Computers in Biology and Medicine 73 (2016) 38–46

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data

Safdar Ali, Abdul Majid *, Syed Gibran Javed, Mohsin Sattar

Department of Computer & Information Sciences, Pakistan Institute of Engineering & Applied Sciences (PIEAS), Nilore 45650, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 22 October 2015
Received in revised form
31 March 2016
Accepted 2 April 2016

Keywords:
Breast cancer
Tissue protein
Imbalanced data
Cost-sensitive classifier
GentleBoost ensemble

ABSTRACT

Early prediction of breast cancer is important for effective treatment and survival. We developed an effective Cost-Sensitive Classifier with GentleBoost Ensemble (Can-CSC-GBE) for the classification of breast cancer using protein amino acid features. In this work, first, discriminant information of the protein sequences related to breast tissue is extracted. Then, the physicochemical properties hydrophobicity and hydrophilicity of amino acids are employed to generate molecule descriptors in different feature spaces. For comparison, we obtained results by combining Cost-Sensitive learning with conventional ensemble of AdaBoostM1 and Bagging. The proposed Can-CSC-GBE system has effectively reduced the misclassification costs and thereby improved the overall classification performance. Our novel approach has highlighted promising results as compared to the state-of-the-art ensemble approaches.

1. Introduction

Worldwide, cancer is the second most fatal disease and breast cancer is the second main cause of cancer related deaths. The International Agency for Research on Cancer has reported for 2012 that 1,677,000 women were diagnosed with breast cancer and 577,000 women died with this disease [1]. In Pakistan, due to lack of treatment facilities approximately 17,552 women would die with breast cancer in 2015. Like other diseases, however, breast cancer can be successfully treated if diagnosed in the early stages [2–4]. For the early classification of breast cancer, development of an effective decision support system is a critical task.

Several conventional, statistical, and machine learning (ML) techniques are used for the detection/prediction of breast cancer. Mammography/imaging-diagnosis is one of the methods used for the detection of breast cancer; however, it has considerable variation in interpreting the results of graphs. For this reason, in recent times, many other statistical and ML approaches have been proposed for the prediction of cancers.

The fields of sequencing of human genome and proteins are well established. The growth of proteome data increased rapidly. For example, various international projects such as The Human Proteome Project and The Human Genome Project has been generated huge amount of data. This data is being used in various medical applications of diagnostic, prognostic, therapeutic, and preventive. The knowledge extracted from these projects increasingly promises the potential for future widespread adoption in personalized medicine and diagnostics. As the development in human genome and proteins has increased rapidly, so in recent times, researchers and practitioners would use protein data for clinic/research. In the view of fact, the ML based approaches (such as the proposed approach) will have very bright practical applications and benefits.

The sequencing of human genome and proteins are well established that are now utilized to find new cancer related molecular descriptors and biomarkers [5]. These markers are employed to develop improved decision support systems, since the mutated proteins are associated with breast cancer. Therefore, features extracted from such protein disorders would be used for cancer classification, treatment, and drug discovery. Recently, protein sequences are employed for the classification of ovarian cancer [6], lung cancer [7], colon cancer, and breast cancer [8]. Xin et al. used sequence-based features for the prediction of DNA binding residues in protein sequences [9].

As cancer is a genetic disease, investigating various kinds of mutations in genetic material such as DNA, RNA, and proteins, can reveal important underpinnings of carcinogenesis and cancer proliferation. In this paper, for classification of breast cancer, we have utilized the discriminant information of mutated protein

* Corresponding author.
E-mail addresses: safdarali_11@pieas.edu.pk (S. Ali), abdulmajiid@pieas.edu.pk (A. Majid), gibranjaved_11@pieas.edu.pk (S.G. Javed), mohsin_14@pieas.edu.pk (M. Sattar).

http://dx.doi.org/10.1016/j.compbiomed.2016.04.002
0010-4825/

molecules with physicochemical properties hydrophobicity (H_b) and hydrophilicity (H_p) of amino acids. These physicochemical properties exhibit excellent discriminant capability in different feature spaces for amino acid sequences classification [10].

Usually, in cancerous data, number of cancer and non-cancer patients is inherently imbalanced i.e., the particular diagnosis class is not easily achievable. Thus, the decision boundary of conventional classifier is biased towards majority class. To address this problem, different techniques are suggested; either processing the input data or use the cost sensitive learning (CSL). In this study, we have employed CSL technique, which minimized the misclassification costs.

The over-sampling technique is used to put more novel data examples to the rare class to balance dataset. The new data examples are produced using synthetic techniques or by duplicating the data examples of the minority class. Synthetic minority oversampling technique (SMOTE) adds new synthetic samples to the minority class by randomly interpolating pairs of the closest neighbors [11]. Usually, SMOTE engages in making copies of samples and consequently lead to overfitting [12]. For small medical dataset, Mega-Trend diffusion method adds new synthetic samples to the minority class by employing membership function rather than normal distribution to compute the possibility values of synthetic [13]. On the other hand, under-sampling can discard potentially useful medical and biological information of the majority data class that could be important for the induction process. For example, given imbalance ratio of 100:5, in order to get a close match for the minority class, it might be undesirable to throw away 95% of majority class instances. Therefore, to avoid the risk of deleting useful information form majority data class and to prevent overfitting in case of over-sampling, we preferred to use cost-sensitive learning (CSL) approach.

Previously, Wang et al. [14] constructed five-year breast prognosis models by combining Logistic Regression (LR) and Decision Tree (DT) with cost-sensitive classifier (CSC) technique, Bagging, and Boosting. Their proposed CSC ensemble models showed improved performance than the original models. They reported accuracy up to 91.30%. Liu et al. have applied under-sampling approach with DT to deal with imbalanced problem for breast cancer survivability and reported 86.52% survival rate of patients [15]. Zhang et al. utilized gene expression profiles for the prediction of breast cancer by employing LR, Support Vector Machine (SVM), AdaBoost, LogitBoost and Random Forest (RF) [16]. They achieved maximum value of Area Under the receiver operating characteristic Curve (AUC) measure of 88.6% and 89.9% for SVM and RF models, respectively. Delen et al. used surveillance and epidemiology results for prediction of breast cancer [17]. They employed three classification models of DT, LR, and Artificial Neural Network (ANN) based learning approaches. They obtained the highest value AUC of 84.9% and 76.9% for LR and DT models, respectively. In another study, Khalilia et al. developed prediction models from highly imbalanced data using SVM, Bagging, Boosting and RF [18]. They demonstrated that, in terms of AUC measure, RF model (91.2%) outperformed SVM (90.6%), Bagging (90.5%), and Boosting (88.9%).

It is a challenging task for an individual learner to develop an improved prediction models for breast cancer [4,19]. Therefore, Boosting and Bagging based ensemble systems were developed for cancer dataset. These ensemble systems were constructed by a set of trained classifiers with the same learning classifier. They attempt to enhance the performance by iteratively retraining the base classifiers with a subset of most informative data and then combining their predictions with novel examples. These systems have limited performance due to small number of samples and class imbalance. The main novelty in this study is the development of CSL based GentleBoost ensembles (Can-CSC-GBE) using physicochemical properties H_b and H_p of amino acids as molecule descriptors in different feature spaces. To the best of our knowledge, previously, this aspect has not been explored for imbalanced data in the context of breast cancer classification.

In the proposed study, first, molecule descriptors are generated in four different feature spaces of (i) Amino Acid Composition (AAC) of dimensions 20, (ii) Split Amino Acid Composition (SAAC) of dimensions 60, (iii) Pseudo Amino Acid Composition-Series (PseAAC-S) of dimensions 40, and (iv) Pseudo Amino Acid Composition-Parallel (PseAAC-P) of dimensions 60. The CSL technique is then employed in feature spaces to reduce the misclassification costs. In the next step, we employed GentleBoost ensemble to construct Can-CSC-GBE system (GBE_{FS}^{CSC}) for different feature spaces (FS) using 10-fold jackknife technique. For comparison purpose, ensemble system of AdaBoostM1, and Bagging are implemented using CSC technique to develop $AdaM1_{FS}^{CSC}$ and Bag_{FS}^{CSC} models. The experimental results demonstrate that $GBE_{PseAAC-S}^{CSC}$ model using PseAAC-S feature space is superior to individual, $AdaM1_{FS}^{CSC}$, and Bag_{FS}^{CSC} models.

2. Material and methods

Framework of the proposed CSC based GentleBoost ensemble cancer classification is shown in Fig. 1. The proposed system consists of three main modules: the feature space, the CSC development, and the ensemble development. The proposed system is assessed using two datasets of protein amino acid sequences for cancer/non-cancer (C/NC) and breast/non-breast cancer (B/NBC). These datasets are borrowed from [8]. These datasets consist of 1056 protein sequences. First C/NC dataset is composed of 865 non-cancer and 191 cancerous protein sequences, whereas the second B/NBC dataset is containing 865 non-cancer and 122 breast-cancer related protein sequences. The next subsection describes the feature space generation of protein primary sequences.

2.1. Feature generation

Proper input representation of protein primary sequences make easier for a classifier to recognize underlying regularities in the sequences. The native twenty amino acids in a protein sequence are usually illustrated by set of single letter codes of English letters. A protein of length L_r is formally represented as an ordered sequence $p = (a_1, a_2, \ldots, a_L)$ with elements a_i, from the finite set ={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, where A stand for Alanine, C stand for Cysteine, E stand for Glutamic acid, etc. The raw information given by the protein sequence is customarily restructured for prediction such that each representation of protein sequences is suitable for a feature space. For cancer prediction, we used correlation based discriminant feature extraction strategies of AAC, SAAC, PseAAC-S, and PseAAC-P.

In AAC feature space (20-dimensions), a vector of the relative frequencies of the 20 native amino acids represents each protein in its sequence as:

$$f_i = \frac{n_i}{L_r} \qquad (i = 1, 2, \ldots, 20) \tag{1}$$

where f_i represents the occurrence frequency of the i-th native amino acid in the protein, n_i is the number of the i-th native amino acid in sequence. Then, the AAC feature vector is expressed as:

$$\mathbf{x}_{AAC} = [f_1, f_2, \ldots, f_{20}]^T \tag{2}$$

In SAAC feature space generation, the given protein sequence is split into three dissimilar sections, named the N-terminal, the Internal segments and the C-terminal [20,21]. The amino acid

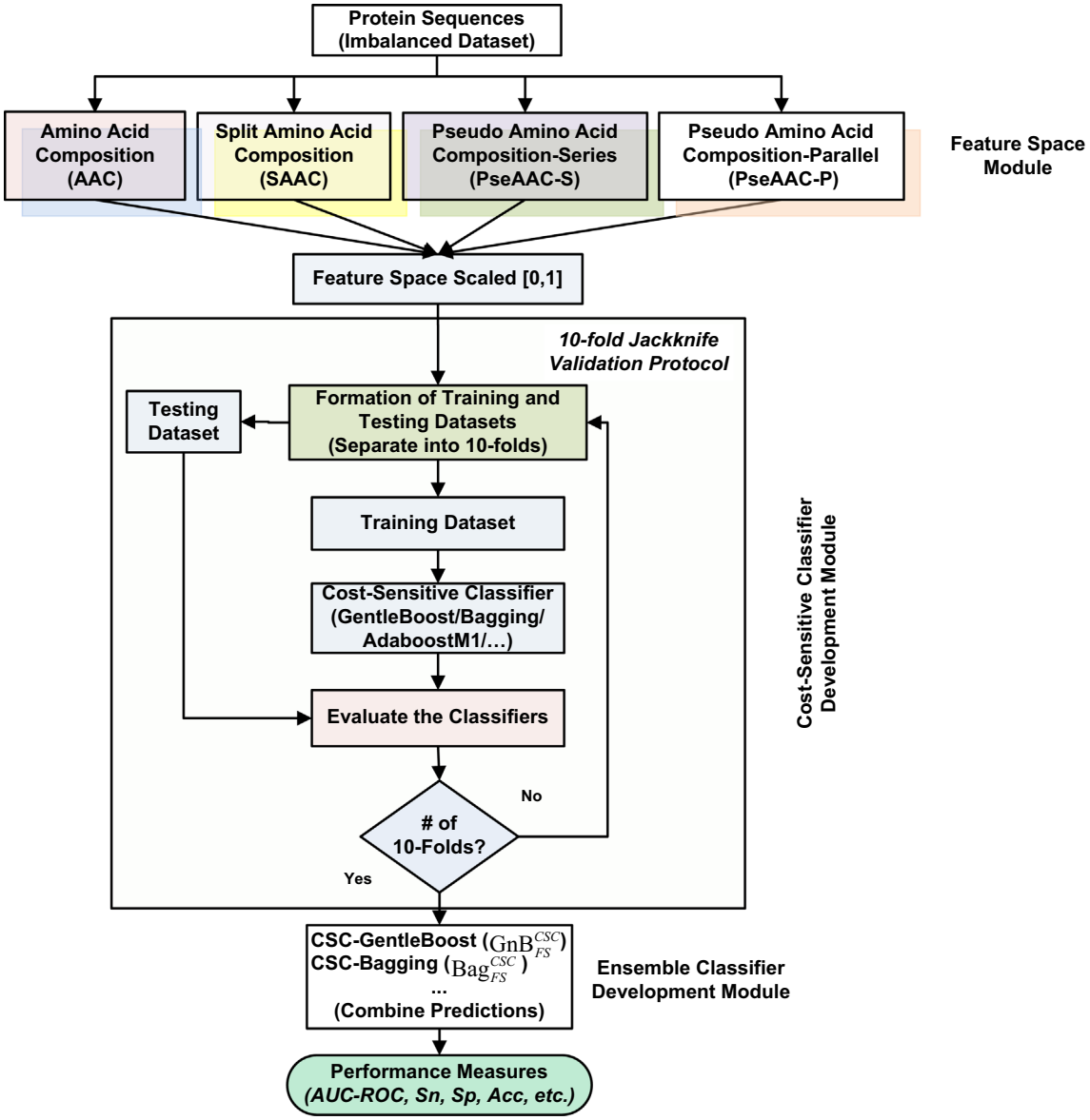

Fig. 1. Framework of the proposed CSC-GBE system for cancer classification.

compositions of these three parts are calculated individually using Eqs. (1) and (2). In this study, the length of each part is equal to 20. Consequently, the dimension of each feature vector in SAAC model is 60. The SAAC feature vector is given as:

$$\mathbf{x}_{SAAC} = \left[nt_1, \ldots, nt_{20}, i_1, \ldots, i_{20}, c_1, \ldots, c_{20} \right]^T \quad (3)$$

where $\mathbf{N} = [nt_1, \ldots, nt_{20}]^T$, $\mathbf{I} = [i_1, \ldots, i_{20}]^T$, and $\mathbf{C} = [c_1, \ldots, c_{20}]^T$ are the N-terminal, the Internal segments and the C-terminal feature vectors of a protein, respectively.

The PseAAC based feature spaces have capability to carry core and essential information concealed in complicated protein sequences without losing its sequence-order [13,22]. We generated a protein feature vector in series correlation (PseAAC-S) using $20 + i \times \lambda$ discrete components, where 'i' is the selected number of amino acid properties. The protein vector in series correlation of dimensions 60, with $\lambda = 20$ and $i = 2$ for hydrophobic and hydrophilic values, is expressed as:

$$\mathbf{x}_{PseAAC-S} = \left[p_1 \cdots p_{20} p_{21} \cdots p_{20+\lambda}\, p_{20+\lambda+1} \cdots p_{20+2\lambda} \right]^T \quad (\lambda < L_r) \quad (4)$$

with

$$p_u = \begin{cases} \dfrac{g_u}{\sum_{i=1}^{20} g_i + w \sum_{j=1}^{2\lambda} \tau_j} & \text{for} \quad 1 < \mu < 20 \\ \dfrac{w\tau_u}{\sum_{i=1}^{20} g_i + w \sum_{j=1}^{2\lambda} \tau_j} & \text{for} \quad 21 < \mu < 20 + 2\lambda \end{cases} \quad (5)$$

where, $g_i (i = 1, 2, \ldots, 20)$ be the normalized occurrence frequencies of 20 native amino acids in the protein and τ_j the jth-tier sequence-correlation factor is calculated according to $\tau_{2\lambda-1} = \frac{1}{L_r - \lambda} \sum_{i=1}^{L_r-\lambda} H_{b_{i,i+\lambda}}$ and $\tau_{2\lambda} = \frac{1}{L_r - \lambda} \sum_{i=1}^{L_r-\lambda} H_{p_{i,i+\lambda}}$, L_r is the amino acid residues. Here, weighting factor w is empirically set equal to 0.05.

The protein feature vector in parallel correlation (PseAAC-P) is given by $20+\lambda$ discrete components as:

$$\mathbf{x}_{PseAAC-P} = \left[p_1 \cdots p_{20} p_{20+1} \cdots p_{20+\lambda} \right]^T \quad (6)$$

with

$$p_u = \begin{cases} \dfrac{g_u}{\sum_{i=1}^{20} g_i + \omega \sum_{j=1}^{\lambda} \theta_j} & \text{for} \quad 1 \le \mu \le 20 \\ \dfrac{\omega \theta_{u-20}}{\sum_{i=1}^{20} g_i + \omega \sum_{j=1}^{\lambda} \theta_j} & \text{for} \quad 21 \le \mu \le 20+\lambda \end{cases} \tag{7}$$

where, θ_λ be the λ-tier correlation factor that reveals the sequence order correlation between all the λ for most contiguous residues along a protein chain. It is determined using the following equation.

$$\theta_\lambda = \frac{1}{L_r - \lambda} \sum_{i=1}^{L_r-\lambda} \Theta(R_i, R_{i+\lambda}) \tag{8}$$

Thus, for $\lambda=20$, we created a feature vector of 40 dimensions. The value of $\Theta(R_i, R_j)$ is computed as follow:

$$\Theta(R_i, R_j) = \frac{1}{2}\left\{ \left[H_b(R_j) - H_b(R_i)\right]^2 + \left[H_p(R_j) - H_p(R_i)\right]^2 \right\} \tag{9}$$

where, H_b and H_p are hydrophobicity and hydrophilicity of i-th amino acid R_i and j-th amino acid R_j, respectively.

2.2. Handling imbalanced data with CSL technique

Generally, the intelligent computational approaches perform well on balanced datasets. However, if algorithms are implemented with imbalanced data, they are overwhelmed by examples in the majority class, and examples in the minority class are ignored, resulting in high accuracy for the majority class but poor accuracy for the minority class. We handled this imbalanced data problem with CSL technique, which impose higher cost to the misclassification examples. The aim is to develop an effective and reliable model with the least misclassification costs. Table 1 shows the detail information of cost matrix.

In Table 1, True Positive (TP) and True Negative (TN) represent the correct classifications; the cost of False Negative (FN) indicates misclassifying of actual positive example (minority class) as a negative (majority class), and False Positive (FP) denotes misclassifying of actual negative instance (majority class) as a positive (minority class). The $C(i, j)$ indicates the cost of misclassifying of example from actual class i as predicted class j. For our binary classification problem, we denote majority class by '0' and minority class by '1'

A classifier should classify an example $\mathbf{x}$ into the class i, with the minimum expected cost. This expected cost $R(i/\mathbf{x})$ for a given cost matrix is given as

$$R(i/\mathbf{x}) = \sum_j P(j/\mathbf{x}) C(i, j) \tag{10}$$

where $P(j/\mathbf{x})$ denotes the posterior probability of classifying an example $\mathbf{x}$ into class j. We consider no cost for accurate classifications. The CSL is implemented by predicting the class with the minimum expected misclassification cost using the values in the cost matrix. The next section explains development of the proposed ensemble classifier.

2.3. Development of ensemble classifier

In computational intelligent approaches, an ensemble classifier combines the preliminary predictions of several individually trained classifiers (base-learners) with a suitable combining scheme. Previous studies have revealed that often the performance of an ensemble is better than any of the base-learners. Currently, several ensembles approaches have been proposed for real life problems. The most popular approaches for generating ensembles are Bagging [23] and Boosting [24,25]. The most popular way of generating of an ensemble is integrating weak decision trees into strong model. In this paper, we use GentleBoost classification algorithm, a variant of Boosting algorithm, for breast cancer prediction. GentleBoost minimizes the exponential loss and is implemented using DT as base learners. Detailed information on GentleBoost ensemble algorithm is available in [26].

Table 1
Cost matrix for binary problem.

		Predicted class	
		Minority	**Majority**
Actual class	**Minority**	C(1,1) or TP	C(1,0) or FN
	Majority	C(0,1) or FP	C(0,0) or TN

For the development of cost-sensitive classifier, GentleBoost ensemble is trained using N samples in training dataset, $S_t = \{\mathbf{x}^{(n)}, t^{(n)}\}_{n=1}^{N}$, where $\mathbf{x}^{(n)}$ represents the nth feature vector from feature spaces (FS) of $\{FS_1, FS_2, \ldots, FS_m\}$ that are extracted using protein sequences correspond to target $t^{(n)}$. Here, $Fs_1, Fs_2, \ldots, Fs_m$ are m different feature spaces. In this work, we develop GentleBoost ensemble using the best number of learners found (see Fig. 2; 78 and 34 numbers of learners for C/NC and B/NBC datasets, respectively). The proposed approach works by: learning a base-

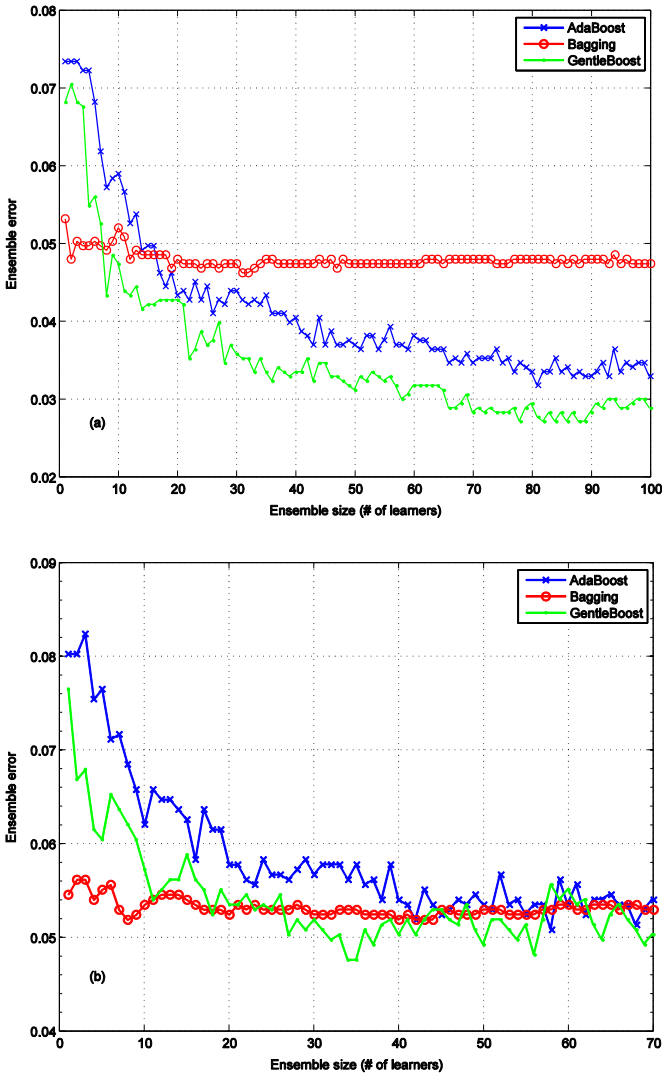

Fig. 2. Ensemble error as a function of ensemble size using PseAAC-S feature space for (a) C/NC and (b) B/NBC datasets.

Table 2
Performance of conventional ensembles for imbalanced datasets.

Approach	C/NC dataset					B/NBC dataset				
	AUC (%)	Acc (%)	Sn (%)	Sp (%)	G_{mean} (%)	AUC (%)	Acc (%)	Sn (%)	Sp (%)	G_{mean} (%)
$AdaM1_{AAC}$	87.73 ± 2.33	87.78 ± 2.62	53.93 ± 2.40	95.26 ± 2.13	71.67 ± 3.51	76.84 ± 2.73	88.26 ± 2.90	9.84 ± 4.77	98.50 ± 1.97	31.13 ± 2.97
$AdaM1_{SAAC}$	95.16 ± 2.45	92.80 ± 2.38	75.39 ± 3.78	96.65 ± 2.59	85.36 ± 2.66	86.52 ± 2.94	88.54 ± 2.81	15.57 ± 5.07	98.01 ± 2.11	39.08 ± 5.82
$AdaM1_{PseAAC-S}$	**97.14** ± 1.29	**94.03** ± 2.69	81.68 ± 2.93	96.76 ± 2.61	88.90 ± 2.23	**93.08** ± 2.67	**91.57** ± 2.89	59.84 ± 4.64	95.72 ± 2.10	75.68 ± 3.86
$AdaM1_{PseAAC-P}$	93.07 ± 2.37	91.29 ± 2.54	68.06 ± 2.46	96.42 ± 2.90	81.01 ± 2.47	88.81 ± 2.93	89.20 ± 2.74	13.11 ± 5.65	99.1 ± 2.90	36.06 ± 3.99
Bag_{AAC}	91.54 ± 2.85	89.30 ± 2.91	60.73 ± 2.57	95.61 ± 2.87	76.20 ± 2.98	89.23 ± 2.87	89.87 ± 2.54	40.98 ± 3.94	96.25 ± 2.56	62.81 ± 3.95
Bag_{SAAC}	92.73 ± 2.89	90.81 ± 2.61	69.63 ± 2.85	95.49 ± 2.55	81.54 ± 2.69	90.36 ± 2.64	**91.38** ± 2.65	54.92 ± 3.63	96.15 ± 2.48	72.66 ± 3.86
$Bag_{PseAAC-S}$	**95.78** ± 2.84	**93.09** ± 2.23	81.68 ± 2.68	95.61 ± 2.62	88.37 ± 2.25	**92.37** ± 2.82	90.63 ± 2.70	59.84 ± 3.85	94.65 ± 2.26	75.25 ± 3.13
$Bag_{PseAAC-P}$	93.96 ± 2.92	90.81 ± 2.83	65.97 ± 3.91	96.30 ± 2.70	79.70 ± 2.83	91.52 ± 2.59	90.63 ± 2.56	46.72 ± 3.87	96.36 ± 2.29	67.10 ± 3.43
GBE_{AAC}	89.94 ± 2.99	88.73 ± 2.81	59.69 ± 3.78	95.14 ± 2.91	75.36 ± 2.94	76.52 ± 2.88	88.73 ± 2.38	9.84 ± 4.72	99.04 ± 1.98	31.21 ± 4.61
GBE_{SAAC}	96.28 ± 2.32	93.84 ± 2.69	79.06 ± 2.83	97.11 ± 2.58	87.62 ± 2.86	90.89 ± 2.92	90.06 ± 2.49	32.79 ± 4.13	97.54 ± 2.15	56.55 ± 3.57
$GBE_{PseAAC-S}$	**97.30** ± 1.27	**94.22** ± 2.10	80.63 ± 2.87	97.23 ± 1.86	88.54 ± 2.15	**92.52** ± 2.67	**92.14** ± 2.14	59.84 ± 3.17	96.36 ± 2.02	75.93 ± 2.97
$GBE_{PseAAC-P}$	93.4 ± 2.77	90.91 ± 2.97	68.06 ± 2.88	95.95 ± 2.61	80.81 ± 2.17	89.83 ± 2.81	89.02 ± 2.51	33.61 ± 3.34	96.25 ± 2.34	56.87 ± 3.89

Table 3
Prediction performance of CSC – ensembles in different feature spaces.

Approach	C/NC dataset				B/NBC dataset			
	Acc (%)	Sn (%)	Sp (%)	G_{mean} (%)	Acc (%)	Sn (%)	Sp (%)	G_{mean} (%)
$AdaM1^{CSC}_{AAC}$	83.61 ± 1.67	79.58 ± 2.14	84.51 ± 2.41	82.01 ± 2.78	81.82 ± 2.87	70.49 ± 2.89	83.30 ± 2.66	76.63 ± 2.56
$AdaM1^{CSC}_{SAAC}$	87.97 ± 2.73	89.01 ± 2.05	87.75 ± 2.11	88.37 ± 2.83	87.78 ± 2.72	81.15 ± 2.77	88.65 ± 2.83	84.82 ± 2.74
$AdaM1^{CSC}_{PseAAC-S}$	**93.56** ± 1.22	88.48 ± 1.79	94.68 ± 2.01	91.53 ± 2.54	**89.02** ± 2.67	79.51 ± 2.75	90.26 ± 2.34	84.71 ± 2.48
$AdaM1^{CSC}_{PseAAC-P}$	86.93 ± 2.17	81.68 ± 2.85	88.09 ± 2.53	84.82 ± 2.87	87.31 ± 2.19	67.21 ± 2.95	89.94 ± 2.65	77.75 ± 3.41
Bag^{CSC}_{AAC}	83.62 ± 2.26	85.34 ± 2.71	83.24 ± 2.86	84.28 ± 2.44	80.30 ± 2.99	82.79 ± 2.71	79.98 ± 2.79	81.37 ± 2.56
Bag^{CSC}_{SAAC}	86.84 ± 1.91	88.48 ± 2.41	86.47 ± 2.54	87.47 ± 2.63	83.24 ± 2.69	84.43 ± 2.56	83.08 ± 2.84	83.75 ± 2.36
$Bag^{CSC}_{PseAAC-S}$	**88.35** ± 2.56	95.29 ± 1.99	86.82 ± 2.36	90.96 ± 2.45	**84.47** ± 2.55	93.44 ± 2.06	83.30 ± 2.97	88.22 ± 2.79
$Bag^{CSC}_{PseAAC-P}$	86.08 ± 2.69	88.48 ± 2.45	85.55 ± 2.45	87.00 ± 2.67	83.62 ± 2.74	86.07 ± 2.90	83.30 ± 2.85	84.67 ± 2.91
GBE^{CSC}_{AAC}	86.27 ± 2.01	82.20 ± 2.98	87.17 ± 2.70	84.65 ± 2.19	85.51 ± 2.81	68.03 ± 2.87	87.79 ± 2.65	77.28 ± 2.96
GBE^{CSC}_{SAAC}	90.53 ± 1.99	86.39 ± 2.07	91.45 ± 2.01	88.88 ± 2.77	89.87 ± 2.46	83.61 ± 2.87	90.69 ± 2.60	87.07 ± 2.83
$GBE^{CSC}_{PseAAC-S}$	**94.41** ± 1.13	85.34 ± 2.56	96.42 ± 2.15	90.71 ± 2.82	**90.15** ± 2.01	72.13 ± 3.45	92.51 ± 2.23	81.69 ± 2.72
$GBE^{CSC}_{PseAAC-P}$	90.34 ± 1.63	81.68 ± 2.74	92.25 ± 2.37	86.80 ± 2.67	88.64 ± 2.44	65.57 ± 3.73	91.65 ± 2.12	77.52 ± 2.97

learner (DT) on training set; computing class probability of every training example by the fraction of weights that receives from the ensemble; using Eq. (1) to consistent each training example with the computed optimal class; and reapplying the classifier to the modified (relabeled) training set. In this way, a classifier is found that has a minimum overall risk based on Eq. (1). For every learner with index t, the GentleBoost estimates the mean-squared error for the ensemble model as:

$$\sum_{n=1}^{N} d_t^{(n)}(\tilde{y}^{(n)} - h_t(X^{(n)}))^2 \tag{11}$$

where $d_t^{(n)}$ represent observation weights at step t, $h_t(X^{(n)})$ specify the predictions of the learner with index t, i.e., model of a weak learner in the ensemble trained on N observations with predictors $X^{(n)}$, fitted to response values $\tilde{y}^{(n)}$.

3. The proposed experimental framework

During training, we employ the cost parameter when classes are asymmetrically presented to classifier. In current study, we want to classify cancer and non-cancer patients using protein sequences. Failure to classify a cancer (false negative) has much more serious consequences than misclassifying non-cancer as cancer (false positive). With this assumption, we should assign high cost to misclassifying cancer as non-cancer and low cost to misclassifying non-cancer as cancer. In the training, we forward misclassification costs in the form of a nonnegative square matrix. For example, if we have faith that the first error is three times worse than the second is. We construct another cost matrix that reflects this faith. The element C(i,j) of cost matrix represents the cost of classifying an example into class j if the true class is i. The diagonal elements C(i,i) of the cost matrix must be 0. For this study, we choose cancer to be class 1 and non-cancer to be class 0. Then we set the cost matrix to where C(i,j) > 1 is the cost of misclassification a cancer as non-cancer. This indicates that the maximum costs are for misclassifying a minority class as a frequent one, and inversely for the minimum. In each simulation run, distinct cost metrics are constructed.

In training phase, CSL technique computes weights of each training example in accordance with the cost assigned to each class. The misclassification cost is assigned to train model as proposed in [14]; explicitly, the cost of misclassifying a cancer patient as non-cancer equals the imbalance ratio, which is the ratio of majority to minority class. Hence, the cost of misclassifying a cancer patient as non-cancer is equal to C(1,0)=4.53 and C(1,0)=7.66 for C/NC and B/NBC datasets, respectively. The cost of misclassifying a non-cancer as cancer patient is given to C(0,1)=1. The cost of correct classification in each class is set to 0, i.e., C(1,1)=C(0,0)=0. For comparison purpose, we also performed experiments in different feature spaces to develop gentle ensemble without CSL technique.

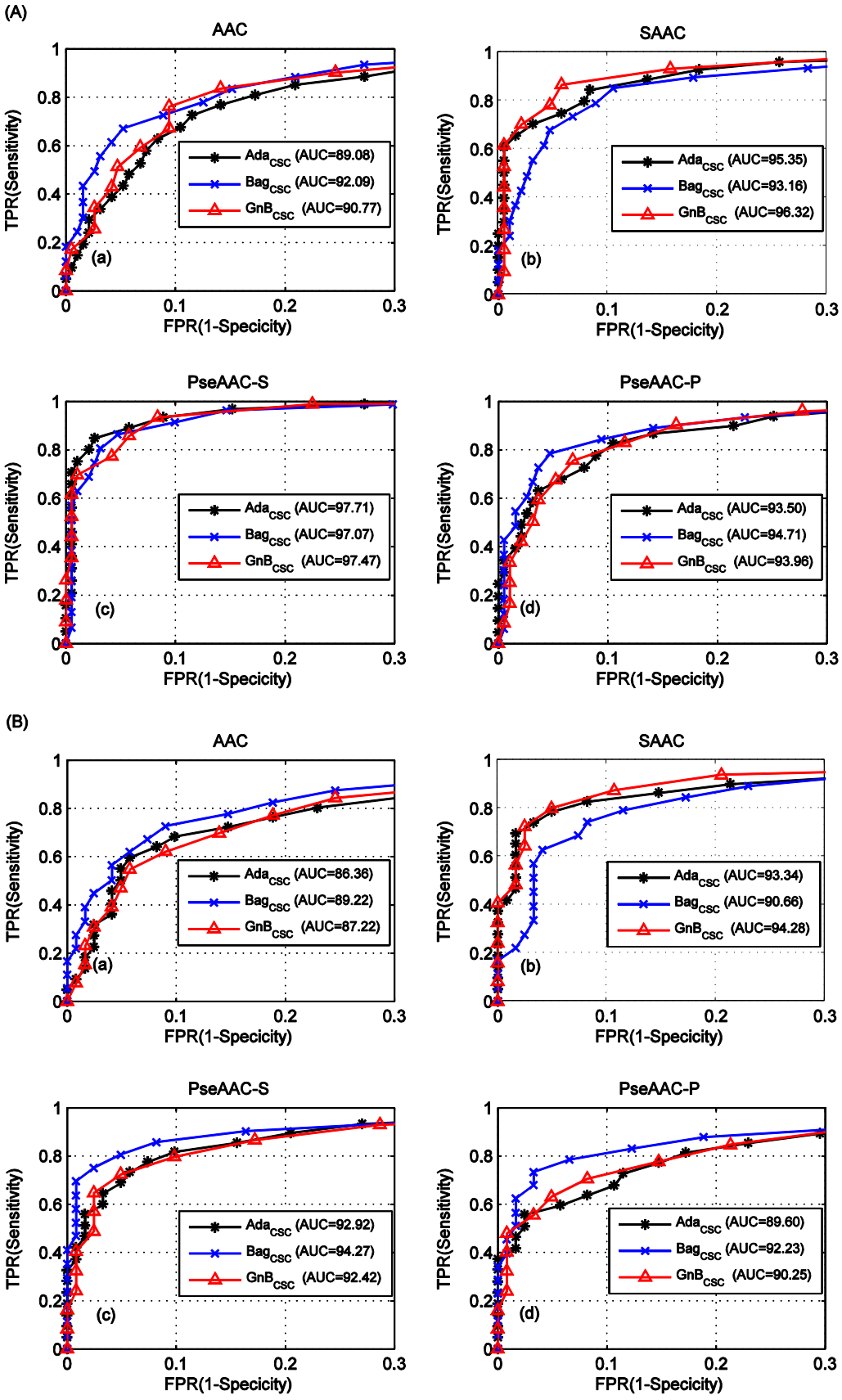

Fig. 3. (A-B) ROC curves of CSC-AdaBoostM1, CSC-Bagging, and CSC-GentleBoost approaches for (A) C/NC and (B) B/BNC datasets using: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P spaces. Note that for better visualization of region of interest, we have plotted partial ROC curves.

Table 4
Performance comparison of the proposed approach with state-of-the-art approaches.

Approach		C/NC dataset				B/NBC dataset			
		AUC (%)	Acc (%)	Sn (%)	Sp (%)	AUC (%)	Acc (%)	Sn (%)	Sp (%)
Quantitative proteome-disease relationship QPDR [8]		NA*	90.50	NA	NA	NA	91.80	NA	NA
Over-sampling	$KNN_{PseAAC\text{-}S}$ [13]	87.77 ± 2.75	96.01 ± 2.36	95.14 ± 2.57	96.88 ± 2.36	81.89 ± 3.37	94.54 ± 2.98	94.43 ± 2.84	94.65 ± 2.91
	SVM_{SAAC} [13]	90.00 ± 2.85	96.71 ± 2.78	97.57 ± 2.54	95.84 ± 2.65	90.00 ± 2.91	95.18 ± 2.63	93.04 ± 2.28	97.32 ± 2.56
	Present study								
Under-sampling	$RUSBoost_{PseAAC\text{-}S}$	93.05 ± 2.63	86.35 ± 2.85	88.03 ± 2.27	86.20 ± 3.64	90.31 ± 2.87	82.02 ± 3.49	83.64 ± 2.76	81.39 ± 3.18
	$RUSBoost_{PseAAC\text{-}P}$	85.29 ± 3.57	73.53 ± 2.98	78.53 ± 3.73	72.43 ± 3.93	82.10 ± 2.95	74.17 ± 3.91	76.31 ± 3.77	73.50 ± 3.98
Without CSL	$GBE_{PseAAC-S}$	97.30 ± 1.53	94.22 ± 2.10	80.63 ± 2.87	97.23 ± 1.86	92.52 ± 2.68	92.14 ± 2.14	59.84 ± 3.17	96.36 ± 2.02
With CSL	$GBE^{CSC}_{PseAAC-S}$	97.47 ± 1.89	94.41 ± 1.13	85.34 ± 2.56	96.42 ± 2.15	92.42 ± 2.81	90.15 ± 2.01	72.13 ± 3.45	92.51 ± 2.23
	$Bag^{CSC}_{PseAAC-S}$	97.07 ± 2.01	88.35 ± 2.56	95.29 ± 1.99	86.82 ± 2.36	94.27 ± 2.33	84.47 ± 2.55	93.44 ± 2.06	83.30 ± 2.97
	$AdaM1^{CSC}_{PseAAC-S}$	97.17 ± 1.38	93.56 ± 1.22	88.48 ± 1.79	94.68 ± 2.01	92.92 ± 2.46	89.02 ± 2.67	79.51 ± 2.75	90.26 ± 2.34

* N/A=Not available.

Table 5
Relative improvement in performance of the proposed over conventional ensembles.

Approaches	RAI (for C/NC dataset) (%)					RAI (for B/NBC dataset) (%)				
	AUC (%)	Acc (%)	Sp (%)	Sn (%)	G_{mean} (%)	AUC (%)	Acc (%)	Sp (%)	Sn (%)	G_{mean} (%)
AdaBoostM1	5.7	−4.8	91.5	−18.55	31.42	23.39	−3.79	1449.05	−29.31	393.97
Bagging	4.79	−2.8	87.72	−16.42	31.65	0.21	−9.23	179.13	−21.62	66.96
GentleBoost	1.74	−6.73	72.84	−18.89	23.63	18.07	−6.43	861.98	−27.16	245.48
CSC-AdaBoostM1	3.16	10.92	−3.2	13.92	5.23	2.19	9.68	−12.18	12.09	−0.36
CSC-Bagging	1.58	19.23	−24.17	29.37	1.55	−2.36	27.18	−65.42	40.01	−16.91
Proposed CSC-GBE	–	–	–	–	–	–	–	–	–	–

In order to choose an appropriate number of learners for an ensemble system, several performance curves of GentleBoost, AdaBoostM1, and Bagging are generated using various feature spaces. Fig. 2 illustrates the ensemble error as a function of ensemble size using PseAAC-S feature space for C/NC and B/NBC datasets. As the number of learners in the ensemble increases the corresponding ensemble error asymptotically decreases up to the certain limit. From Fig. 2(a), we have selected number of learners 78, 81, and 19, which gave minimum ensemble errors for GentleBoost, AdaBoostM1, and Bagging using C/NC dataset, respectively. From Fig. 2(b), we have chosen number of learners 34, 58, and 08, which gave minimum ensemble errors for GentleBoost, AdaBoostM1, and Bagging using B/NBC dataset, respectively. It is found that the GentleBoost approach has smaller ensemble error than the other ensemble approaches.

The implementation of GentleBoost, using DT as base learners, together with CSL is highlighted in Fig. 1. Ten folds cross-validation data resampling is employed to report the average performance in terms of well known performance measures of area under ROC curve (AUC), accuracy (Acc), sensitivity (Sn), specificity (Sp), and *G-mean*.

4. Results and discussion

Table 2 shows the performance comparison of AdaBoostM1, Bagging, and GentleBoost ensemble approaches without taking into deliberation of the class imbalance problem (without CSC). It is observed that, for C/NC dataset, GentleBoost ensembles have provided the highest values of Acc 94.22% and 93.84%, for PseAAC-S and SAAC feature spaces respectively. Overall, the classification models performed better for PseAAC-S feature space and yielded average Acc near to 93.66%. Form Table 2, we found that GentleBoost ensemble has better decision than AdaBoostM1 and Bagging for B/NBC dataset. However, it is observed that, due to the imbalanced nature of the input data, the values of Sp are lower than Sn values. The inflated values of Sp verified that standard ensembles are biased towards majority class for C/NC and B/NBC datasets. The improved predicted results of the proposed CSC-GBE system in different feature spaces is compared with the conventional ensembles (see Table 3).

Detailed results of the CSC based ensemble approaches are provided in Table 3. From this table, it is found that our gentle ensemble system ($GBE^{CSC}_{PseAAC-S}$) using PseAAC-S feature space yields the highest Acc of 94.41% and 90.15% for C/NC and B/NBC datasets, respectively. It is observed that the values of Sn measure are smaller than Sp for standard approaches (Table 2). On the other hand, our approach considerably enhanced an average improvement nearly 15.62% for the values of Sn (see Tables 2 and 3) followed by a small drop in the values of Sp and Acc. This is extremely essential because it demonstrates that the novel approach effectively recognize more examples from the positive class that is very useful in the context of cancer diagnosis. Specifically the most important is to decrease the number of false negatives that is attained through the application of the proposed CSC technique. The elevated values of Sn verified that our proposed approach improved the classification performance by assimilating of CSC technique. Consequently, reduce bias of ensemble classifier towards the majority class for C/NC and B/NBC datasets. However, in CSL technique misclassification costs are normally not known. Darning simulation, several runs are performed to find the minimum misclassification costs.

ROC curves of CSC based models using AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces are depicted in Fig. 3(A and B). It is observed that for C/NC dataset (Fig. 3A (a–d)), all models have provided higher AUC (above 97%) in PseAAC-S feature space. However, for B/NBC dataset (Fig. 3B(a–d)), CSC-GentleBoost has showed the best performance (94.28%) in SAAC feature space,

followed by CSC-Bagging (94.27%) in PseAAC-S and CSC-AdaBoostM1 (93.34%) in SAAC space. On the other hand, in terms of AUC measure, an average enhancement of 2.14%, 0.60%, and 1.78% is observed for CSC-AdaBoostM1, CSC-Bagging, and CSC-GentleBoost ensemble models, respectively. CSL takes care of different misclassification costs of false negatives and false positives, the enhancement in performance demonstrated its effectiveness to handle imbalanced data. Thus, it is an efficiently technique, which reduce the total misclassification costs of the models. It is found that the proposed models have provided higher AUC in PseAAC-S and SAAC feature spaces for C/NC and B/NBC datasets, respectively.

In this study, we used only one type of sequential data belong to proteomic domain i.e., datasets of protein amino acids molecules of cancer and non-cancer. As cancer is a complex mutation disease, useful features (biomarkers) in genomic domain can be extracted using DNA-seq and/or RNA-seq datasets. The genomic domain features could be integrated with proteomic domain to generate more discriminant features. These new discriminant features might be helpful for the improvement of the cancer/non-cancer protein classification.

In Table 4, the performance of the proposed CSL based Gentle ensemble is compared to the state-of-the-art approaches developed using quantitative proteome-disease relationship (QPDR), over-sampling, and under-sampling methods for the same datasets. In [27], conventional QPDR models were developed using multiple linear regression technique. In [13], for small medical dataset, Mega-Trend diffusion method was employed to generate "synthetic" samples for over-sampling the minority class. For under-sampling, RUSBoost approach is employed. This approach uses random-under-sampling (RUS) method with AdaBoost algorithm for re-weighting and developing the ensemble classifier. RUS randomly eliminates instances from the majority class until required class distribution is attained [28]. The over/under-sampling based models are developed by employing tenfolds cross-validation technique.

For C/NC dataset, QPDR models have attained the Acc value of 90.81% [8]. The proposed model has obtained 3.91% improvement as compared to QPDR model. The over-sampling based $KNN_{PseAAC\text{-}S}$ and SVM_{SAAC} models have achieved the maximum AUC value of 87.77% and 90.00%, respectively [13]. Similarly, under-sampling based $RUSBoost_{PseAAC\text{-}S}$ and $RUSBoost_{PseAAC\text{-}P}$ models have yielded the values of AUC 93.05% and 85.29%, respectively. On the other hand, the performance of the proposed $GBE_{PseAAC-S}$ model is the highest among the conventional approaches (without CSL). Similarly, the proposed $GBE^{CSC}_{PseAAC-S}$ has yielded the highest AUC of 97.47% for C/NC datasets and $Bag^{CSC}_{PseAAC-S}$ has attained better AUC of 94.27% for B/NBC dataset.

For C/NC dataset, in terms of AUC, the proposed CSL models has demonstrated an enhancement of 9.70% and 7.47% compared to over-sampling based $KNN_{PseAAC\text{-}S}$ and SVM_{SAAC}, models, respectively. Further, an improvement of 4.42% and 12.18% of the proposed CSL models is found compared to under-sampling based $RUSBoost_{PseAAC\text{-}S}$ and $RUSBoost_{PseAAC\text{-}P}$ models of the reference dataset, respectively. Similar, an improvement of the proposed approach is observed for B/NBC dataset. From this analysis, we summarized that the proposed CSL approach is more effective than the over-sampling and the under-sampling methods.

Table 5 highlights the relative improvement ($RIA = \sum \frac{\alpha_l - \alpha'_i}{\alpha'_i}$)of the proposed CSC based models over conventional approaches, where α_l denotes the performance of proposed Can-CSC-GBE approach in the ith dataset and α'_i denotes the performance of the approach being compared with proposed approach. For C/NC dataset, the proposed Gentle ensemble has gained the highest relative improvement for Sn measure of 91.5% over AdaBoostM1. For Acc (19.23%) and AUC (4.79%) measures, we observed the highest relative improvements over CSC-Bagging and Bagging approaches, respectively. Similarly for B/NBC dataset, relative improvements in terms of Sn are more profound over conventional approaches. This considerable improvement in Sn indicates the effectiveness of CSC technique with GentleBoost approach. Thus the proposed approach system has efficiently reduced the misclassification costs and thereby improves the performance, particular in terms of Sn measure. For better medical decision, higher values of Sn and Sp measures are always demanded. On the other hand, overall reduced accuracy given in Table 5 indicates its performance for minority examples. This highlighted that CSL technique has efficiently computed the misclassification costs of false negatives and false positives. The enhancement has demonstrated its success to handle imbalanced data.

5. Conclusions

The objective of this study was to develop an efficient and reliable ensemble system, Can-CSC-GBE, for the classification of breast cancer using protein amino acids sequences. This system has incorporated CSL technique with conventional ensembles approaches of GentleBoost, AdaBoostM1, and Bagging. Overall, experimental results highlighted that the proposed gentle ensemble outperformed conventional ensembles, over-sampling, under-sampling, and previously individual learning approaches. It is found that, due to unbiased nature of the classifier towards majority class, the values of Sn sufficiently improved for C/NC and B/NBC datasets. The elevated values of Sn verified that the proposed approach improved the classification performance by assimilating of CSC technique.

The proposed study has some limitations. The proposed approach employed one type of sequential data of protein molecules belongs to cancer and non-cancer to develop classification models. However, by incorporating the useful features of the DNA-seq and/or RNA-seq data might enhance the classification performance of the proposed models. The disadvantage with cost-sensitive learning is that misclassification costs are usually unidentified. In each simulation run, distinct cost matrices were constructed. Several runs are required to estimate class probabilities. On the other hand, our novel approach has advantage to integrate and exploit the learning capabilities of CSL and ensemble approaches to ameliorate the performance. The proposed system performed well because: (i) it has used the most informative feature spaces generated from physicochemical properties of amino acids; and (ii) it has successfully incorporated the misclassification costs to the imbalanced data. It is expected that this study could be advantageous for clinical prediction and future research related to proteome, computational biology, biomedical informatics, and drug discovery. In our future work, we plan to incorporate other types of sequences such as DNA-sequence and/or RNA- sequence to obtain cancer prediction.

Acknowledgments

This work is supported by Higher Education Commission, Government of Pakistan under Indigenous Ph.D. Fellowship Program-Batch No. II and VII (PIN no. 213-59474-2PS2-056 and 117-3250-EG7-012).

References

[1] IARC GLOBOCAN 2012, Estimated cancer incidence, mortality and prevalence worldwide, 2012.
[2] S. Ergin, O. Kilinc, A new feature extraction framework based on wavelets for

breast cancer diagnosis, Comput. Biol. Med. 51 (2015) 171–182.
[3] K.S. Sim, F.K. Chia, M.E. Nia, C.P. Tso, A.K. Chong, S.F. Abbas, S.S. Chong, Breast cancer detection from MR images through an auto-probing discrete Fourier transform system, Comput. Biol. Med. 49 (2015) 46–59.
[4] G.H.B. Miranda, J.C. Felipe, Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization, Comput. Biol. Med. 64 (2014) 334–346.
[5] A. Safdar, A. Majid, A. Khan, IDM-PhyChm-Ens: intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids, Amino Acids 46 (2014) 977–993.
[6] Y. Ji-Yeon, K. Yoshihara, K. Tanaka, M. Hatae, H. Masuzaki, H. Itamochi, M. Takano, K. Ushijima, J.L. Tanyi, G. Coukos, Y. Lu, G.B. Mills, R.G.W. Verhaak, Predicting time to ovarian carcinoma recurrence using protein markers, J. Clin. Investig. 123 (2013) 3740–3750.
[7] R.G. Ramani, S.G. Jacob, Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models, PLoS One 8 (2013) e58772.
[8] C.R. Munteanu, A.L. Magalhães, E. Uriarte, H. González-Díaz, Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices, J. Theor. Biol. 257 (2009) 303–311.
[9] M. Xin, J. Guo, H. Liu, J. Xie, X. Sun, Sequence-based prediction of dna-binding residues in proteins with conservation and correlation information, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (2012) 1766–1775.
[10] M.T. Mirza, A. Khan, M. Tahir, Y.S. Lee, MitProt-Pred: predicting mitochondrial proteins of plasmodium falciparum parasite using diverse physiochemical properties and ensemble classification, Comput. Biol. Med. 43 (2013) 1502–1511.
[11] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[12] C. Drummond, R.C. Holte, C4.5 decision tree, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in: Proceedings of Workshop on Learning from Imbalanced Data Sets II, International Conference on Machine Learning, 2003.
[13] A. Majid, A. Safdar, I. Mubashar, K. Nabeela, Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines, Comput. Methods Programs Biomed. 113 (2014) 792–808.
[14] K.J. Wang, B. Makond, K.M. Wang, An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data, BMC Med. Inf. Decis. Mak. 13 (2013) 124.
[15] Y. Liu, W. Cheng, Z. Lu, Decision tree based predictive models for breast cancer survivability on imbalance data, in: Proceedings of IEEE International Conference on Bioinformatics and Biomedical Engineering, Beijing, 2009, pp. 1–4.
[16] W. Zhang, F. Zeng, X. Wu, X. Zhang, R. Jian, A comparative study of ensemble learning approaches in the classification of breast cancer metastasis, in: Proceedings of IEEE International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, IJCBS'09, 2009, pp. 242–245.
[17] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2005) 113–127.
[18] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC Med. Inf. Decis. Mak. 11 (2011) 51.
[19] S. Dhahbi, W. Barhoumi, E. Zagrouba, Breast cancer diagnosis in digitized mammograms using curvelet moments, Comput. Biol. Med. 64 (2015) 79–90.
[20] C. Chen, X. Zhou, Y. Tian, X. Zou, P. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Anal. Biochem. 357 (2006) 116–121.
[21] H. Maqsood, A. Khan, M. Yeasin, Prediction of membrane proteins using split amino acid and ensemble classification, Amino Acids 42 (2012) 2447–2460.
[22] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[23] L. Breiman, Stacked regressions, Mach. Learn. 24 (1996) 49–64.
[24] Y. Freund, R. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 1996, pp. 148–156.
[25] R. Schapire, The strength of weak learnability, Mach. Learn. 5 (1990) 197–227.
[26] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2000) 337–407.
[27] C.R. Munteanu, A.L. Magalhães, E. Uriarte, H. González-Díaz, Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices, J. Theor. Biol. 257 (2009) 303–311.
[28] C. Seiffert, T.M. Khoshgoftaar, J.V. Hulse, A. Napolitano, RUSBoost: a hybrid approach to alleviating class imbalance, IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. 40 (2010) 185–197.

Dr. Safdar Ali Deputy Chief Scientist, received his M.Sc. degree in Physics from University of the Punjab, Pakistan and his Ph.D. degree in computer and information science from Department of Computer and Information Sciences at Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad. He has 25 years experience in research and development. His current research interests include artificial intelligence, machine learning, pattern recognition, image processing, and development of intelligent decision support system, as well as applications in human breast and colon cancers. In particular, Dr. Ali is interested in the development of computational approaches based ensemble systems for cancer and biomedical images.

Dr. Abdul Majid Professor, has 23 years career of research, development, and teaching in the Department of Computer and Information Sciences at PIEAS, Islamabad. He received MSc degree in Electronics from Quaid-i-Azam University in 1991. He obtained his MS and Ph.D. degrees in Computer Systems Engineering from GIK Institute, in 2003 and 2006, respectively. He has completed his Post-Doc Research in the Department of Mechatronics, GIST, South Korea, in 2010. His research areas include bioinformatics, pattern recognition, image processing, and machine learning. His prominent research includes the development of intelligent decision support systems for cancers. He is on the panel of reviewers of internationally reputed journals in the field of computer science, bioinformatics, and image processing.

Syed Gibran Javed did his BS in Computer Science from COMSATS Institute of Information Technology, Islamabad in 2004. He was awarded fellow-ship for MS studies at GIK Institute of Engineering Sciences and Technology, Pakistan. He completed his MS in Computer Systems Engineering in 2006. Currently, he is pursuing his Ph.D. studies at Department of Computer and Information Sciences, Islamabad, PIEAS. His research interests include medical image processing, bioinformatics, computational intelligence, and pattern recognition.

Mohsin Sattar has completed his four-year BCS Hons in Computer Science in 2004. He was awarded MS degree in Nuclear Engineering from PIEAS, Islamabad, Pakistan in 2008. He received Ph.D. scholarship from higher Education Commission in 2014. Currently, he is pursuing his Ph.D. studies at Department of Computer and Information Sciences, PIEAS. He is researching in the development of decision support system for breast cancer diagnosis using computational intelligence approaches. His research interests include medical image processing, computational intelligence, bioinformatics, and pattern recognition.