Can–Evo–Ens: Classifier stacking based evolutionary ensemble system for prediction of human breast cancer using amino acid sequences

Safdar Ali, Abdul Majid

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

Abstract

The diagnostic of human breast cancer is an intricate process and specific indicators may produce negative results. In order to avoid misleading results, accurate and reliable diagnostic system for breast cancer is indispensable. Recently, several interesting machine-learning (ML) approaches are proposed for prediction of breast cancer. To this end, we developed a novel classifier stacking based evolutionary ensemble system “Can–Evo–Ens” for predicting amino acid sequences associated with breast cancer. In this paper, first, we selected four diverse-type of ML algorithms of Naïve Bayes, K-Nearest Neighbor, Support Vector Machines, and Random Forest as base-level classifiers. These classifiers are trained individually in different feature spaces using physicochemical properties of amino acids. In order to exploit the decision spaces, the preliminary predictions of base-level classifiers are stacked. Genetic programming (GP) is then employed to develop a meta-classifier that optimal combine the predictions of the base classifiers. The most suitable threshold value of the best-evolved predictor is computed using Particle Swarm Optimization technique. Our experiments have demonstrated the robustness of “Can–Evo–Ens” system for independent validation dataset. The proposed system has achieved the highest value of Area Under Curve (AUC) of ROC Curve of 99.95% for cancer prediction. The comparative results revealed that proposed approach is better than individual ML approaches and conventional ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Subspace. It is expected that the proposed novel system would have a major impact on the fields of Biomedical, Genomics, Proteomics, Bioinformatics, and Drug Development.

1. Introduction

Cancer is one of the rapidly growing diseases in the world. Nearly, fourteen million people are diagnosed per year with cancer. It is estimated that this figure will increase up to 19 million in 2025. It is assessed that out of 24 million cancer patients, half of them could be prevented in 2035 [1]. There are many types of cancers associated with human organs such as breast, colorectum, lung, and prostate. The most commonly diagnosed cancers are related to lung (1.8 million, 13.0% of the total), breast (1.7 million, 11.9%), and colorectum (1.4 million, 9.7%) [2]. Human breast cancer is the second foremost cause of worldwide cancer related deaths. It is most widespread cancer among women. Like other cancers, breast cancer can be successfully treated if predicted in early stages. Approximately 1.5 million cases of women breast cancer are registered per year, worldwide. About 88% of women diagnosed with breast cancer would survive at least 10 years. In US, due to early detection and treatment, more than 2.9 million women diagnosed with breast cancer were alive in 2012 [3]. In Pakistan, approximately 36,750 new women breast cancers cases are estimated for 2015 and about 17,552 women would die due to breast cancer. In order to increase the survival rate and to reduce cost, early prediction of breast cancer with reliable decision support system is essential. Such a system would also be helpful in avoiding unnecessary toxicity to the patient.

Human body cells normally grow in a regular and controlled pattern. However, during cancer development, body cells grow without control. The causes of uncontrolled growth of abnormal cells are due to mutation in genes. Cancer genomes are very unstable and usually show extensive genomic alterations. These changes vary from intragenic mutations to gross gains and/or loss of chromosomal material [4,5]. Genetic mutations, deletions, and allelic loss of tumor suppressor genes produce aberrant RNA transcript

© 2015 Elsevier Inc. All rights reserved.
that mostly affects normal function of the translated protein. Examples of commonly mutated tumor suppressor genes in human cancer are BRCA1, BRCA2, and PS3. Usually, majority of the mutations influence these genes and effect in protein sequences with deletions, insertions or truncations. This variation in protein sequences of amino acids would be helpful for developing cancer diagnosis system.

Breast cancer refers to a malignant tumor that occurs in human breast cells. It is a genomically complex and heterogeneous disease [6]. Breast cancer is widely studied with genomic technologies with efforts to develop molecular predictors of clinical outcomes and drug response. It is very different from other types of cancers. For instance, breast cancer and prostate cancer are two types of diseases. Their response could vary under different diagnosis and treatment methods. That is why for diagnosis, treatment, and drug discovery of breast cancer different types of signatures/features are required. The molecular signatures of breast cancer are identified using fundamental knowledge of system biology, cell biology, structural biology, genomics, and proteomics. However, in our view, proteins molecules based features that we have extracted using hydrophobicity (Hd) and hydrophilicity (Hb) properties of amino acids of protein would be quite helpful for breast cancer prediction. These physicochemical properties have been used in the study of protein foldings, structures, protein–protein interaction, and sequence-order effects [7].

In literature, several breast cancer prediction systems are developed using different feature extraction strategies [8–13]. A review of different feature extraction and selection techniques related to protein sequence analysis is presented in [14]. Protein sequences of amino acids have been utilized for the prediction of ovarian cancer, lung cancer [15], colon cancer, and breast cancer [16]. Jene-Sanz et al. have identified levels of 1200 genes expression [17]. They are directly controlled by enzyme, EZH2, which is correlated with the aggressiveness of breast cancer cases. Ahmad et al. have used Bayesian network technique to construct a gene regulatory network from microarray data [18].

Research interest is growing exponentially by integrating various classifiers to improve the performance. The integration approach is referred to as ensemble or collective decision-making system. Xin Ma et al. have applied Random Forest based ensemble system using protein sequence-based features for prediction of DNA-binding residues [19]. Goodman et al. have used non-evolutionary methods Optimized-LVQ (Learning Vector Quantization), Big-LVQ, and Artificial Immune Recognition System (AIRS) [20] with clinical features computed from digitized image of a fine needle aspirate (FNA) of breast mass. The combination of Evolutionary Algorithms (EAs) and SVM, with clinical features obtained from digitized image of FNA of breast mass have yielded 97.07% accuracy [21]. Protein features of malignant and benign cancers are evaluated using different screening methods. For example, Decision Tree models, generalized rule induction, and clustering methods for identification of similarity patterns in benign and malignant breast cancer tissues [22]. In another study, Aminizadeh et al. have developed RotBoost ensemble approach with microarray genes and attained 94.39% accuracy [23]. RotBoost ensemble is constructed by integrating the ideas of Rotation Forest and AdaBoost. Lavanya and Rani have applied Bagging and Boosting for ensemble based decision making systems [24]. Their ensemble systems have used Classification and Regression Trees for feature selection and achieved prediction accuracies up to 97.85% and 95.56%, respectively.

Though, several ensemble systems are proposed for various applications. However, it is still a challenging task to develop a high performing ensemble system for cancer prediction. Earlier, Boosting, Bagging, and RF based ensemble systems are developed by generating a set of classifiers that are trained from single learning algorithm. These approaches attempt to improve the prediction performance by iteratively retraining the base classifiers with a subset of most informative training data. These approaches have limited performance due to small number of biological samples and class imbalance. Another limitation is that these approaches have merely one level by taking the original input data to give single output prediction. On the other hand, our proposed stacking ensemble system has generated a set of classifiers that are trained from different learning algorithms.

Our approach effectively combines multiple level models for prediction [25–27]. It has two level classifier structures; (i) base classifiers are used for generation of preliminary predictions, and (ii) meta-classifiers are used for fusion of base classifiers. The 1st level base classifiers are trained on the original input dataset, and their predicted results are extracted called meta-data. Then, 2nd level meta-classifier is trained on this new dataset to obtain final prediction. The role of 2nd level classifier is to find out how best to combine the results of the base classifiers. However, to the best of our knowledge, in previous studies, GP based evolutionary ensemble as stacking classifier is not tailored for early prediction of breast cancer using amino acid sequences. GP technique is based on the principles of natural selection and recombination under defined fitness criterion. It is a powerful evolutionary approach, which searches for possible solutions in the defined problem space. This approach was used effectively in different applications of pattern recognition [28–31]. Here, we employed GP to develop a new ensemble system of ameliorated performance by taking advantage of its ability to explore and exploit the search space in an effective way for breast cancer prediction.

The proposed evolutionary ‘Can–Evo–Ens’ system is developed by exploring diversities in both feature and decision spaces. The input data of protein primary sequences is converted into features of Amino Acid Composition (AAC), Split Amino Acid Composition (SAAC), Pseudo Amino Acid Composition-Series (PseAAC-S), and Pseudo Amino Acid Composition-Parallel (PseAAC-P). We have chosen four diverse types of ML algorithms of Naïve Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and Random Forest (RF) as base-level classifiers/predictors. These predictors are trained in different feature spaces. Next, predictions of base-level predictors are stacked. During GP evolution process, predictions of the base-level predictors are combined using fitness criterion of Area Under Curves of Receiver Operating Characteristic (AUCS-ROC). Finally, the performance of best individual evolved at the end of GP simulations is computed using optimal threshold obtained from Particle Swarm Optimization (PSO) algorithm. The performance is evaluated on two benchmark cancer datasets using various quality measures of AUC, Accuracy (Acc), Specificity (Sp), Sensitivity (Sn), G-mean, F-score, and Mathews Correlation Coefficient (MCC). The overall performance of individual predictors is reported using 10-fold cross-validation data sampling technique [32]. Our results show that evolutionary ensemble systems are more accurate than either Bagging or Boosting ensembles. The comparative performance highlights that the proposed approach is superior to individual, conventional ensemble, and previous approaches. It is observed that PseAAC-S feature space has better discrimination power over the rest of feature extraction strategies for cancer prediction.

2. Material and methods

Fig. 1 shows the basic block diagram of the proposed system for cancer prediction. Fig. 1a indicates data preprocessing, development and predictions stacking of base-level predictors (1st level). Fig. 1b demonstrates the working principle of GP (2nd level), PSO based optimal threshold, and system performance evaluation module.
2.1. Input protein sequences

We performed experiments on two real datasets of human protein amino acid sequences for cancer/non-cancer (C/NC) and breast/non-breast cancers (B/NBC) datasets. These datasets were obtained from Sjoblom and Dobson groups [33,34]. These groups have extracted cancer-related proteins from the experimental analysis of 13,023 genes in 11 breast and 11 colorectal cancers. They extracted total of 1056 proteins sequences, in which 865 are non-cancer protein sequences (first dataset) and 191 cancer related protein sequences (2nd dataset). The 191 cancer protein sequences contain 122 protein sequences that are related to breast-cancer sequences.

In the present study, we merely focused on the variation of amino acid compounds in cancerous protein primary sequences using physicochemical properties of amino acids in early stages of cancer development. Supplementary Fig. S1(a–c) highlights the percentage change in concentration of amino acid compounds for general-cancer and breast-cancer protein with respect to non-cancer protein sequences. From this figure, it is obvious that composition of all amino acids is disturbed in cancerous proteins.

2.2. Feature generation strategies

Suitable representation of protein sequences makes informative patterns for a predictor to recognize underlying regularities in the sequences. Protein sequence of amino acids holds intrinsic dependencies between their contiguous constituent elements. Such dependencies in the data increase the richness of the representation. Therefore, in order to develop appropriate representation of protein sequences, the dependencies in sequence data are modeled in the form of mathematical expressions. These expressions satisfactorily reflect the inherent correlation with corresponding target labels. These expressions are used to generate different types of feature spaces for protein system. Features generated from physicochemical properties of amino acids are quite useful in the prediction of cancer. Due to differences in side chains, each amino acid has different physicochemical properties. These properties are used to convert amino acid codes into numerical values to extract effective protein features [35,36].

We used four correlation based discriminant feature extraction strategies of AAC, SAAC, PseAAC-S, and PseAAC-P for the prediction of breast cancer. The details of feature extraction strategies are provided in Table 1. Simple AAC based features are unable to find important concealed information of protein sequences. However, PseAAC based features of a protein are represented without losing its sequence-order information [37]. These features are widely employed to predict various attributes of proteins. PseAAC is popular for the predictions of protein related problems such as predicting enzyme family class, protein subcellular localization, outer membrane proteins, and protein structural class [38,39]. We have employed both series correlation and parallel correlation based PseAAC features. In series correlation (PseAAC-S), a protein feature vector is extracted by $20 + r + \lambda$ discrete components. We employed hydrophilic and hydrophobic as amino acid attributes to form 60D feature vector ($\lambda = 20$ and $r = 2$). In this series correlation, set of $20 + 2\lambda$ components is termed as the amphiphatic pseudo amino acid composition. In parallel correlation (PseAAC-P), a protein feature vector is represented by $20 + \lambda$ discrete components.
Table 1
Feature space extraction strategies and protein vectors.

<table>
<thead>
<tr>
<th>Feature space</th>
<th>Protein vector</th>
<th>Dim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Amino acid composition (AAC) [38,40]</td>
<td>( P_{\text{AAC}} = [p_1, p_2, \ldots, p_m]^T ) where ( p_1, p_2, \ldots, p_m ) are the native composition discrete numbers of 20 amino acids of a protein ( P )</td>
<td>20</td>
</tr>
<tr>
<td>2. Split amino acid composition (SAAC) [41]</td>
<td>The protein primary sequence is split into three parts, i.e., N-terminal, C-terminal, and internal segments. The amino acid compositions at N-terminal and C-terminal segments were separately calculated. In this work, length of each segment, N-terminal, internal and C-terminal is set at 20. In this way, the vector length for SAAC becomes 60D.</td>
<td>60</td>
</tr>
<tr>
<td>3. Pseudo amino acid composition-series (PseAAC-S) [37]</td>
<td>( P_{\text{min}} = [p_1 \cdot p_2 p_3 \cdot \ldots \cdot p_{20 \cdot 1} + \ldots + p_{20 \cdot 2} \cdot p_{20 \cdot 2} ]^T ) with ( p_i = \frac{\sum_{s=1}^{N} x_{hs} x_{is}}{\sum_{s=1}^{N} x_{hs}^2} ) for ( 1 &lt; \mu &lt; 20 ) and ( \mu = 21 ) where ( f(j = 1, 2, \ldots, 20) ) are the normalized occurrence frequencies of 20 native amino acids in the protein ( P ) and ( \tau_j ) the jth sequence correlation factor computed according to ( \tau_{2i-1} = \frac{\sum_{s=1}^{N} x_{hs} x_{is}}{\sum_{s=1}^{N} x_{hs}^2} ) and ( \tau_{2i} = \frac{\sum_{s=1}^{N} x_{hs} x_{is}}{\sum_{s=1}^{N} x_{hs}^2} ) for ( 1 &lt; \mu &lt; 20 + 2 \cdot \tau ).</td>
<td>60*</td>
</tr>
<tr>
<td>4. Pseudo amino acid composition-parallel (PseAAC-P) [37]</td>
<td>( P_{\text{parallel}} = [p_1 \cdot p_2 p_3 \cdot \ldots \cdot p_{20 \cdot 1} \cdot p_{20 \cdot 1}]^T ) with ( p_i = \frac{\sum_{s=1}^{N} x_{hs} x_{is}}{\sum_{s=1}^{N} x_{hs}^2} ) for ( 1 &lt; \mu &lt; 20 ) and ( \mu = 21 ) where ( \theta_i ) represents ( i )-tier correlation factor, which reflects the sequence order correlation between all the ( i ) most contiguous residues along a protein chain. It is computed according to ( \theta_i = \frac{\sum_{s=1}^{N} x_{hs} x_{is}}{\sum_{s=1}^{N} x_{hs}^2} ). The value of ( \theta(R_i, R_j) ) is calculated as follow: ( \theta(R_i, R_j) = 0.5 \left( \frac{H(R_i) - H(R_j)}{H(R_i) + H(R_j)} + \frac{H(R_i) - H(R_j)}{H(R_i) + H(R_j)} \right) ) where ( H(R_i) ) and ( H(R_j) ) are hydrophobicity and hydrophilicity of ( i )-th amino acid ( R_i ) and ( j )-th amino acid ( R_j ), respectively</td>
<td>40**</td>
</tr>
</tbody>
</table>

* \( 20 = 2 \cdot \tau \) and "20 + 2 \cdot \tau \) where \( \tau \) is tiers level and " \( \tau \) is number of amino acid attributes, we used \( \tau = 20 \) and \( \tau = 2 \).

2.3. Data preprocessing

In preprocessing module, we performed various tasks of data balancing, class labeling, and training/testing datasets formation. Generally, in medical applications data samples are imbalanced that results in poor prediction performance for the minority class. The decision boundary of the predictor is biased toward the majority class. Oversampling/under-sampling techniques can be utilized to handle the imbalanced problem. However, under-sampling can discard useful medical and biological information of the majority class data that could be important for the induction process. Therefore, to avoid the risk of deleting useful information form majority data class, we employed Mega Trend Diffusion (MTD) function. This function generates diffuse samples of the majority class in feature space. The details of MTD is available in our previous work [32]. We scaled feature’s values in the datasets between [0,1] so that features with large values could not dominate the features with small values.

The dataset of protein amino acid sequences is randomly divided into two separate parts: (i) training dataset (Trn) and (ii) validation or testing dataset (Tst). On the average, about \( (1-e^{-1}) \approx 2/3 \) data is used for training and about \( e^{-1} \approx 1/3 \) data is leaving for the network evaluation. Each base predictor (classifier) is train for Trn dataset and thereby obtained a new meta-data (Trn_pred) for the design of GEP based evolutionary ensemble “Evo–Ens”. The performance of base level prediction models is reported using Trn dataset. On the other hand, independent validation dataset (Tst) is used to evaluate the performance of Can–Evo–Ens system (i.e. the optimal model generated using PSE to find the optimal threshold). Here, the “Evo–Ens” represents the output of GP module that is developed at the end of GP process. However, “Can–Evo–Ens” denotes the complete cancer evolutionary ensemble system.

2.4. Classifier stacking

The predictions of base-level predictors are stacked to develop Evo–Ens model. A set of base-level predictors \( \{C_1, C_2, \ldots, C_m\} \) is constructed on training dataset of \( N \) samples, \( S_i = \{x^{(0)}, t^{(0)}\}_{n=1}^{N} \), where \( x^{(0)} \) represents the \( n \)th feature vector of protein sequences correspond to target \( t^{(0)} \). We obtained a set of predictions \( \{x_1, x_2, \ldots, x_m\} \) in numerical form using \( m \) base predictors i.e., \( x_j = C_j(x) \) represents \( j \)th predictor on \( x^{(0)} \) sample. Note that, these predictions in the form of numerical values are very important related to the evolved GP function because the base prediction values are used as the leaf nodes of individuals in GP. The most commonly used arithmetic, trigonometric, logarithmic, etc. functions are defined in GPlab and it is assured and automatically care that no leaf node becomes a negative value (e.g. \( \sin(x_{<2}) \) in Fig. 4). Hence, at meta level training, m-dimensional feature vector is formed \( X_m = (x_1, x_2, \ldots, x_m) \). In this way, for GP training a new meta data of \( N \) sample points is constructed, i.e., \( S_m = \left\{X^{(0)}, t^{(0)}\right\}_{n=1}^{N} \). GP technique maps prediction vectors \( X^{(0)} \) of base-level predictors to target labels \( t^{(0)} \). At the end of GP process, the best numerical Evo–Ens function, represented \( F_e(X) \), is developed.

2.5. GP evolution process

GP evolutionary approach can effectively exploit the search space to find the best candidate solution [42]. In GP evolution process, first, initial population of predictor functions is constructed, i.e., \( \phi = I_{i}^{(0)}(X) \), where \( X \in \mathbb{R}^{m} \), \( \phi \in \mathbb{R} \), \( m \) is m dimensional real vector and \( \theta \) denotes set of selected GP parameters. We provided a set of functions \{plus, minus, times, divide, log, sin, cos, exp, power\}, variables \( (X_1, X_2, X_3, X_4) \), and randomly generated constants, in order to find suitable structure of target function. The candidate solutions \( I_{i}^{(0)}(X) \) are representing in the form of tree structure. This tree-like representation consists of variable size. Adaptable tree representation automatically discovers the underlying useful pattern within data. The terminal set of tree comprises of useful feature vectors and random constants generated with uniform distribution. The most informative values of parameters and variables are chosen. The initial population of 100 individuals is generated using ramped half-and-half method.

In second step, the fitness scores of individual candidates \( I_{i}^{(0)}(X) \) are evaluated. The fitness score demonstrates how well GP individual moves toward the optimal solution. The success of evolutionary approach depends upon the accurate design of the fitness function. In this study, AUC-ROC is used as fitness-criterion. The area is calculated using the trapezoids method, defined as:

\[
I_s = \frac{1}{N} \sum_{k} \left( \frac{TP_{k+1} - TP_{k}}{2} + \frac{FP_{k+1} - FP_{k}}{2} \right) \tag{1}
\]
where $N_k$ denotes the number of class thresholds, and $TP_k$ and $FP_k$ represent the true positive and the false positive at class threshold $k$. The equation sums the area of the individual trapezoids of heights $h_1$ and $h_2$ and width $w$, fitted under the ROC points. This measure returns values between 0 and 1; higher the value, better the performance. The AUC corresponds to the probability that a minority class example is correctly predicted. Fig. 2a and b represents computation of GP solution on input examples and ROC curve for each threshold $T$. The maximum fitness score $I_p$ indicates how successfully an individual $I_p(X)$ moves towards the optimal point.

During third step, current population is used to choose the best candidates. In population of size $P$, the fitness probability of individual candidates, $I_p(X)$ is obtained as

$$Pr(I_p) = \frac{I_p}{\sum_p I_p}$$

where $\sum_p I_p$ denotes the total fitness of the total population size. The individual with higher probability values has greater possibility to produce offspring. Fourth step creates new population by applying crossover, mutation, and replication operators to individual parents. GP process is initiated to automatically speed up the convergence process while maintaining the population diversity. During simulation, we selected the genetic operator probabilities, crossover, and mutation rates to ‘variable’. The GPLAB software adopts the rates of these search operators to reflect their performance, based on the procedure as described in [43]. The rate of the operator will increase that has performed well in the previous search processes. That operator has more chance to produce offspring again. We used two selection stopping criteria:

(i) maximum generations reach up to 200, or

(ii) maximum fitness score ($I_p \geq 0.999$).

Ultimately, the best individual $F_{\max}(\mathbf{X})$ in the population, i.e., $I_p(\mathbf{X} \rightarrow F_{\max}(\mathbf{X}))$ is chosen. Since values of $F_{\max}(\mathbf{X})$ varies with different threshold, it is desirable to choose the most suitable threshold to classify the dataset as cancer and non-cancer.

2.6. Computing optimal threshold

Conventional search techniques such as grid search can be used to obtain the threshold values for Evo–Ens functions. However, for these search techniques, we have to adjust manually suitable grid range and step size. The computational complexity of the problem depends on grid range and its step size. However, for efficient computation, we preferred to use Particle Swarm Optimization based intelligence technique to find optimal threshold for the Evo–Ens functions in different feature spaces. The dimensionality of the search space is the same as the dimension of the feature space. In PSO, initial population is started with a random set of threshold (particle) for the GP expression. The position of each particle refers to a candidate solution. PSO finds the fitness value of each particle to determine personal best (Pbest) and global best (Gbest) particles. The particles are moved toward the optimal area by updating their position and the velocity according to the algorithm [44]. PSO has selected the best threshold values $\gamma_{\text{pso}}^{FS}$ for different feature space ($FS$) of AAC (20 dimensions), SAAC (60 dimensions), PseAAC-S (40 dimensions), and PseAAC-P (60 dimensions). The best predictions $\gamma_{\text{Ens}}^{FS}$ are computed for Cancer and Non-Cancer classes as:

$$\gamma_{\text{Ens}}^{FS} = \begin{cases} \text{Cancer,} & \text{if } f_{\text{Ens}}(\mathbf{X}) \geq \gamma_{\text{pso}}^{FS} \\ \text{Non-Cancer,} & \text{otherwise} \end{cases}$$

3. Parameter settings

In the following subsections, first, we explain selection procedures of parameter setting of individual predictors and then it is describe how Evo–Ens is develop using proper parameter setting in different feature spaces. Several simulations were carried out to select these parameters. A summary of the parameters used for the development of individual predictors (base level classifiers) and Evo–Ens are provided in Table 2.

3.1. Parameter settings of individual predictors

The individual predictors are trained using $Trn$ datasets in AAC, SAAC, PseAAC-S and PseAAC-P feature spaces. In the design of Evo–Ens, we selected four diverse types of base learners NB, KNN, SVM, and RF. Here, RF learner is selected, instead of Decision Tree, as a base learner, because Decision Tree learner has low accuracy and higher variance. On the other hand, Bayesian approach is computationally less complex and this approach has proved to be very effective in biological data classification problem. Detail information of NB, KNN, SVM, and RF learning algorithms is available in the literature of computational intelligence. We have implemented these algorithms in MATLAB R2013 environment. LibSVM software is used for the development of SVM models [45].

Description of parameters selection of different parametric individual predictors is given in Table 2 and their performance is depicted in Supplementary Fig. S2. KNN predictor is tuned by varying the values of $k$ for different feature spaces (Supplementary Fig. S2a). We have selected those values of $k$, which gave minimum prediction errors. Supplementary Figs. S2a and S2b depict

---

Fig. 2. (a) GP solution in the form of numeric outputs, where ‘+’ and ‘-’ be the cancer and non-cancer output classes, respectively, and $T_m$ and $T_n$ be the two different class thresholds. (b) ROC curve, which indicates two thresholds points $T_m$ and $T_n$ and corresponding area of a trapezoid.
prediction error of KNN predictors against number of nearest neighbors for C/NC and B/NBC datasets, respectively. In these figures, number of neighbors (K) is approximately evenly spaced on a logarithmic scale. These figures helped to select the best nearest neighbors in different spaces. Supplementary Fig. S2c illustrates the response of SVM predictor by varying the number of samples of minority class for breast cancer dataset. Table 2 shows the optimal parameter values of SVM predictors of the error term C and width of the Gaussian kernel σ. Similarly, for RF predictor this Table indicates the number of trees (ntree) and number of selected variables (mtry) in various feature spaces. For B/NBC dataset, Supplementary Fig. S2d shows their prediction accuracy against number of trees. Similar is the case for C/NC dataset, Table 2 gives the suitable number of trees used in different feature spaces.

3.2. Parameter setting of Evo–Ens

Table 2 shows the summary of necessary parameters to generate the best GP expressions. Several GP simulations were carried out to adjust these parameters using GPLAB toolbox (http://gplab.sourceforge.net/download.html) in MATLAB R2013a environment. Approximately 15–20 GP runs were carried out to obtain the best individual. The computational time of each GP simulation depends on several input parameters of input data size, population size, and number of generations. While, developing $F^{P\text{AAC}}_0(\hat{X})$ function, the most informative arithmetic and trigonometric functions are selected during GP evolution. Its performance depends on the useful combination of local information of the protein amino acid sequence. For C/NC dataset, during GP evolution process, we observed the progress of the best individual generations in different feature extraction strategies. A considerable improvement in fitness of the best GP individual is observed up to 8 generations for PseAAC-S and 50 generations for PseAAC-P feature spaces. For AAC and SAAC spaces, enough improvement is found in the best GP individual up to 40 generations. After about 50 generations, there occur smooth fitness transition and best-fit individual converges to near optimal for AAC, SAAC, and PseAAC-P. It is noticed that after 18 generations, best-fit individual for PseAAC-S space converges to optimal or near optimal.

For C/NC dataset, the best numerical functions are developed using AAC, SAAC, PseAAC-S, and PseAAC-P spaces. These functions in the prefix form are given below:

$F^{P\text{AAC}}_0(\hat{X}) = \text{minus}(\text{minus}(\text{minus}(\text{times}(X_4, \text{sin}(X_2))), \text{times}(\text{minus}(\text{times}(X_1, X_3)), \text{log}(\text{cos}(\text{plus}(\text{plus}(X_1, X_3), \text{sin}(X_2))))))), \text{cos}(\text{divide}(\text{sin}(X_2), 0.63288))), X_4), X_3)$

(4)

$F^{P\text{SAAC}}_0(\hat{X}) = \text{divide}(\text{times}(\text{minus}(X_3, \text{times}(\text{plus}(\text{plus}(X_1, X_3), X_2)), \text{cos}(\text{divide}(\text{minus}(\text{abs}(\text{divide}(X_4, X_3))), X_3, \text{times}(\text{times}(X_1, X_3), \text{cos}(X_3))), \text{times}(\text{times}(\text{cos}(X_1), X_3), \text{sin}(X_1)))))), X_3), \text{log}(\text{cos}(X_3)))$

(5)

$F^{P\text{PseAAC-S}}_0(\hat{X}) = \text{minus}(\text{sin}(\text{divide}(X_1, X_1))), \text{minus}(\text{plus}(\text{sin}(X_1)), \text{cos}(\text{minus}(\text{times}(X_4, \text{times}(\text{log}(\text{log}(X_1)), 0.49498))), X_4)))$

(6)

and

$F^{P\text{PseAAC-P}}_0(\hat{X}) = \text{plus}(\text{plus}(\text{cos}(X_2), \text{times}(\text{minus}(\text{minus}(\text{minus}(\text{cos}(\text{minus}(\text{minus}(\text{times}(X_4, X_3), X_3), X_3)), \text{cos}(\text{sin}(\text{times}(X_1))), X_1)), X_3), \text{abs}(\text{sin}(\text{cos}(\text{minus}(\text{cos}(X_2), X_2)), \text{cos}(\text{minus}(\text{abs}(X_3), X_3))))), X_3), \text{cos}(\text{minus}(\text{cos}(X_3), X_3)), \text{cos}(\text{minus}(\text{abs}(X_3), X_3))))$

(7)

The tree structure of Evo–Ens for SAAC feature space is shown in the Supplementary Fig. S3. This graphical representation demonstrated its functional dependency on the predictions of base-level predictors $X = \{X_1, X_2, X_3, X_4\}$ along with selected arithmetic and trigonometric functions.
For breast cancer, optimal numerical functions $F_{opt}^+$ ($\mathbf{X}$) are developed for AAC, SAAC, PseAAC-S, and PseAAC-P spaces. These functions are given below:

$$F_{opt}^{AAC} (\mathbf{X}) = \min(\cos(\times(X_1, \times(\min(\cos(X_2), X_2))), \cos(X_1), 0.75846)), X_1)$$

For breast cancer, optimal numerical functions $F_{opt}^+$ ($\mathbf{X}$) are developed for AAC, SAAC, PseAAC-S, and PseAAC-P spaces. These functions are given below:

$$F_{opt}^{AAC} (\mathbf{X}) = \min(\cos(\times(X_1, \times(\min(\cos(X_2), X_2))), \cos(X_1), 0.75846)), X_1)$$

$$F_{opt}^{SAAC} (\mathbf{X}) = \min(\cos(\times(X_1, \times(\min(\cos(X_2), X_2))), \cos(X_1), 0.75846)), X_1)$$

$$F_{opt}^{PseAAC-S} (\mathbf{X}) = \min(\cos(\times(X_1, \times(\min(\cos(X_2), X_2))), \cos(X_1), 0.75846)), X_1)$$

$$F_{opt}^{PseAAC-P} (\mathbf{X}) = \min(\cos(\times(X_1, \times(\min(\cos(X_2), X_2))), \cos(X_1), 0.75846)), X_1)$$

and

$$F_{opt}^{PseAAC-P} (\mathbf{X}) = \min(\cos(\times(X_1, \times(\min(\cos(X_2), X_2))), \cos(X_1), 0.75846)), X_1)$$

$$F_{opt}^{PseAAC-P} (\mathbf{X}) = \min(\cos(\times(X_1, \times(\min(\cos(X_2), X_2))), \cos(X_1), 0.75846)), X_1)$$

Generally, the complex structure of predictor functions may not be analyzed/understood by human expert. However, these empirical functions can be easily computed by machine. For example, Fig. 4 and Supplementary Fig. S4 show the tree structures of the best individual functions for PseAAC-P and SAAC spaces, respectively. These graphical representations highlight the functional dependency of GP expressions on predictions of base predictors.

4. Results

In this section, we present the experiment results of the proposed Can–Evo–Ens system for C/NC and B/NBC datasets. The performance of individual and proposed predictors is assessed using different quality measures in various feature spaces (see supplementary material). The overall results are obtained using 10-fold cross-validation data sampling technique. We compared the overall performance of our Can–Evo–Ens system with individual (base-level predictors) and approaches from previous studies.

4.1. Performance of individual predictors

Table 3 highlights the performance achieved, using 10-fold cross-validation data sampling technique, by individual base predictors in different feature spaces for C/NC and B/NBC datasets. For C/NC dataset, NB predictor has obtained Acc value of 93.12% in PseAAC-P and overall highest MCC value of 74.97% in SAAC space compared to other feature spaces. KNN predictor has attained the best values of Acc 95.43%, Sn 96.53%, Sp 94.34%, Gmean 95.43%, F-score 95.38%, and MCC 64.36% for PseAAC-S feature space. The
prediction \text{Acc} 96.99\% of SVM is better in AAC space compared to other spaces. However, it is observed that RF predictor in PseAAC-S space has achieved the highest values of \text{Acc} 97.92\%, \text{Sn} 98.26\%, \text{Sp} 97.57\%, \text{Gmean} 97.91\%, \text{Fscore} 97.91\%, and \text{MCC} 68.01\%. For B/NBC dataset, SVM predictor in PseAAC-P space has provided the best \text{Acc} value 94.33\%. NB predictor has gained \text{Acc} value 94.33\%, again, in PseAAC-P feature space compared to other spaces. But, in SAAC space it has best \text{MCC} value 72.89\% among other feature spaces.

\textbf{Fig. 5} (a–d) and B(a–d) demonstrates the partial ROC curves of individual predictors, using 10-fold cross-validation technique, for different feature spaces of C/NC and B/NBC datasets. From \textbf{Fig. 5} (a–d), it is observed that PseAAC-S based RF predictor has provided the best AUC value 99.57\%, followed by PseAAC-P (99.48\%), SAAC (99.41\%), and AAC (99.32\%) spaces. It is observed that predictors have provided the highest values of AUC measure in different feature spaces, for instance, NB (97.47\%) in PseAAC-P space, KNN (98.55\%) in PseAAC-S space, SVM (99.26\%) in SAAC space. The average prediction performances of RF are higher 0.81\%, 2.87\%, and 2.01\% than SVM, KNN, and NB predictors, respectively.

\textbf{Fig. 5} B(a–d) shows that RF has the best value of AUC for all feature spaces. From this figure, again, it is observed that PseAAC-S space has provided the best AUC value 99.19\% for RF predictor. However, other predictors have provided highest values of AUC in different spaces, for example, NB (98.71\%) in SAAC space, KNN (98.86\%) in PseAAC-P space, SVM (98.88\%) in SAAC space.

<table>
<thead>
<tr>
<th>Method/feature</th>
<th>C/NC dataset</th>
<th>B/NBC dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC</td>
<td>Sp</td>
</tr>
<tr>
<td>NB</td>
<td>90.75</td>
<td>82.08</td>
</tr>
<tr>
<td>AAC</td>
<td>88.96</td>
<td>77.92</td>
</tr>
<tr>
<td>SAC</td>
<td>90.81</td>
<td>96.42</td>
</tr>
<tr>
<td>PseAAC-S</td>
<td>93.12</td>
<td>86.24</td>
</tr>
<tr>
<td>KNN</td>
<td>93.12</td>
<td>93.3</td>
</tr>
<tr>
<td>AAC</td>
<td>91.04</td>
<td>90.75</td>
</tr>
<tr>
<td>SAC</td>
<td>95.43</td>
<td>96.53</td>
</tr>
<tr>
<td>PseAAC-P</td>
<td>94.51</td>
<td>96.3</td>
</tr>
<tr>
<td>SVM</td>
<td>96.99</td>
<td>95.49</td>
</tr>
<tr>
<td>AAC</td>
<td>96.65</td>
<td>95.26</td>
</tr>
<tr>
<td>SAC</td>
<td>92.14</td>
<td>84.51</td>
</tr>
<tr>
<td>PseAAC-S</td>
<td>96.19</td>
<td>93.18</td>
</tr>
<tr>
<td>PseAAC-P</td>
<td>97.92</td>
<td>98.26</td>
</tr>
<tr>
<td>RF</td>
<td>96.76</td>
<td>96.76</td>
</tr>
<tr>
<td>SAC</td>
<td>96.82</td>
<td>97.23</td>
</tr>
<tr>
<td>PseAAC-S</td>
<td>97.92</td>
<td>98.26</td>
</tr>
<tr>
<td>PseAAC-P</td>
<td>96.94</td>
<td>97.69</td>
</tr>
</tbody>
</table>

Bold values indicate the highest value of the predictors.
4.2. Performance of the proposed predictor

Table 4 shows the values of average $Q$ statistics of individual predictors in different feature spaces for C/NC and B/NBC datasets (2nd and 3rd columns). The lower value of $Q$ gives the higher improvement in the proposed predictor. Further, Table 4 shows PSO algorithm based optimal threshold $c_{FS}^{pso}$ values for the best GP individuals, in different feature spaces, using C/NC and B/NBC datasets (last two columns).

<table>
<thead>
<tr>
<th>Feature space</th>
<th>Average Q</th>
<th>$c_{FS}^{pso}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C/NC</td>
<td>B/NBC</td>
</tr>
<tr>
<td></td>
<td>dataset</td>
<td>dataset</td>
</tr>
<tr>
<td>AAC</td>
<td>0.3979</td>
<td>0.3985</td>
</tr>
<tr>
<td>SAAC</td>
<td>0.3971</td>
<td>0.3921</td>
</tr>
<tr>
<td>PseAAC-S</td>
<td>0.3965</td>
<td>0.3923</td>
</tr>
<tr>
<td>PseAAC-P</td>
<td>0.3987</td>
<td>0.3989</td>
</tr>
</tbody>
</table>

Fig. 5. (A and B) ROC curves (partial) of individual predictors, NB, KNN, SVM, and RF for (A) C/NC and (B) B/NBC datasets using: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P spaces. (Partial ROC curves are plotted for better visualization of region of interest. High sensitivity levels are desirable in a medical decision.)
represents the performance of the proposed approach in P/C0 denotes the performance of the approach (99.01%), and 98.39%, and has gained a

Analysis of variance (ANOVA) to estimate the significance improvement of models in different feature spaces. Table 6 shows the results of this analysis for C/NC and B/NBC datasets. These results highlight a significant difference among the models because the returned -value is near to zero. This is evidence that the performance varies from one model to another. The graph of Supplementary Fig. S6a shows that PseAAC-S feature space is significantly different from AAC and SAAC spaces. In case of B/NBC dataset, the -value is near to zero. This is evidence that the performance varies from one model to another. The graph of Supplementary Fig. S6b shows that PseAAC-S feature space is significantly different from AAC and SAAC feature spaces for B/NBC dataset.

Additionally, to assess the scalability of the approach, we have computed the computational time for different feature spaces. The average training time of the proposed approach is computed to be 1235.8 and 570.4 s in the feature spaces of PseAAC-S (60 dimensions) and PseAAC-P (40 dimensions), respectively, while keeping all other parameters constant. Therefore, the ensemble models in PseAAC-S space consumed about twice more time than PseAAC-P space.

From Fig. 6a, for C/NC prediction, it is observed that our system in PseAAC-S space has provided the best AUC value 99.95%, followed by SACC (99.93%), AAC (99.91%), and PseAAC-P (99.87%) spaces. For B/NBC prediction, Fig. 6b demonstrates that, again, PseAAC-S based predictor has provided the best AUC value 99.89%, followed by PseAAC-P (99.86%), ACC (99.82%), and SAAC (99.70%).

Supplementary Fig. S5 shows performance comparison of the proposed system in different feature spaces for C/NC and B/NBC datasets. This figure highlighted that our system is better for C/NC dataset in all feature spaces, except PseAAC-P. In case of PseAAC-P, our system has provided similar results for both datasets.

4.3. Overall comparison

In this subsection, we carried out a performance comparison of the proposed approach with individual and approaches from previous studies. Tables 3 and 5 summarize the overall performance of individual predictors and the proposed predictors. Our approach outperformed individual predictors in terms of Acc, Sp, MCC, and -score for C/NC and B/NBC datasets. NB predictor has shown slight progress over our predictor in terms of Sn and MCC for all feature spaces except PseAAC-S. SVM models show improvement in terms of MCC using PseAAC-S space for C/NC and B/NBC datasets. By comparing Figs. 5 and 6, it is also observed that among individual predictors, RF has achieved better performance, and outputs the highest AUC value 99.95% in PseAAC-S space for C/NC dataset.

For comparison purpose, we developed four well-known conventional ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Subspace. AdaBoostM1 and GentleBoost are implemented using Decision Tree as base learners. However, Bagging and Random Subspace are implemented using Discriminant Analysis as base learning algorithm. Supplementary Table S1 presents the results of ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Subspace. Fig. 7 shows performance comparison of the proposed ensemble system with conventional ensemble approaches in the best feature space PseAAC-S. For C/NC dataset, from Fig. 7a, it is evident that our proposed system has outperformed the conventional ensemble by providing the best AUC value of 99.95%. For B/NBC dataset, Fig. 7b highlights the comparison of our approach with other ensemble approaches in the best performing feature space (PseAAC-S). Again, it is observed that our approach has outperformed the previous ensemble approaches by producing the best AUC value of 99.89%.
conventional ensemble. From Table 7, it is evident that our approach has the highest values of RIA measure over individual predictors, particularly, for NB approach with C/NC and B/NBC datasets. In case of conventional ensembles, again our approach has the highest RIA values, particularly, over Random Subspace ensemble for C/NC and B/NBC datasets.

In Table 8, we carried out accuracy comparison of the proposed approach with previous well-known approaches for breast cancer. Optimized-LVQ and Big-LVQ models have provided accuracy near to 96.80% [47]. The prediction performance using clinical features has enhanced accuracy in the range of 97.07–97.51% for Fuzzy-GA, AR + NN, SVM + EAs, and SBS-BPPSO approaches. Ensemble (NF, KNN, QC) using information gain based selected clinical features has provided accuracy of 97.14% [48]. Topological indices based QPDR models have achieved maximum accuracy of 90.81% [16]. On the other hand, for C/NC dataset, our system using PseAAC-S feature space has given the best prediction of 99.02%. Similarly, for B/NBC dataset, we have achieved the best prediction of 98.39%.

5. Discussion

Our experimental results have demonstrated that the proposed evolutionary ensemble approach is more effective than individual and conventional ensembles approaches in predicting cancerous protein sequences. For the development of ensemble system, accurate and diverse characteristics of base predictors are vital. From Table 3 and Fig. 5, we found that the performance of individual base predictors varies in different feature spaces. RF predictor in PseAAC-S space has achieved the highest Acc compared to other spaces. This accurate response of predictors is helpful for our proposed system. Additionally, the lower values of average Q-statistic (Table 4) highlight the generation of useful diversity of base predictors in different feature spaces. Such diverse-type of response of individual predictors are beneficial for development of the proposed system. Through GP evolution process, we exploit effectively the useful diversity of base predictors. As a result the performance of the proposed system is ameliorated in different feature spaces.

We observed improved results of the proposed predictor (Table 5) than individual predictors (Table 3) using C/NC and B/NBC datasets. It is observed (Table 5) that using variation of amino acid compounds in cancerous protein primary sequences (see Supplementary Fig. S1), our proposed predictor has attained the highest Sp values of 99.42% and 98.40% for C/NC and B/NBC datasets, respectively. For B/NBC dataset, the average Sp of the proposed approach is higher with respect to Sn measure. Hence, our predictor has predicted cancerous protein sequences more precisely. It is evident from Tables 3 and 5 that after combining the predictions of base predictors, the value of Sp and Sn considerably ameliorate. It is found (Table 5) that although the Sn decreases slightly (Table 3) when applying GP module, the Sp is improved. Supplementary Fig. S2c also accentuates that the Sn of the classification is considerably improved by varying the imbalance ratio of the classes.

Fig. 6. ROC curves (partial) of the proposed predictors for (a) C/NC dataset and (b) B/NBC dataset using AAC, SAAC, PseAAC-S, and PseAAC-P spaces.

Fig. 7. Performance comparison of the proposed ensemble system with well-known ensemble approaches in the best PseAAC-S space for (a) C/NC dataset and (b) B/NBC datasets.
higher values of Sn and Sp are desired for medical decision. G_{mean} and F-score are also improved because these values depend on Sn and Sp measures.

The prediction performance of proposed system is analyzed in terms of ROC curves for C/NC and B/NBC datasets. Similarly, Fig. 6a and b shows the improved ROC curves of our system in different feature spaces. The improved ROC curve is helpful in selecting operating point of the predictor. The performance comparison of our system is conducted in different feature spaces for C/NC and B/NBC datasets. For C/NC datasets, Table 7 shows sufficient improvement over previous approaches in terms of ROC curves for C/NC and B/NBC datasets. On the other hand, our approach has shown sufficient improvement over previously approaches in terms of MCC (32.75%), AUC (14.81%), Acc (36.61%), G_{mean} (32.75%), F-score (13.48%), and Sn (13.48%). In terms of B/NBC datasets, we observed similar performance trend.

In case of conventional ensembles (Table 7), for C/NC dataset, the proposed approach has obtained relative improvement in AUC-ROC measure of 10.01% (AdaBoostM1), 9.62% (Bagging), 8.88% (GentleBoost), and 19.05% (Random Subspace). The relative improvement of accuracy values of 21.77%, 18.76%, 17.40%, and 45.82% has been observed with AdaBoostM1, Bagging, GentleBoost, and Random Subspace, respectively. The proposed approach has gained the highest relative improvement over Random Subspace in terms of Sn (55.38%), Acc (39.41%), G_{mean} (47.49%), and MCC (88.26%). Our approach has attained the smallest relative improvement over GentleBoost using AUC-ROC (8.88%), Acc (17.40%), and MCC (88.26%). Our approach has attained the smallest relative improvement over GentleBoost using AUC-ROC (8.88%), Acc (17.40%), and MCC (88.26%). Similar behavior is observed for B/NBC datasets. On the other hand, our approach has shown sufficient improvement over previous approaches (see Table 8) for both of C/NC and B/NBC datasets. During GP evolution process, complex structure of predictor functions is developed. These functions in the form of equations and figures accentuated the functional dependency on the predictions of base predictors. Therefore, the proposed approach has potential to exploit most informative feature spaces, which are extracted from the numerical descriptors based on physicochemical properties of amino acids. In the study, we found that our predictor in PseAAC-S space has provided the best performance for cancer prediction. This is because PseAAC-S feature space carries

### Table 7
Relative Improvement (RIA) of the proposed approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>RAI (C/NC dataset) (%)</th>
<th>RAI (B/NBC dataset) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>Acc</td>
</tr>
<tr>
<td>Base predictor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>10.79</td>
<td>34.77</td>
</tr>
<tr>
<td>KNN</td>
<td>8.1</td>
<td>22.63</td>
</tr>
<tr>
<td>RF</td>
<td>1.98</td>
<td>6.90</td>
</tr>
<tr>
<td>Conventional ensemble</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagging</td>
<td>9.62</td>
<td>18.76</td>
</tr>
<tr>
<td>GentleBoost</td>
<td>8.88</td>
<td>17.40</td>
</tr>
<tr>
<td>Random Subspace</td>
<td>19.05</td>
<td>45.82</td>
</tr>
<tr>
<td>Proposed approach</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Table 8
Prediction comparison of our proposed approach with other approaches.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Feature extraction strategy</th>
<th>Accuracy (B/NBC) (%)</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble (AdaboostM1 + SVM-SMO)</td>
<td>Clinical features</td>
<td>82.0</td>
<td>[49]</td>
</tr>
<tr>
<td>Boosting</td>
<td></td>
<td>95.43</td>
<td>[24]</td>
</tr>
<tr>
<td>Bagging</td>
<td></td>
<td>96.96</td>
<td>[24]</td>
</tr>
<tr>
<td>Optimized-LIQq</td>
<td></td>
<td>96.70</td>
<td>[20]</td>
</tr>
<tr>
<td>Big-LIQq</td>
<td></td>
<td>96.80</td>
<td>[20]</td>
</tr>
<tr>
<td>Fuzzy-GA</td>
<td></td>
<td>97.36</td>
<td>[50]</td>
</tr>
<tr>
<td>AR + NN</td>
<td></td>
<td>97.40</td>
<td>[51]</td>
</tr>
<tr>
<td>SVM + EAs</td>
<td></td>
<td>97.07</td>
<td>[21]</td>
</tr>
<tr>
<td>SBS-RPPS</td>
<td></td>
<td>97.51</td>
<td>[52]</td>
</tr>
<tr>
<td>Fuzzy-SVM</td>
<td></td>
<td>96.35</td>
<td>[53]</td>
</tr>
<tr>
<td>Ensemble (NF KNN QC)</td>
<td>Information gain based clinical features</td>
<td>97.14</td>
<td>[48]</td>
</tr>
</tbody>
</table>

# Table 8
<table>
<thead>
<tr>
<th>Approach</th>
<th>Feature extraction strategy</th>
<th>Accuracy (B/NBC) (%)</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/NC</td>
<td>B/NBC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QPDR</td>
<td>pTle^* (embedded)</td>
<td>90.00</td>
<td>[16]</td>
</tr>
<tr>
<td>RF</td>
<td>PseAAC-S</td>
<td>97.92</td>
<td>Present study</td>
</tr>
<tr>
<td>GentleBoost</td>
<td>PseAAC-S</td>
<td>96.36</td>
<td>94.75</td>
</tr>
<tr>
<td>Bagging</td>
<td>PseAAC-S</td>
<td>94.54</td>
<td>98.39</td>
</tr>
<tr>
<td>Proposed approach</td>
<td>PseAAC-S</td>
<td>99.02</td>
<td></td>
</tr>
</tbody>
</table>

^ With embedded star graph, dTle = same topological indices (TIs) minus average TIs for cancer, pTle = cancer probability TIs; find out more details from [16].
the most discriminant information. This discriminant information is due to effective use of Hd and Hb properties of amino acids of Proline, Tyrosine, Serine, Arginine, Asparagine, Isoleucine, and Cysteine (Supplementary Fig. S1).

The overall performance of the proposed system is impressive due to two reasons. First, the utilization of most informative features derived from the physicochemical properties of amino acids. At feature level, these features have a potential to accommodate the variation of amino acid composition in cancer and breast-cancer protein sequences with reference to non-cancer proteins. Second, our GP approach is designed differently from conventional approaches. At decision level, our approach has optimal combined the predictions of diverse types of learning algorithms and thereby ameliorate the performance. On the other hand, the proposed study has some limitations due to the stochastic nature of the GP technique. In order to find the best parametric values from the large search space, during GP training process, we had to run GP simulation several times. Further, the candidate solutions may convergence slowly near the global optima.

6. Conclusion

In this study we have proposed classifier stacking based evolutionary ensemble approach and developed Can–Evo–Ens system for the reliable prediction of breast and non-breast cancer. GP evolution process has combined effectively the diverse-type of useful information of base predictors by generating better decision space than individual and conventional ensemble approaches. The performance of the GP function also depends on useful information extracted from protein primary sequences in different feature spaces. The proposed system has demonstrated its robustness for independent validation dataset.

The proposed system using PseAAC-S space has achieved the highest values of accuracies 99.02% and 98.39% for C/NC and B/NBC datasets, respectively. Our approach has yielded excellent discrimination power in PseAAC-S space over other feature extraction strategies. In PseAAC-S space, the proposed system has provided the highest AUC values of 99.95% and 99.89% for C/NC and B/NBC datasets, respectively. Comparative analysis has demonstrated that our approach has performed superior than conventional ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Forests. Our approach has performed superior than conventional ensemble approaches. Comparative analysis has demonstrated that our approach has performed superior than conventional ensemble approaches.

Acknowledgments

Authors are grateful to Pakistan Institute of Engineering and Applied Sciences (PIEAS) for providing resources. We are also thankful to Munteanu, Sjoblom and Dobson groups for providing the cancer protein datasets to carry out this research.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.01.004.

References