values of 94.43 and 64.64 %, respectively. Ensemble-KNN, for AAC+SAAC features space, has provided the highest values of Acc (94.54 %), Sp (89.72 %), G-mean (94.42 %), and F-score (94.79 %). However, for PseAAC-S+PseAAC-P feature spaces, ensemble-KNN has given the highest values for Sn 99.89 % and MCC 73.19 %. It is noted that decision of ensemble-KNN is enhanced by 8.55 % (MCC), when the predicted values of KNN classifier in PseAAC-S feature space is combined with the predicted values in PseAAC-P feature space. The combined feature spaces of PseAAC hold the most useful information for breast/non-breast cancer dataset.

Overall performance comparison

Table 4 shows a comparative analysis, in term of accuracy, of individual classifiers RF, SVM, KNN, and Naïve Bayes

Table 2 Performance of SVM-based individual and ensemble classifiers using different feature spaces

Method	Feature space	Cancer/non-cancer						Breast/non-breast cancer					
		Acc	Sn	Sp	G_{mean}	F-score	MCC	Acc	Sn	Sp	G_{mean}	F-score	MCC
Individual-SVM	AAC	95.72	96.88	94.57	95.72	95.77	67.65	94.59	92.72	96.47	94.57	94.49	62.50
	SAAC	95.72	96.76	94.68	95.72	95.77	67.49	95.07	91.97	91.97	95.02	94.92	61.55
	PseAAC-S	**96.71**	97.57	**95.84**	**96.70**	**96.73**	68.35	**95.18**	**93.04**	97.32	**95.16**	**95.08**	**62.80**
	PseAAC-P	96.47	**97.69**	95.26	96.47	96.52	**68.58**	94.97	91.97	**97.97**	94.92	94.81	61.57
Ensemble-SVM	AAC+SAAC	96.71	93.76	99.65	96.66	96.61	67.80	96.57	99.68	93.47	96.52	96.68	71.46
	AAC+PseAAC-S	97.11	94.57	99.65	97.08	97.03	68.14	95.77	99.68	91.86	95.69	95.93	71.75
	AAC+PseAAC-P	97.23	94.68	99.77	97.19	97.15	68.18	96.15	99.25	93.04	96.10	96.26	70.96
	SAAC+PseAAC-S	97.11	94.45	99.77	97.07	97.03	68.08	96.31	99.57	93.04	96.25	96.42	71.39
	SAAC+PseAAC-P	97.23	94.57	99.88	97.19	97.15	68.12	**96.95**	99.68	**94.22**	**96.91**	**97.03**	71.33
	PseAAC-S+PseAAC-P	**97.63**	**95.38**	**99.88**	**97.60**	**97.58**	**68.47**	96.20	**99.79**	92.61	96.13	96.33	**71.76**

Bold values indicate the maximum value of the methods

Table 3 Performance of KNN-based individual and ensemble classifiers using different feature spaces

Method	Feature space	Cancer/non-cancer						Breast/non-breast cancer					
		Acc	Sn	Sp	G_{mean}	F-score	MCC	Acc	Sn	Sp	G_{mean}	F-score	MCC
Individual-KNN	AAC	94.74	93.29	96.18	94.73	94.66	63.18	93.36	93.58	93.15	93.36	93.38	63.74
	SAAC	93.99	94.34	93.64	93.99	94.01	64.62	92.56	92.61	92.51	92.56	92.56	62.63
	PseAAC-S	**96.01**	**95.14**	96.88	**96.01**	**95.98**	**65.28**	94.54	**94.43**	94.65	94.54	94.53	**64.64**
	PseAAC-P	95.49	93.87	**97.11**	95.48	95.42	63.77	**94.59**	94.33	**94.86**	**94.59**	**94.58**	64.50
Ensemble-KNN	AAC+SAAC	93.87	**98.84**	99.31	**99.07**	**99.07**	**70.22**	92.88	**99.89**	85.87	92.61	93.35	**73.19**
	AAC+PseAAC-S	94.22	88.67	99.77	94.06	93.88	65.94	94.00	99.79	88.22	93.83	94.33	72.58
	AAC+PseAAC-P	93.76	87.51	**100.0**	93.55	93.34	65.59	94.06	99.68	88.44	93.89	94.37	72.38
	SAAC+PseAAC-S	**94.68**	89.71	99.65	94.55	94.40	66.28	93.68	99.68	87.69	93.49	94.04	72.52
	SAAC+PseAAC-P	94.10	88.44	99.77	93.93	93.75	65.87	93.79	99.68	87.90	93.61	94.14	72.48
	PseAAC-S+PseAAC-P	94.57	89.36	99.77	94.42	94.27	66.16	**94.54**	99.36	**89.72**	**94.42**	**94.79**	71.69

Bold values indicate the maximum value of the methods

Table 4 Overall performance comparison of individual classifiers

Method	Feature space	Accuracy (%)	
		Cancer/non-cancer	Breast/non-breast cancer
Individual-RF	AAC	95.78	94.65
	SAAC	**96.76**	94.49
	PseAAC-S	95.85	**95.13**
	PseAAC-P	95.61	94.49
Individual-SVM	AAC	95.72	94.59
	SAAC	95.72	95.07
	PseAAC-S	**96.71**	**95.18**
	PseAAC-P	96.47	94.97
Individual-KNN	AAC	94.74	93.36
	SAAC	93.99	92.56
	PseAAC-S	**96.01**	94.54
	PseAAC-P	95.49	**94.59**
Individual-NB	AAC	90.75	93.84
	SAAC	88.96	93.47
	PseAAC-S	91.04	88.06
	PseAAC-P	**93.12**	**94.16**

Bold values indicate the maximum value of the methods

(NB) for cancer datasets. Form Table 4, we found that individual-RF performed better decision than the other individual classifiers for cancerous and non-cancerous related protein sequences. The highest accuracy given by individual-RF, individual-SVM, individual-KNN and individual-NB are 96.76, 96.71, 96.01, and 93.06 %, respectively. Individual-RF has yielded 3.7 % higher accuracy than that of NB. However, in case of breast/non-breast cancer dataset, individual-SVM has provided 0.05, 0.59, and 1.02 % higher accuracy than individual-RF, individual-KNN and individual-NB.

Figures 10 and 11 highlighted the overall performance comparison of individual and ensemble classifiers using the most informative feature spaces in terms of G-mean and F-score. For cancer/non-cancer dataset, Fig. 10 shows that individual-SVM in PseAAC-S feature space and ensemble-RF in SAAC+PseAAC-S feature spaces have performed better. For breast/non-breast cancer, Fig. 11 highlights that individual and ensemble of RF classifiers have performed better in PseAAC-S and AAC+SAAC feature spaces, respectively. Figure 12 demonstrates the overall performance comparison of ensemble classifiers for the highest performing feature spaces in terms of MCC measure. For cancer/non-cancer dataset, from Fig. 12, it is observed that ensemble-RF performs better than other ensemble classifiers using SAAC+PseAAC-S feature spaces. However, for breast/non-breast cancer dataset, ensemble-KNN is better than other ensemble classifiers using combined feature spaces of AAC+−SAAC. Therefore, the classification method developed using RF algorithm has shown good performance for the classification of breast/non-breast cancer. From Figs. 11 and 12, we observed that using the combined feature space of AAC+SAAC, ensemble-RF (in terms of G-mean and F-score) and ensemble-KNN (in terms of MCC) have provided better performance for breast/non-breast cancer dataset. Further, we observed that ensemble-RF has performed well in SAAC+PseAAC-S and AAC+SAAC feature spaces to classify cancer/non-cancer and breast/non-breast cancer datasets, respectively.

Performance comparison with previous studies

In Table 5, we carried out a performance comparison of the proposed IDM-PhyChm-Ens approach with previous approaches for breast cancer. Multiple sonographic and textural features based classifiers have provided accuracy

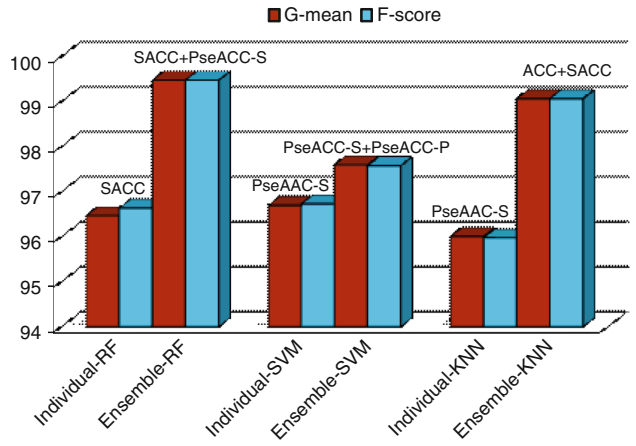

Fig. 10 Comparison of individual and ensemble classifiers based on RF, SVM, and KNN algorithms using the highest performing feature spaces in terms of G-mean and F-score for cancer/non-cancer dataset

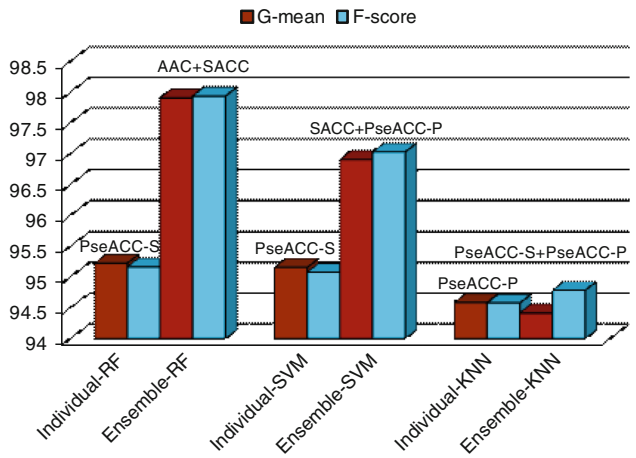

Fig. 11 Comparison of individual and ensemble classifiers based on RF, SVM, and KNN algorithms using the highest performing feature spaces in terms of G-mean and F-score for breast/non-breast cancer dataset

in the range 83.80–86.92 % (Liao et al. 2010). On the other hand, mammographic features using SVM classifier enhanced accuracy up to 93.73 % (Krishnan et al. 2010). Socio-demographic and other cancer-specific information-based features have classification accuracy in range 93.60–94.70 % (Eshlaghy et al. 2013).

Topological indices-based QPDR models have maximum accuracy of 90.5 % (Munteanu et al. 2009). Clinical features computed from a digitized image of FNA of a breast mass were also used for breast cancer classification along with various machine-learning approaches (Ster and Dobnikar 1996; Goodman et al. 2002). Optimized-LVQ, Big-LVQ, AIRS, and LDA models provided accuracy near to 96.80 % (Ster and Dobnikar 1996). The classification performance using clinical features enhanced the accuracy

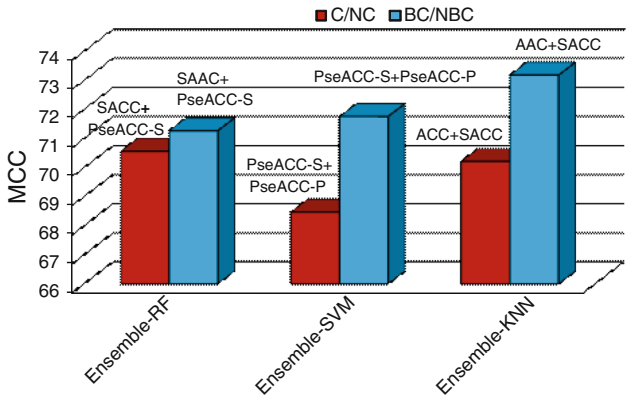

Fig. 12 Comparison of ensemble classifiers using the highest performing feature spaces in terms of MCC

in the range of 97.07–97.51 % for Fuzzy-GA, AR+NN, SVM+EAs, and SBS-BPPSO approaches.

On the other hand, our ensemble-RF using Hd and Hb properties of amino acid-based combined feature space SAAC+PseAAC-S has given the best classification accuracy of 99.48 % for cancer/non-cancer (CNC). Ensemble-RF using combined feature space AAC+SAAC has given the best classification accuracy of 97.91 % for breast/non-breast cancer (BNBC). Similarly, ensemble-SVM using feature spaces PseAAC-S+PseAAC-P and SAAC+PseAAC-P based on Hd and Hb properties of amino acids has yielded an improved accuracy of 97.63 and 96.95 % than previous SVM based approaches (Eshlaghy et al. 2013; Liao et al. 2010; Krishnan et al. 2010; Li et al. 2011; Bennett and Blue 1998; Ruxandra and Stoean 2013). Ensemble-NB using Hd and Hb properties of amino acids in the SAAC+PseAAC-S feature space has yielded the best classification accuracy of 98.32 % for CNC dataset. Similarly, ensemble-NB in AAC+SAAC feature space has yielded the best classification accuracy of 98.88 % for BC dataset.

Therefore, ensemble-RF and ensemble-NB yielded comparable performance for cancer datasets. This analysis shows that NB and RF approaches-based classification methodologies are quite effective for cancerous protein sequences prediction problem.

The performance of the proposed IDM-PhyChm-Ens method is impressive due to two reasons. The first reason is the use of descriptors derived from physicochemical properties of amino acids in protein primary sequences. These descriptors have a potential to accommodate the variation of amino acids composition in cancer and breast cancer proteins sequences with reference to non-cancer proteins. The second reason is the use of ensemble classifier, which combines the predictions of a specific learning algorithm at decision level using different feature spaces through the majority-voting.

Table 5 Comparison of classification accuracies achieved from our proposed ensemble-based classification method with other classifiers from literature

Method	Feature extraction strategy	Accuracy (BNBC) (%)	References
KNN	Sonographic and textural features	83.80	(Liao et al. 2010)
ANN		86.60	(Liao et al. 2010)
SVM		86.92	(Liao et al. 2010)
ANN	Incidence and population based features	91.20	(Dursun et al. 2005)
Decision tree		93.60	(Dursun et al. 2005)
Decision tree	Socio-demographic and cancer-specific information-based features	93.62	(Eshlaghy et al. 2013)
ANN		94.70	(Eshlaghy et al. 2013)
SVM		95.70	(Eshlaghy et al. 2013)
SVM	Mammographic features	93.73	(Krishnan et al. 2010)
Decision tree	Clinical features	94.74	(Quinlan 1996)
Optimized-LVQ	(10 features for each cell nucleus computed form a digitized image of a fine needle aspirate (FNA) of a breast mass)	96.70	(Goodman et al. 2002)
Big-LVQ		96.80	(Goodman et al. 2002)
AIRS		97.20	(Goodman et al. 2002)
LDA		96.80	(Ster and Dobnikar 1996)
SVM		97.20	(Bennett and Blue 1998)
Fuzzy-GA		97.36	(Pena-Reyes and Sipper 1999)
AR+NN		97.40	(Karabatak and Ince 2009)
SVM+EAs		97.07	(Ruxandra and Stoean 2013)
SBS-BPPSO	Entropy-Based selected clinical features	97.51	(Huang et al. 2010)
Fuzzy-SVM	Clinical features extracted with Principal Component Analysis	96.35	(Li et al. 2011)
Ensemble (NF KNN QC)	Information gain-based selected clinical features	97.14	(Sheau-Ling et al. 2012)

	CNC	BNBC	CNC	BNBC	
QPDR	*pTIe*[a] (embedded)	*TIe*+*dTIe*[a] (embedded)	90.0	91.80	(Munteanu et al. 2009)
Ensemble-RF	SAAC+PseAAC-S	AAC+SAAC	99.48	97.91	Present study
Ensemble-SVM	PseAAC-S+PseAAC-P	SAAC+PseAAC-P	97.63	96.95	Present study
Ensemble-KNN	SAAC+PseAAC-S	PseAAC-S+PseAAC-P	94.68	94.54	Present study
Ensemble-NB	SAAC+PseAAC-S	AAC+SAAC	98.32	98.88	Present study

[a] *dTIe* Difference between the same topological indices (TIs) and the average of the TIs for each type of cancer with embedded star graph, *pTIe* cancer probability TIs with embedded star graph, for more detail see (Munteanu et al. 2009)

Conclusions

We proposed a new and effective prediction method to distinguish between breast and non-breast cancer proteins. The proposed methodology exploits the physicochemical properties of Hd and Hb of amino acids and it extracts useful information from protein primary sequences in different feature spaces. In this study, we observed that proline, serine, tyrosine, cysteine, arginine, and asparagine amino acids offer high discrimination for cancer and healthy proteins using physicochemical properties.

We have developed new and ensemble classifiers using a specific learning algorithm such as RF, SVM, and KNN trained on different feature spaces. Our analysis demonstrates that RF-, SVM-, and KNN-based ensemble are more effective than their individual counterparts in different feature spaces. We observed that ensemble-RF have performed better than ensemble-SVM and ensemble-KNN. The proposed classification method has achieved an accuracy of 99.48 and 97.63 % for ensemble-RF and ensemble-SVM, respectively. Ensemble-RF has provided the highest F-score values of 99.48 and 97.94 % for cancer datasets. We observed that combined feature spaces of SAAC+PseAAC-S and AAC+SAAC show the best discrimination using ensemble-RF and ensemble-NB.

The comparative analysis highlights the improved performance of our proposed IDM-PhyChm-Ens method over existing approaches. This study indicates that the proposed methodology can be readily used for the development of clinical decision support system for breast cancer diagnosis.

Acknowledgments Authors are very grateful to Pakistan Institute of Engineering and Applied Sciences (PIEAS) for providing useful resources for this work.

Conflict of interest None.

References

American Cancer Society (2013) Cancer Facts & Figures. American Cancer Society Inc. http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-036845.pdf. Accessed 4 Aug 2013

Balmain A, Gray J et al (2003) The genetics and genomics of cancer. Nat Genet 33:238–244

Benediktsson JA, Swain PH (1992) Consensus theoretic classification methods. IEEE Trans Syst Man Cabernet 22:688–704

Bennett KP, Blue JA (1998) A support vector machine approach to decision trees. In: Neural networks proceedings. IEEE world congress on computational intelligence. The 1998 IEEE international joint conference, Anchorage, pp 2396–2401

Bing-Yu S, Zhu Z-H, Li J, Linghu B (2011) Combined feature selection and cancer prognosis using support vector machine regression. EEE/ACM Trans Comput Biol Bioinform 8(6):1671–1677

Bray F, McCarron P, Parkin DM (2004) The changing global patterns of female breast cancer incidence and mortality. Breast Cancer Res 6(6):229–239

Caroline D, Brasseur K, Leblanc V, Parent S, Asselin É, Bérubé G (2012) SAR study of tyrosine–chlorambucil hybrid regioisomers; synthesis and biological evaluation against breast cancer cell lines. Amino Acids 43(2):923–935

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(27):1–27

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357(1):116–121

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21(1):10–19

Chou KC, David WE (1999) Prediction of membrane protein types and subcellular locations. Proteins: Struct, Funct, Bioinf 34(1):137–153

Dobson PD, Cai YD, Stapley BJ, Doig AJ (2004) Prediction of protein function in the absence of significant sequence similarity. Curr Med Chem 11(16):2135–2142

Dursun D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 34(2):113–128

Džeroski S, Ženko B (2004) Is combining classifiers with stacking better than selecting the best one? Mach Learn 54:255–273

Einipour A (2011) A fuzzy-ACO method for detect breast cancer. Glob J Health Sci 3(2):195–199

Emmanuel M, Alvarez MM, Trevino V (2010) Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. Comput Biol Chem 34(4):244–250

Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, Ahmad LG (2013) Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform 4(2):124. doi:10.4172/2157-7420.1000124

Goodman DE, Boggess L, Watkins A (2002) Artificial immune system classification of multiple-class problems. In: Proceedings of the artificial neural networks in engineering 2002, pp 179–183

Hastie T, Tibshirani R, Friedman J (eds) (2001) The elements of statistical learning. Springer, New York

Hayat M, Khan A (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J Theor Biol 271:10–17

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Nat Acad Sci 78(6):3824–3828

Huang M-L, Hung Y-H, Chen W-Y (2010) Neural network classifier with entropy based feature selection on breast cancer diagnosis. J Med Syst 34(5):865–873

Jene-Sanz A, Váraljai R, Vilkova AV, Khramtsova GF, Khramtsov AI, Olopade OI, Lopez-Bigas N, Benevolenskaya EV (2013) Expression of polycomb targets predicts breast cancer prognosis. Mol Cell Biol 33(19):3951–3961

Ji-Yeon Y, Yoshihara K, Tanaka K, Hatae M, Masuzaki H, Itamochi H, Takano M, Ushijima K, Tanyi JL, Coukos G, Lu Y, Mills GB, Verhaak RGW (2013) Predicting time to ovarian carcinoma recurrence using protein markers. J Clin Investig 123(9):3740–3750

Karabatak M, Ince MC (2009) An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 36(2, Part 2):3465–3469

Khan A, Majid A, Tae-Sun C (2010) Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers. Amino Acids 38(1):347–350

Khan A, Majid A, Hayat M (2011) CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. Comput Biol Chem 35(4):218–229

Krishnan MMR, Banerjee S, Chakraborty C, Ray AK (2010) Statistical analysis of mammographic features and its classification using support vector machine. Expert Syst Appl 37:470–478. doi:10.1016/j.eswa.2009.05.045

Li DC, Wu CS, Tsai TI, Lina YS (2007) Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. Comput Oper Res 34:966–982

Li D-C, Liu C-W, Hu SC (2010) A learning method for the class imbalance problem with medical data sets. Comput Biol Med 40(5):509–518

Li DC, Liu CW, Hu SC (2011) A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artif Intell Med 52:45–52. doi:10.1016/j.artmed.2011.02.001

Liao R, Wan T, Qin Z (2010) Classification of benign and malignant breast tumors in ultrasound images based on multiple sonographic and textural features. In: Proceedings international conference on intelligent human-machine systems and cybernetics 2011 (IHMSC-2011). IEEE, Hangzhou, 26–27 Aug 2010, pp 71–74

Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252(2):350–356

Maqsood H, Khan A, Yeasin M (2012) Prediction of membrane proteins using split amino acid and ensemble classification. Amino Acids 42(6):2447–2460

Milenković J, Hertl K, Košir A, Žibert J, Tasič JF (2013) Characterization of spatiotemporal changes for the classification of dynamic contrast-enhanced magnetic-resonance breast lesions. Artif Intell Med 58(2):101–114

Mohabatkar H (2010) Prediction of cyclin proteins using Chous pseudo amino acid composition. Protein Pept Lett 17(10):1207

Muhammad T, Khan A, Majid A, Lumini A (2013) Subcellular localization using fluorescence imagery: utilizing ensemble classification with diverse feature extraction strategies and data balancing. Appl Soft Comput 13(11):4231–4243

Munteanu CR, Magalhães AL, Uriarte E, González-Díaz H (2009) Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J Theor Biol 257(2):303–311

Nasim FU, Ejaz S, Ashraf M, Asif AR, Oellerich M, Ahmad G, Malik GA, Attiq-ur-Rehman (2012) Potential biomarkers in the sera of breast cancer patients from Bahawalpur, Pakistan. Biomark Cancer 10(4):19–34

Pena-Reyes CA, Sipper M (1999) A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 17:131–155

Phang JM, Liu W (2012) Proline metabolism and cancer. Front Biosci: J Virtual Libr 17:1835

Pierrick C, Joseph AP, Poulain P, Brevern AGd, Rebehmed J (2013) Cis-trans isomerization of omega dihedrals in proteins. Amino Acids 45(2):279–289

Qiu JD, Huang JH, Shi SP, Liang RP (2010) Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept Lett 17(6):715–722

Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 4:77–90

Ramani RG, Jacob SG (2013a) Improved classification of lung cancer tumors Based on structural and physicochemical properties of proteins using data mining models. PLoS One 8(3):e58772. doi:10.1371/journal.pone.0058772

Ramani RG, Jacob SG (2013b) Prediction of cancer rescue p53 mutants in silico using Naïve Bayes learning methodology. Protein Pept Lett 20(11):1280–1891

Ramani RG, Jacob SG (2013c) Prediction of P53 mutants (multiple sites) transcriptional activity based on structural (2D&3D) properties. PLoS One 8(2):e55401

Richardson A (2011) Proline metabolism in metastatic breast cancer. http://cbcrp.org.127.seekdotnet.com/research/PageGrant.asp?grant_id=6922. Accessed 23 Sept 2013

Ruxandra S, Stoean C (2013) Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. Expert Syst Appl 40:2677–2686

Şahan S, Polat K, Kodaz H, Güneş S (2007) A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med 37(3):415–423

Sahu SS, Panda G (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. J Comput Biol Chem 34(5):320–327

Saima R, Hussain M, Ali A, Khana A (2013) A recent survey on colon cancer detection techniques. IEEE/ACM Trans Comput Biol Bioinform 10(3):545–563

Sheau-Ling H, Hsieh S-H, Cheng P-H, Chen C-H, Hsu K-P, Lee I-S, Wang Z, Lai F (2012) Design ensemble machine learning model for breast cancer diagnosis. J Med Syst 36(5):2841–2847

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314(5797):268–274

Ster B, Dobnikar A (1996) Neural networks in medical diagnosis: Comparison with other methods. In: Proceedings of the international conference on engineering applications of neural networks, pp 427–430

Tanford C (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 84(22):4240–4247

Vapnik VN (1995) The nature of statistical learning theory. Springer Verlag, New York

William CC (ed) (2010) An omics perspective on cancer research. Springer, Netherlands. ISBN: 978-90-481-2674-3

Xin M, Guo J, Liu H, Xie J, Sun X (2012) Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. IEEE/ACM Trans Comput Biol Bioinform 9(6):1766–1775

Xu R, Anagnostopoulos GC, Wunsch DC (2007) Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. IEEE/ACM Trans Comput Biol Bioinform 4(1):65–77

Yvan S, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517