Arabian Journal for Science and Engineering (2019) 44:3197–3211
https://doi.org/10.1007/s13369-018-3468-8

RESEARCH ARTICLE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

Lung Cancer Classification Models Using Discriminant Information of Mutated Genes in Protein Amino Acids Sequences

Mohsin Sattar[1] · **Abdul Majid**[1]

Received: 27 March 2018 / Accepted: 12 July 2018 / Published online: 31 July 2018

Abstract
Lung cancer is a heterogeneous disease based on uncontrollable growth of cells. Lung cancer is major cause of cancer-related deaths. Early diagnosis of lung cancer is important for its treatment and survival of patients. In this study, through the statistical analysis of cancerous proteins sequences, we observed the mutated genes associated with etiology of lung cancer. Our analysis revealed most frequent mutated genes TP53, EGFR, KMT2D, PDE4DIP, ATM, ZNF521, DICER1, CTNNB1 RUNX1T1, SMARCA4, FBXW7, NF1, PIK3CA, STK11, NTRk3, APC, PTPRB, BRCA2, MYH11 and AMER1. We observed abnormal mutations in genes contributed toward variations in the composition of amino acid sequences. This variation was described in various feature spaces using statistical and physicochemical properties of amino acids. These influential features have provided sufficient discrimination power for the development of effective lung cancer classification models (LCCMs). The main advantage of proposed novel approach is the effective utilization of the discriminant information of mutated genes. Experimental results showed that SVM model has the best performance in split amino acid composition. In the study, we explored a new dimension of early lung cancer classification using discriminant information of mutated genes revealed through the statistical analysis of the mutated genes. It is anticipated that the proposed approach would be useful for practitioners and domain experts for early lung cancer diagnosis and prognosis.

Keywords Lung cancer · Amino acids · Classification · Diagnosis · Prognosis · Machine learning

1 Introduction

Lung cancer is the major cause of cancer deaths in both men and women worldwide [1,2]. Lung cancer is common in smokers but non-smokers are also affected due to second-hand smoke [1]. Approximately 10% of lung cancers occur in non-smokers each year [3]. American Cancer Society (ACS) has estimated that about 27% of all cancer deaths were due to lung cancer. In the year 2017, it has been shown that death rate due to lung cancer was higher than other cancer-related deaths. They have reported, nearly 222,500 cases of lung cancer, out of which 155,870 people died due to lung cancer. Due to lack of early detection and prognosis, lung cancer incidence and mortality rates are increasing significantly [4,5]. There is need for developing effective classification model for treatment of lung cancer in early stages.

Generally, lung cancer is categorized into two groups: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [6–8]. NSCLC incidence rate is higher than SCLC but SCLC develops rapidly. However, NSCLC is further grouped into adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. SCLC is categorized into the limited and extensive stage. In the limited stage, part of lung and nearby lymph nodes are affected due to lung cancer. In the extensive stage, lung cancer is proliferated to other lymph nodes and body parts. In early stages of NSCLC, tumor size is restricted to 3 cm, however in later stages, tumor is expanded up to 5 cm [9].

The recent technological advancements have rapidly increased the proteomic and genomic sequential data. Such

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13369-018-3468-8) contains supplementary material, which is available to authorized users.

✉ Abdul Majid
abdulmajiid@pieas.edu.pk

Mohsin Sattar
mohsin_14@pieas.edu.pk

[1] Biomedical Informatics Research Lab, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, P.O. Nilore, Islamabad, Pakistan

databases contain useful information of various genes that are responsible for healthy function of body organs [10]. The healthy lung function is highly dependent on the normal function of genes. For example, TP53, EGFR, KMT2D, PDE4DIP, ATM, ZNF521, DICER1, CTNNB1, RUNX1T1, SMARCA4, FBXW7, NF1, PIK3CA, STK11, NTRk3, APC, PTPRB, BRCA2, MYH11, and AMER1 genes are mostly responsible for normal functions of cell growth, division, and apoptosis [10,11]. These genes play significant role in tumor suppression, transcriptional activation, DNA protein binding, hydrolase activity, phosphodiesterase, formation of adherens junctions, transcriptional repression, phosphorylation-dependent ubiquitination, negative regulation of the signal transduction pathway, activation of signaling pathways, and regulation of cell processes [10,11]. These activities are performed for smooth functioning of different cellular processes. However, abnormal mutations dysregulate the normal function that indicate the uncontrollable growth of lung cells.

There are many predominant factors that hinder the normal function of genes, for example, metabolic mutations, hereditary mutations, tumor suppressor genes mutations, malnourishment, tobacco use and other environmental factors [12,13]. Various types of mutations occur in different genes over life span. The abnormal mutations acquired in lung cells result from epigenetic and environmental factors [12,13]. The single nucleotide and missense mutations in tumor suppressor gene TP53 are universal across different cancer types [10,14,15]. Loss of tumor suppression activity is recognized as large deleterious events, in frame-shift mutations and/or premature stop codons. Due to cancer-related mutations in BRCA2, heterozygous features in the organism are lost [10,15]. The mutations in EGFR gene may cause lung cancer [10,15,16]. Most of the EGFR mutations are somatic, while a few germ-line mutations also exist [10,15,16]. The KMT2D is major mutated genes in SCLC that causes the perturbation of transcriptional enhancer control [10,15–17]. The most common mutations like missense, nonsense, frame-shift deletions and insertions cause lung cancer [10,15–17]. These cancer-driving somatic mutations change the protein amino acids composition. If we could detect such cancer/non-cancer protein amino acid sequences in early stages, the survival rate of patients would be increased.

Researchers have proposed different lung cancer classification models employing different data modalities, and feature selection/extraction techniques. Proteins structural and physicochemical properties were employed to develop different learning algorithms for the classification of lung cancer [18]. They have reported the best accuracy with Bayesian Network using hybrid feature selection. In another work [19], features based on the structural and physicochemical properties of protein sequences were employed to classify lung tumor. Gene expressions data was employed to classify lung cancer in [20]. Decision tree (DT) model was developed using DNA methylation markers to predict lung cancer [21]. In another study, Micro-RNA expression profiling was used to classify lung cancer [22]. In [23], authors classified cancer and non-cancer genes using radiographic signatures with clinical model. The structural and physiochemical properties of protein amino acid sequences were employed for the classification of lung cancer [18], colon cancer, ovarian cancer [24], and breast cancer [25–27]. The clinical features using bronchoscopy, lung needle biopsy are also common for lung cancer detection [28]. In these invasive techniques, a sample of lung cells is extracted for microscope analysis.

Mostly, the previous cancer detection models have used the physical features like geometry/size of tumor that appear in the later stages. In the later stages, the detrimental effects occur to lung cells, lymph nodes, and other body organs. These techniques are prone to increase the risk of lung cancer. In this scenario, for early lung cancer classification, there is need to extract discriminant features of mutated genes present in protein amino acids. But the problem is how to extract useful information in a meaningful way. For this purpose, we have carried out statistical analysis of mutated genes in protein amino acid sequences. This analysis revealed the most frequent mutated genes TP53, EGFR, KMT2D, PDE4DIP, ATM, ZNF521, DICER1, CTNNB1 RUNX1T1, and SMARCA4, etc. The mutation in these genes is the major cause of lung cancer. The discriminant information of mutated genes is important in understanding the cancer driven biological processes. In this study, we have used this discriminant information in order to model the risk of lung cancer in early stages. This information is described using various mathematical formulation-based feature spaces of amino acid composition (AAC), dipeptide composition (DC), split amino acid composition (SAAC), pseudo-amino-acid-composition-series (PseAAC-S), and pseudo-amino-acid-composition-parallel (PseAAC-P). These feature spaces were used to develop various classification models support vector machines (SVM), random forest (RF), Naïve Bayes (NB), and K-nearest neighbor (KNN). Our results demonstrated that SVM model outperformed other classification models. In this study, the main contribution is the exploration of a new dimension for early cancer classification using discriminant information of the mutated genes in protein amino acid sequences. The proposed novel approach would be useful to increase the survival rate of lung cancer patients.

This paper is organized in different sections such that Sect. 2 provides the description of materials and methods, Sect. 3 gives the overview of feature spaces, Sect. 4 explains the development of models, Sect. 5 discusses various performance measures, Sect. 6 observes results and discussion. Finally, Sect. 7 concludes the study.

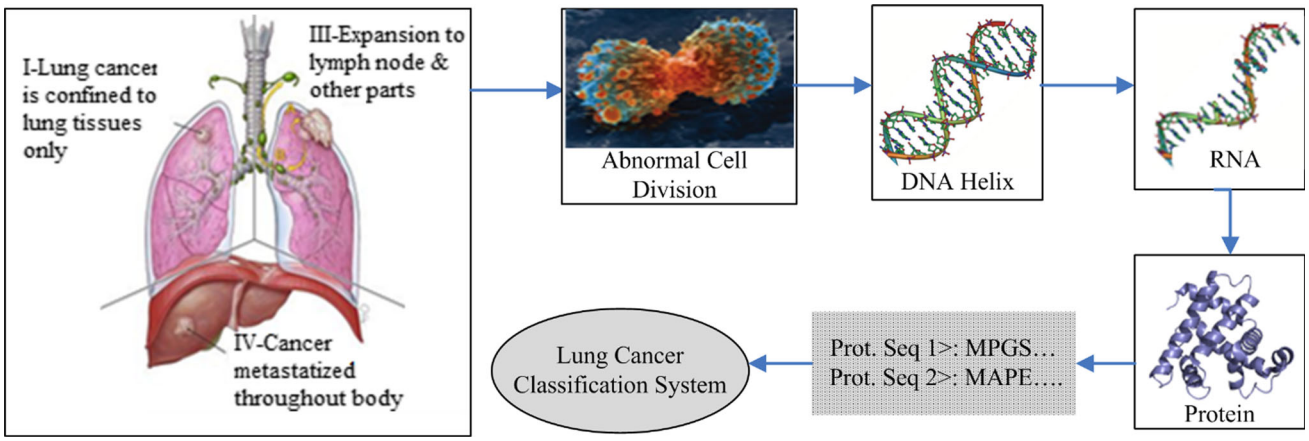

Fig. 1 Different phases of lung cancer development in proposed approach

2 Materials and Methods

Figure 1 depicts different phases of lung carcinogenesis. Lung carcinogenesis is the formation of lung cancer in which lung normal cells are converted to cancer cells as a result of mutagenesis. This causes the transcription of DNA to damaged RNA and replication of abnormal functions. The cancerous protein sequences are formed as a result of translation of this damaged RNA to specific amino acid chain, or polypeptide. In the final stage, LCCMs are developed to classify protein sequences as cancer or non-cancer.

The block diagram of proposed approach is shown in Fig. 2. The primary sequences of amino acids are given to the data formulation/preprocessing module. The preprocessed sequences are split into training data (70%) and testing data (30%). In the next step, various features are computed in different feature spaces. These features are used to develop various lung cancer classification algorithms. These models are evaluated using different performance measures on testing data.

2.1 Description of Dataset Formation

In order to develop lung cancer classification models, a valid updated dataset of somatic mutations in protein amino acids sequence is essential. For this purpose, we explored various proteomic and genomic data sources [10,15,16,29–31] and retrieved protein amino acids sequences related to lung cancer. In the first step, the mutations data were retrieved from COSMIC [15] and TCGA [16] databases. We processed to retain those samples that were confirmed-somatic-variant. The dataset was then filtered using primary site: "lung", primary histology: "carcinoma" terms. The neutral samples were excluded from the dataset. The protein sequences related to lung cancer genes were retrieved from UniProt [10]. The individual protein sequences were validated using

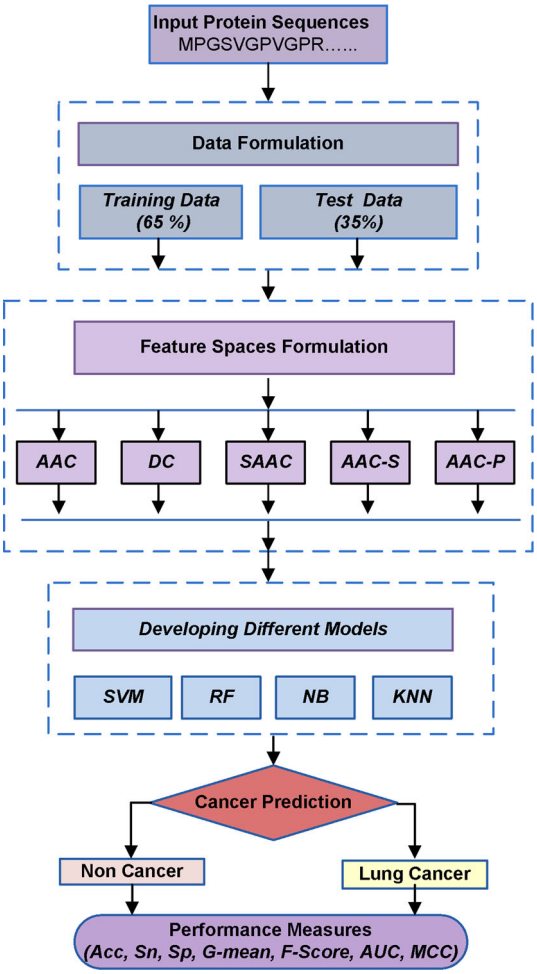

Fig. 2 Proposed lung cancer classification models (LCCMs)

Ensembl [31]. The preprocessed 865 sequences were labeled as cancer samples.

Similarly, normal genes were identified from COSMIC [15] and TCGA [16] databases. The 1800 protein sequences

Table 1 Variation of the most frequent mutated genes using COSMIC [15] and TCGA [16]

Gene name	Variants	f-mutation (%)	Gene name	Variants	f-mutation (%)
TP53	942	26.0006	FBXW7	116	7.4121
EGFR	346	12.5499	NF1	105	6.9814
KMT2D	288	11.6176	PIK3CA	104	7.1576
PDE4DIP	210	09.2962	STK11	99	7.0563
ATM	192	09.0737	NTRK3	96	7.0796
ZNF521	168	08.4379	APC	88	6.7073
DICER1	150	07.9660	PTPRB	83	6.5149
CTNNB1	148	08.2682	BRCA2	82	6.6075
RUNX1T1	126	07.4161	MYH11	82	6.7881
SMARCA4	120	07.3665	AMER1	78	6.6496

related to normal genes were retrieved from UniProt [10]. These sequences were labeled as non-caner samples. The phenotype-gene relationship of the mutated and normal samples was confirmed from OMIM database [29]. The BioMart was used for high-throughput analysis [30]. We selected 865 examples from the set of negative examples to form a balanced dataset. An exemplary cancer and non-cancer protein amino acid sequences in FASTA format are given in the "Supplementary File S1".

These peptide sequences are represented in the text-based FASTA format in which amino acid sequences start with a single line description followed by lines of sequence data. The description line is separated from sequence data using symbol ">" at the beginning. Each protein sequence is composed of twenty protein amino acids. These twenty amino acids are represented by single letter code as: Alanine (A), Cysteine (C), Aspartic Acid (D), Glutamic Acid (E), Phenylalanine (F), Glycine (G), Histidine (H), Isoleucine (I), Lysine (K), Leucine (L), Methionine (M), Asparagine (N), Proline (P), Glutamine (Q), Arginine (R), Serine (S), Threonine (T), Valine (V), Tryptophan (W), and Tyrosine (Y). For example, TP53 protein sequence in FASTA format is represented as follows:

> sp|P04637-M8|P53_HUMAN mutated protein
MFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQH
MTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVE
YLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
S CMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACP
GRDRRTEEENLLKKGEPHHELPPGSTKRALPNNTSSS
PQPKKK RLDGEYFTLQDQTSFQKENC

2.2 Statistical Analysis of Mutated Genes

In order to detect the most frequent mutated genes, we have carried out the statistical analysis of the genes. The mutated genes related to the lung tissues were found from COSMIC [15] and TCGA [16]. The functional information of the most frequent mutated genes is available from UniProt Consortium [10]. In the next step, the mutation frequency of the mutated genes was calculated and ranked according to their mutation frequency. The top most 20 mutated genes of TP53, EGFR, KMT2D, PDE4DIP, ATM, ZNF521, DICER1, CTNNB1, RUNX1T1, SMARCA4, FBXW7, NF1, PIK3CA, STK11, NTRk3, APC, PTPRB, BRCA2, MYH11, and AMER1 were selected. Table 1 indicates the most frequent mutated genes with their mutation frequency.

These genes play important roles to maintain normal functions of cell growth and division [10,15,16]. For example, both TP53 and BRCA2 are identified as tumor suppression genes that encode tumor suppressor proteins to regulate various cell functions [10,15,16]. BRCA2 protein repairs the damaged DNA and maintains the genomic stability through homologous recombination, transcription-coupling, and double-stranded-break repairing [10,15,16]. The EGFR gene makes receptor that binds ligand [10,15,16]. This mechanism allows the cell to receive signals and promotes cell growth, division, and survival [10,15,16]. The protein encoded by KMT2D is responsible for cell differentiation and embryonic development [10,15,16]. It is important in regulating transition metabolism, and tumor suppression. The PDE4DIP gene forms a protein to anchor specific regions in the cell [10,15,16]. The ATM gene produces a protein to regulate protein functions [10,15,16]. The ZNF521 gene is responsible for repression of gene expression [10,15,16]. The DICER1 protein is involved in the cell growth, division (proliferation) and differentiation [10,15,16]. The CTNNB1 protein is responsible to regulate cell growth [10,15,16]. RUNX1T1 gene performs transcriptional repression via its association with DNA-binding transcription factors [10,15,16]. It also functions to recruit other co-repressors and histone-modifying enzymes [10,15,16]. The SMARCA4 protein involves in regulating the genes transcription [10,15,16]. Similarly, FBXW7, NF1, PIK3CA, STK11, NTRk3, APC, PTPRB, MYH11 and AMER1 involve in important functions of ubiquitination, negative regulation of signal transduction pathway,

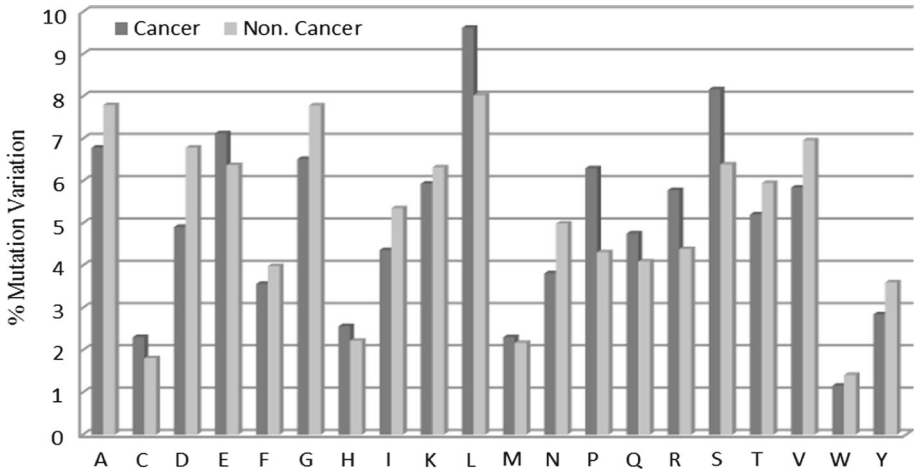

Fig. 3 Variation in lung cancer versus non-cancer sequences

phosphorylation, tumor suppression, cell differentiation, tumor suppression including cell migration, adhesion, mitotic cycle and oncogenic transformation, energy conversion, and un-regulation of transcriptional activation, respectively [10,15,16]. The somatic mutation in these genes causes variation in amino acid composition. These mutated genes mutated genes are shown in Table 1. It is observed that TP53, EGFR, and KMT2D genes have the highest mutation frequencies 26, 13, and 11%, respectively. Their variants are obtained 942, 346, and 288 respectively. The PDE4DIP and ATM genes have their mutation frequencies to 9% with variants 210 and 192, respectively. ZNF521, DICER1, and CTNNB1 genes gave their mutation frequencies 8%. These genes show variants 168, 150, and 148, respectively. Other genes have given relatively lower mutation frequency 7% and their variants are in the range of 78–126. The mutation frequencies information was exploited to compute the discriminant molecular descriptors using various physiochemical properties of amino acids.

We computed the individual variation between cancer and non-cancer with respect to 20 amino acids. To minimize the effect of length variation, we divided the computed variation with the corresponding length of sequence. For each cancer and non-cancer protein sequence, the variation of individual amino acid is computed as follows:

$$\begin{aligned} \mathrm{AAC}_{i,j} &= \frac{\mathrm{counts}(\mathrm{AA}_j)}{\mathrm{LS}_i} * 100, \\ \mathrm{AAC}'_{i,j} &= \frac{\mathrm{counts}(\mathrm{AA}'_j)}{\mathrm{LS}'_i} * 100 \end{aligned} \tag{1}$$

In Eq. 1, $\mathrm{AAC}_{i,j}$ is the amino acid composition (in percent) of ith non-cancer sequence related to jth amino acid and LS_i is the length of that sequence. Similarly, $\mathrm{AAC}'_{i,j}$ is the amino acid composition (in percent) of ith cancer sequence related to jth amino acid and LS'_i is the length of that sequence. AA_j and AA'_j are the amino acid in the non-cancer and cancer sequence, respectively. The overall composition of each amino acid in cancer and non-cancer sequence is computed as follows:

$$\mathrm{AAC}_j = \frac{\sum_{i=1}^{865} \mathrm{AAC}_{i,j}}{N_\mathrm{s}}, \quad \mathrm{AAC}'_j = \frac{\sum_{i=1}^{865} \mathrm{AAC}'_{i,j}}{N'_\mathrm{s}} \tag{2}$$

In Eq. 2, $N_\mathrm{s} = 865$ and $N'_\mathrm{s} = 865$ are the total number of sequences in non-cancer and cancer, respectively. The overall variation between cancer and non-cancer sequences is computed as follows:

$$\text{Total variation} = \sum_{j=1}^{20} \left| AAC_j - AAC'_j \right| \tag{3}$$

In Eq. 3 AAC_j and AAC'_j are the amino acid composition of jth amino acid in non-cancer and cancer protein sequence, respectively. Using Eq. 3, the value of total variation of twenty amino acids was obtained 479. Figure 3 demonstrates the variation of twenty amino acids between cancer and non-cancer sequences. These abnormal mutations hinder the normal function of genes. We have exploited these somatic mutations to form discriminant molecular descriptors using various feature spaces.

3 Formation of Feature Spaces

Table 2 demonstrates five feature spaces of different dimensions. Primary sequences of amino acids determine the structure and functions in the cell. Various amino acids have diverse physical and chemical properties due to variation in the side chains [32]. The physiochemical properties are expressed using different statistical and mathematical formulations. The numerical values of these descriptors give the discrimination power between cancer and non-caner amino acid molecules. These properties are more discriminant as

Table 2 Dimension of different feature spaces

Feature spaces	Dimension (D)
Dipeptide composition (DC)	400
Split amino acid composition (SAAC)	60
Pseudo-amino-acid-composition-series (PseAAC-S)	60
Pseudo-amino-acid-composition-parallel (PseAAC-P)	40
Amino acid composition (AAC)	20

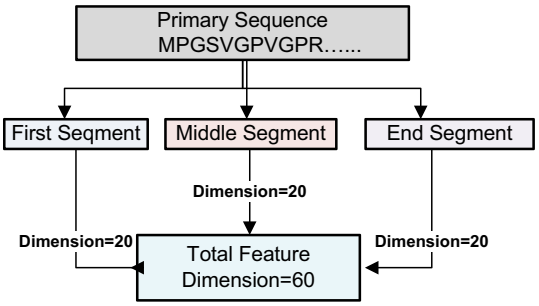

Fig. 4 Split amino acid composition

compared to other physiochemical properties of polarity, solubility, melting point and chemical reactions etc. These properties are expressed in different feature spaces of AAC, DC, SAAC, PseAAC-P, and PseAAC-S. The brief description about various feature spaces is given below.

3.1 Dipeptide Composition (DC)

The dipeptide composition is in the form of fractions of 400 dipeptides and these components are calculated in Eq. 4 as:

$$\mathrm{DC}_i = \frac{\mathrm{DC}_{\mathrm{total}}(i)}{400} \tag{4}$$

DC_i is the ith dipeptide of 400 dipeptides, $i = 1, 2, 3, \ldots,$ 400. This feature space in vector form is represented in Eq. 5 as:

$$X_{\mathrm{DC}} = [\mathrm{DC}_1, \mathrm{DC}_2, \mathrm{DC}_3, \ldots, \mathrm{DC}_{400}]^{\mathrm{T}} \tag{5}$$

3.2 Split Amino Acid Composition (SAAC)

In SAAC features the primary protein sequence is split into three parts, namely N-terminus, internal terminus and C-terminus [33]. Figure 4 depicts the formation of split amino acid composition by splitting the original primary protein sequence into three mentioned parts. The amino acid composition of each part is calculated separately and concatenated to form the feature vector of 60-dimensional (60D).

3.3 Pseudo-Amino-Acid Composition in Series (PseAAC-S)

PseAAC-S features represent both the compositional and positional effects of primary protein sequences [34]. The composition is calculated by pairwise relationships of chemical properties of hydrophobicity and hydrophilicity to unfold the various characteristics of protein amino acid sequences [35]. In this composition, a protein sequence is transformed into feature vector of $20 + i * t$ dimensions. The protein feature vector in series composition is represented in Eq. 6 as:

$$X_{\mathrm{s}} = \left[S_{\mathrm{C}_1} \ldots S_{\mathrm{C}_{20}} S_{\mathrm{C}_{21}} \ldots S_{\mathrm{C}_{20+t}} S_{\mathrm{C}_{20+t+1}} \ldots S_{\mathrm{C}_{20+2t}}\right]^{\mathrm{T}} \tag{6}$$

$$S_{C_r} = \begin{cases} \frac{f_r}{\sum_{i=1}^{20} a_i + .05 \sum_{j=1}^{2t} T_j} & \text{for } 1 < r < 20 \\ \frac{.05 T_r}{\sum_{i=1}^{20} a_i + .05 \sum_{j=1}^{2t} T_j} & \text{for } 21 < r < 20 + 2t \end{cases} \tag{7}$$

where a_i is the normalized proportion of amino acids in a protein sequence and T_j is the ordered correlation factor [27], that is based on numerical values of the hydrophilicity and hydrophobicity properties. The value of T_j is computed using Eqs. 8 and 9 as follows:

$$T_{2t-1} = \frac{1}{L-t} \sum_{i=1}^{L-t} H_{d\,i,i+t} \tag{8}$$

$$T_{2t} = \frac{1}{L-t} \sum_{i=1}^{L-t} H_{b\,i,i+t} \tag{9}$$

3.4 Pseudo-Amino-Acid Composition in Parallel (PseAAC-P)

For a given sequence, PseAAC-P features are computed using pairwise relationships related to chemical properties of hydrophobicity and hydrophilicity. In PseAAC-P, a protein sequence is expressed into feature vectors each of 40 dimensions. The individual components Ψ_t is calculated using the sequence order correlation as follows

$$\Psi_t = \frac{1}{L-t} \sum_{i=1}^{L-t} \phi\left(A_i, A_{i+t}\right) \tag{10}$$

The value of $\phi\left(A_i, A_{i+t}\right)$ is calculated as follows:

$$\phi\left(A_i, A_{i+t}\right) = \frac{1}{3}\left\{\left[H_d\left(A_j\right) - H_d\left(A_i\right)\right]^2 + \left[H_b\left(A_j\right) - H_b\left(A_i\right)\right]^2\right\} \tag{11}$$

In Eq. 11, H_d and H_b represents the numerical values of the hydrophilicity and hydrophobicity properties.

Table 3 Optimal parameter values of learning algorithms

Sr. no.	Learning models	Optimal parameter values
1	SVM with Poly and RBF kernels	$C = 2048$ and degree = 2 for Poly, $C = 256$ and gamma = 16 for RBF
2	RF	Num-trees = 200, criteria = GINI
3	KNN	Nearest neighbor = 7
4	NB	Gaussian function with prior probability $P(\text{cancer}) = 0.5$, $P(\text{non-cancer}) = 0.5$

3.5 Amino Acid Composition (AAC)

AAC features are computed using protein sequence of lung cancer or non-cancer. The dimension of each sequence shows the occurrence frequency of individual amino acids. The percentage occurrence of $\text{AAC}_{i,j}$ of the amino acid i in the jth protein is computed for amino acids in protein sequence using Eq. 12.

$$\text{AAC}_{i,j} = \frac{n_{i,j}}{n_{\text{aa}_j}} * 100 \tag{12}$$

where $n_{i,j}$ is the number of amino acids of type i observed to be present in protein j and n_{aa_j} is the total number of amino acids in protein j. In AAC features dataset, the jth protein sequence is expressed as 20-dimensional (20D) feature vector, in Eq. 13.

$$X_j = \left[\text{AAC}_{1,j}, \text{AAC}_{2,j}, \ldots, \text{AAC}_{20,j}\right]^{\text{T}} \tag{13}$$

In Eq. 13, $\text{AAC}_{1,j}, \text{AAC}_{2,j}, \ldots, \text{AAC}_{20,j}$, represent the percent composition of amino acids.

4 Development of Classification Models

Lung cancer classification models are developed using diverse learning algorithms SVM, RF, KNN, and NB. The theoretical background of these learners is available in the literature of machine learning [36–39]. Here, we have mainly focused on implementation details. These algorithms were implemented using Python 3.5 software (Python Software Foundation, https://www.python.org/) [40]. The optimal parameters values of these learners were computed using Grid Search. The optimal tuned parameters of five learning algorithms are given in Table 3.

4.1 SVM Algorithm

SVM performs classification by constructing a hyperplane that has maximum margin between two closest points [39]. The block diagram of SVM model is shown in Fig. 5. A set of 1210 training examples, $S = \{(X_i, t_i)\}_{i=1}^{1210}$ is given to SVM algorithm that finds optimal hyperplane according to input data distribution between two classes. Here, $t_i \in \{0, 1\}$ and $X_i \in \left\{\Re^{400}, \Re^{60}, \Re^{60}, \Re^{40}, \Re^{20}\right\}$ for DC, SAAC, PseAAC-S, PseAAC-P and AAC feature spaces, respectively. The decision surface for SVM was defined as follows:

$$Y(X) = \sum_{i=1}^{n} \alpha_i t_i X_i^{\text{T}}.X + \text{bias} \tag{14}$$

Here α_i is the Lagrange multiplier. The data samples X_i correspond to $\alpha_i > 0$ are called support vectors. For the classification of non-separable data, the solution to the objective function is defined as:

$$\Psi(W, \zeta) = 0.5W^{\text{T}}W + C\sum_{i=1}^{n} \zeta_i \tag{15}$$

Subject to condition $t_i(W^{\text{T}}\psi(X_i) + \text{bias}) \geq 1 - \zeta_i$, where C is penalty factor for the error term $\sum_{i=1}^{n} \zeta_i$ and $\psi(X)$ is nonlinear mapping. The nonlinear decision boundary is now defined by:

$$Y(X) = \sum_{i=1}^{s_v} \alpha_i t_i K(X_i, X) + \text{bias}$$
$$\text{where } K(X_i, X) = \Psi(X_i)^{\text{T}}.\Psi(X) \tag{16}$$

The parameter C defines the trades off between the misclassification of input training instances and complexity of the decision surface. A low value of this parameter makes the decision boundary smooth. On the other hand, a high value tailors the boundary of decision surface across the input training data by selecting more training examples as support vectors. As a result, the trained model may not perform well on unseen data. However, the lower value of RBF kernel parameter gamma (σ) parameter indicates the higher influence of the neighbor training examples. For Polynomial kernel, $K(X_i, X) = (X_i.X + 1)^d$, the optimal values of $C = 2048$ and degree $= 2$ were found using grid search. For RBF kernel, $K(X_i, X) = \exp(-\|X_i - X\|^2/(2\sigma^2))$, $\sigma = 16$ and C $= 256$ are the best parameters.

4.2 Random Forest (RF)

This algorithm generates different decision trees and then combines outputs for final decision. Due to good

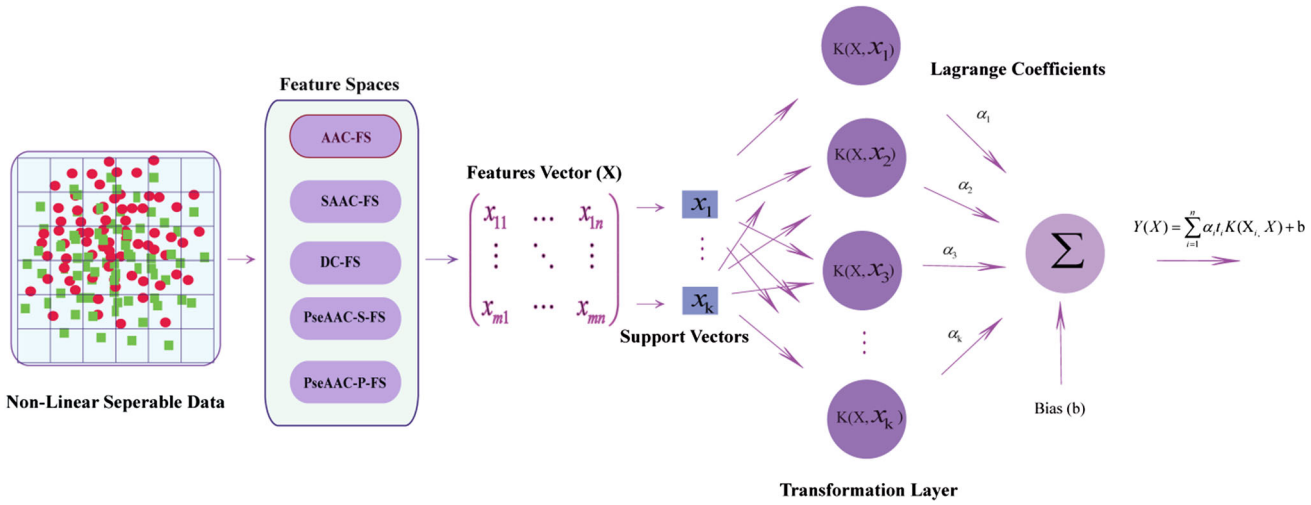

Fig. 5 Support vector machine (SVM) model

performance and good generalization on high-dimension input dataset, we have employed this algorithm for lung cancer classification. These trees are constructed using training dataset $S_{tr} = \{(X_i, t_i)\}_{i=1}^{1210}$. The remaining testing dataset, $S_{ts} = \{(X_i, t_i)\}_{i=1211}^{1730}$ is kept for model evaluation. For best node split, we choose $m_f = \sqrt{D}$ features randomly from the D-dimensional input feature vector, $D \in \left\{\sqrt{\Re^{400}}, \sqrt{\Re^{60}}, \sqrt{\Re^{60}}, \sqrt{\Re^{40}}, \sqrt{\Re^{20}}\right\}$. The Gini-fitness criteria, $G(d, u_i) = \sum_{i=1}^{c} \frac{a_i}{n_{fv}} I_G(d_{ui})$, has computed the best node split to evaluate the importance of each feature. Here, c represents the number of children at node d and n_{fv} shows the number of feature vectors selected for training. The Gini impurity $I_G(d_{ui})$ gives class label distribution. For a feature variable $u_i \in U$ with c values at node d, $u_i = \{v_1, v_2, \ldots v_p\}$, the value of $I_G(d_{vi})$is computed as: $I_G(d_{vi}) = 1 - \sum_{i=0}^{c_i} \left(\frac{n_{ci}}{a_i}\right)^2$, where n_{ci} is the number of samples with values, v_i belong to class c_i and a_i indicates the number of samples with the value v_i at node d. The testing data, $S_{ts} = \{(X_i, t_i)\}_{i=1211}^{1730}$, were used to evaluate the RF model classification performance. During implementation, we found the optimal number of random trees $n_t = 200$.

4.3 *K*-Nearest Neighbor (KNN) Algorithm

This is instance-based learning algorithm. It does not explicitly model the complexity of the input data. It tries to memorize the input training examples, $S_{tr} = \{(X_i, t_i)\}_{i=1}^{1210}$ to extract knowledge for classification. In KNN classification, an input instance is classified according to distance function and the majority of k-nearest neighbors [39]. We have used Euclidean distance to measure distance between examples. The Euclidean distance $d(X_i, X_j)$ between two vectors X_i and X_j each of m-dimension is calculated as follows:

$$d(X_i, X_j) = \sqrt{(x_{i,1} - x_{j,1})^2 (x_{i,2} - x_{j,2})^2 + (x_{i,3} - x_{j,3})^2 + \cdots (x_{i,m} - x_{j,m})^2} \tag{17}$$

The input example is assigned to the class which has more class neighbors. During implementation, we found the optimal value of $K = 7$.

4.4 Naïve Bayes (NB) Algorithm

This learning algorithm is modeled by applying Bayes rules with the assumption of independent attributes/features. The features were denoted by $X_{fi} \in \left\{\Re^{400}, \Re^{60}, \Re^{60}, \Re^{40}, \Re^{20}\right\}$ for five feature spaces. An instance is assigned to the class of highest posterior probability. We used the Gaussian function with equal prior probability $P(X_f) = 0.5$ to train the model as follows:

$$P(X_{f1}, X_{f2}, \ldots, X_{fn}|c) = \prod_{i=1}^{n} P(X_{fi}|c) \tag{18}$$

$$P(X_f|c_i) = \frac{P(c_i|X_f)P(X_f)}{P(c_i)} \quad c \in \{\text{cancer, non-caner}\} \tag{19}$$

The testing data is classified according to probability of association as follows:

$$c_{nb} = \text{argmax } P(c_k) \prod_{i=1}^{n} P(X_{fi}|c_k), \text{ for } k = 1, 2 \tag{20}$$

In Eq. 20, $c_i \in \{\text{cancer, non-caner}\}$.

5 Performance Measures

The performance of the classification models is evaluated using various measures of accuracy (Acc), sensitivity (Sn), specificity (Sp), *G*-mean, *F*-measure, MCC, ROC, and AUC. These measures are commonly used for reporting the performance of machine learning algorithms. These measures evaluate classification performance in different aspects [41].

$$\text{Acc} = \left(\frac{\text{TP}+\text{TN}}{n}\right) \quad \text{Sn} = \left(\frac{\text{TP}}{\text{TP}+\text{FN}}\right)$$

$$\text{Sp} = \left(\frac{\text{TN}}{\text{TN}+\text{FP}}\right),$$

$$\text{Pre} = \left(\frac{\text{TP}}{\text{TP}+\text{FP}}\right) \quad G_{\text{mean}} = \sqrt{Sn * \text{Pre}},$$

$$F_{\text{measure}} = 2\frac{\text{Pre} * \text{Sn}}{(\text{Pre}+\text{Sn})}, \quad \text{FPR} = 1 - \text{Sp}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{[\text{TP}+\text{FP}][\text{TP}+\text{FN}][\text{TN}+\text{FP}][\text{TN}+\text{FN}]}}$$

The values of Sn, Sp, Pre, *G*-mean, *F*-measure, and MCC measures are based on the computed values of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Further, the classification performance was also analyzed using measures of ROC curve [42]. The ROC curve shows the association between sensitivity and false positive rate (FPR). For each threshold value, a point in the form of (FPR, Sn) is plotted and corresponding ROC curve is obtained by connecting all such points. ROC curve close to the top-left corner indicates that classifier has a good performance. To measure the performance of the model as a single value of area under the curve (AUC) is computed.

6 Results and Discussion

This section describes the performance of various models in terms of accuracy, sensitivity, specificity, *G*-mean, *F*-score, AUC, and MCC. The comparative performance of the proposed models was carried out with previous approaches.

6.1 Performance of Proposed Models

Table 4 depicts the performance of classification models in different feature spaces. It is observed that, in SAAC feature space, SVM-Poly, SVM-RBF, RF, KNN, and NB models have yielded accuracy values of 0.9938, 0.9943, 0.9850, 0.9871, and 0.9684 respectively. Overall, classification models have obtained higher performance values in SAAC feature space. However, SVM models have given comparatively better performance. This is because, in SACC feature space, margin-based SVM model has higher generalization capability than RF, KNN and NB models to classify cancer protein sequences from non-cancer. Further, SVM-RBF and SVM-Poly have shown higher accuracy values than other models. This is because, in SAAC feature space, both SVM-RBF and SVM-Poly models have generated better decision surface than that of other classification models. Probabilistic NB model has obtained relatively lower accuracy for this feature space. This is because, on the complex binary class problem, NB model has generated less discriminant data distribution boundary than other models. On the same feature space, we observed improved sensitivity values of 0.9922, 0.9932, 0.9792, 0.9822, and 0.9756 for SVM-Poly, SVM-RBF, RF, KNN, and NB, respectively. Here, SVM-RBF and SVM-Poly models have retained the highest sensitivity values of 0.9932 and 0.9922, respectively. Therefore, SVM has developed more discriminant and generalized model in SAAC feature space. It differentiated the complexity of lung cancer dataset effectively. In this case, SVM-RBF has performed better than SVM-Poly. This is because proximity (distance)-based SVM-RBF has developed better discriminant boundary than SVM-Poly. In SAAC feature space SVM-Poly, SVM-RBF, RF, KNN, and NB models provided improved Sp values of 0.9954, 0.9954, 0.9905, 0.9919, and 0.9613, respectively. SVM-RBF and SVM-Poly models have better Sp value (0.9954) than other models. Further, we observed that in SAAC feature space, SVM-RBF model also outperformed other models using *G*-mean, *F*-score, AUC, and MCC quality measures.

In DC feature space, we observed that SVM-Poly, SVM-RBF, RF, KNN, and NB models have given higher sensitivity values of 0.9759, 0.9834, 0.9894, 0.9909, and 0.9766 respectively. KNN model has obtained the highest sensitivity values. On the other hand, SVM-Poly model has relatively lower sensitivity value. Therefore, to classify cancer patients with high sensitivity, instance-based KNN model is a better choice in DC feature space. Further, in PseAAC-S feature space, KNN model has values of 0.9616, 0.9588, 0.9644, 0.9610, 0.9608, 0.9827, and 0.9239, for Acc, Sn, Sp, *G*-mean, *F*-score, AUC, and MCC measures, respectively. In PseAAC-S feature space, KNN model again outperformed other models for these performance measures. In PseAAC-P feature space, we found SVM-RBF outperformed other models in terms of these measures. Overall, it was observed that SVM model has effectively exploited the genes mutation information in different feature spaces of protein amino acid sequences. The overall performance comparison of classification models is visually demonstrated in Fig. 6.

Figure 7 demonstrates ROC performance curves of SVM-RBF and SVM-Poly models in AAC, SAAC, PseAAC-P, and PseAAC-S feature spaces. Figure 7a–d shows the performance curves for SVM-Poly model using AAC, SAAC, PseAAC-P, and PseAAC-S feature spaces, respectively. Figure 7e–h shows the ROC curves for SVM-RBF using AAC,

Table 4 Performance comparison of proposed models using different measures

Input dataset	Proposed models	Acc	Sn	Sp	*G*-mean	F-Score	AUC	MCC
SVM-Poly	SVM_{AAC}	0.8187	0.8261	0.8115	0.8191	0.8188	0.8935	0.6385
	SVM_{SAAC}	0.9938	0.9922	0.9954	0.9938	0.9938	0.9997	0.9877
	SVM_{DC}	0.9707	0.9759	0.9656	0.9707	0.9706	0.9928	0.9418
	$SVM_{PseAAC\text{-}P}$	0.9201	0.9244	0.9160	0.9204	0.9200	0.9640	0.8417
	$SVM_{PseAAC\text{-}S}$	0.9301	0.9290	0.9312	0.9295	0.9290	0.9695	0.8616
SVM-RBF	SVM_{AAC}	0.8789	0.8835	0.8744	0.8791	0.8789	0.9290	0.7585
	SVM_{SAAC}	0.9943	0.9932	0.9954	0.9943	0.9943	0.9997	0.9888
	SVM_{DC}	0.9794	0.9834	0.9756	0.9794	0.9794	0.9952	0.9592
	$SVM_{PseAAC\text{-}P}$	0.9411	0.9444	0.9379	0.9413	0.9410	0.9755	0.8833
	$SVM_{PseAAC\text{-}S}$	0.9534	0.9527	0.9541	0.9530	0.9527	0.9797	0.9077
	$AdB_{PseAAC\text{-}S}$	0.9463	0.9423	0.9502	0.9454	0.9451	0.9757	0.8762
RF	RF_{AAC}	0.9092	0.9080	0.9105	0.9086	0.9084	0.9476	0.8190
	RF_{SAAC}	0.9850	0.9792	0.9905	0.9845	0.9844	0.9962	0.9704
	RF_{DC}	0.9810	0.9894	0.9728	0.9814	0.9812	0.9966	0.9627
	$RF_{PseAAC\text{-}P}$	0.9155	0.8632	0.9667	0.9001	0.8885	0.9732	0.8475
	$RF_{PseAAC\text{-}S}$	0.9552	0.9519	0.9585	0.9545	0.9542	0.9798	0.9112
KNN	KNN_{AAC}	0.9222	0.9211	0.9233	0.9217	0.9215	0.9551	0.8449
	KNN_{SAAC}	0.9871	0.9822	0.9919	0.9867	0.9866	0.9967	0.9746
	KNN_{DC}	0.9837	0.9909	0.9767	0.9840	0.9839	0.9971	0.9680
	$KNN_{PseAAC\text{-}P}$	0.9276	0.8828	0.9714	0.9144	0.9044	0.9770	0.8693
	$KNN_{PseAAC\text{-}S}$	0.9616	0.9588	0.9644	0.9610	0.9608	0.9827	0.9239
	$ET_{PseAAC\text{-}S}$	0.9664	0.9639	0.9689	0.9659	0.9657	0.9848	0.9334
NB	NB_{AAC}	0.9200	0.9218	0.9183	0.9199	0.9198	0.9456	0.8406
	NB_{SAAC}	0.9684	0.9756	0.9613	0.9704	0.9698	0.9759	0.9387
	NB_{DC}	0.9699	0.9766	0.9634	0.9702	0.9701	0.9804	0.9405
	$NB_{PseAAC\text{-}P}$	0.9238	0.8943	0.9527	0.9144	0.9065	0.9623	0.8591
	$NB_{PseAAC\text{-}S}$	0.9557	0.9537	0.9576	0.9551	0.9549	0.9720	0.9120

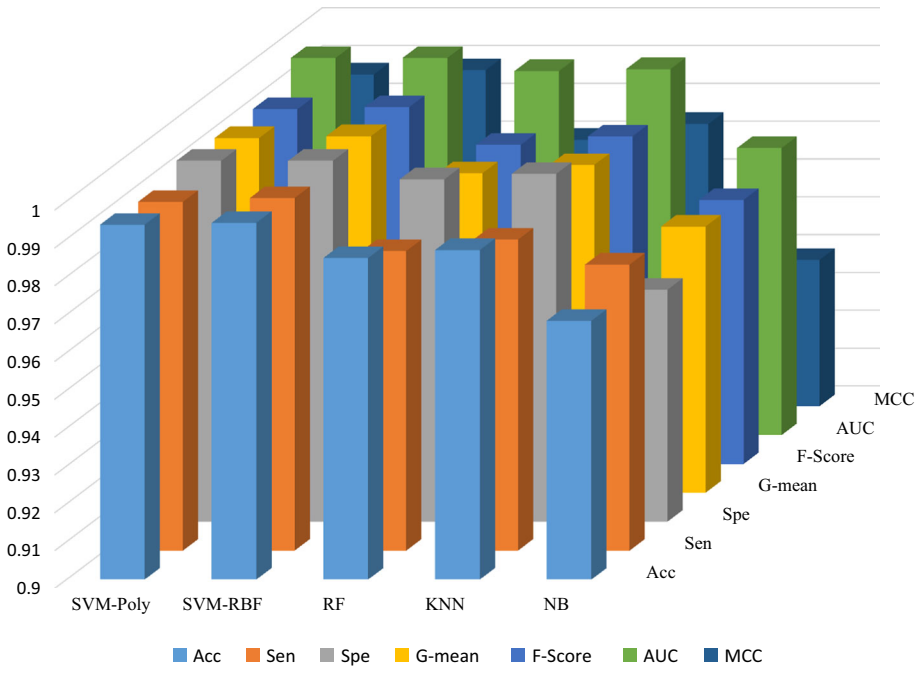

Fig. 6 Performance comparison of proposed models in SAAC feature space

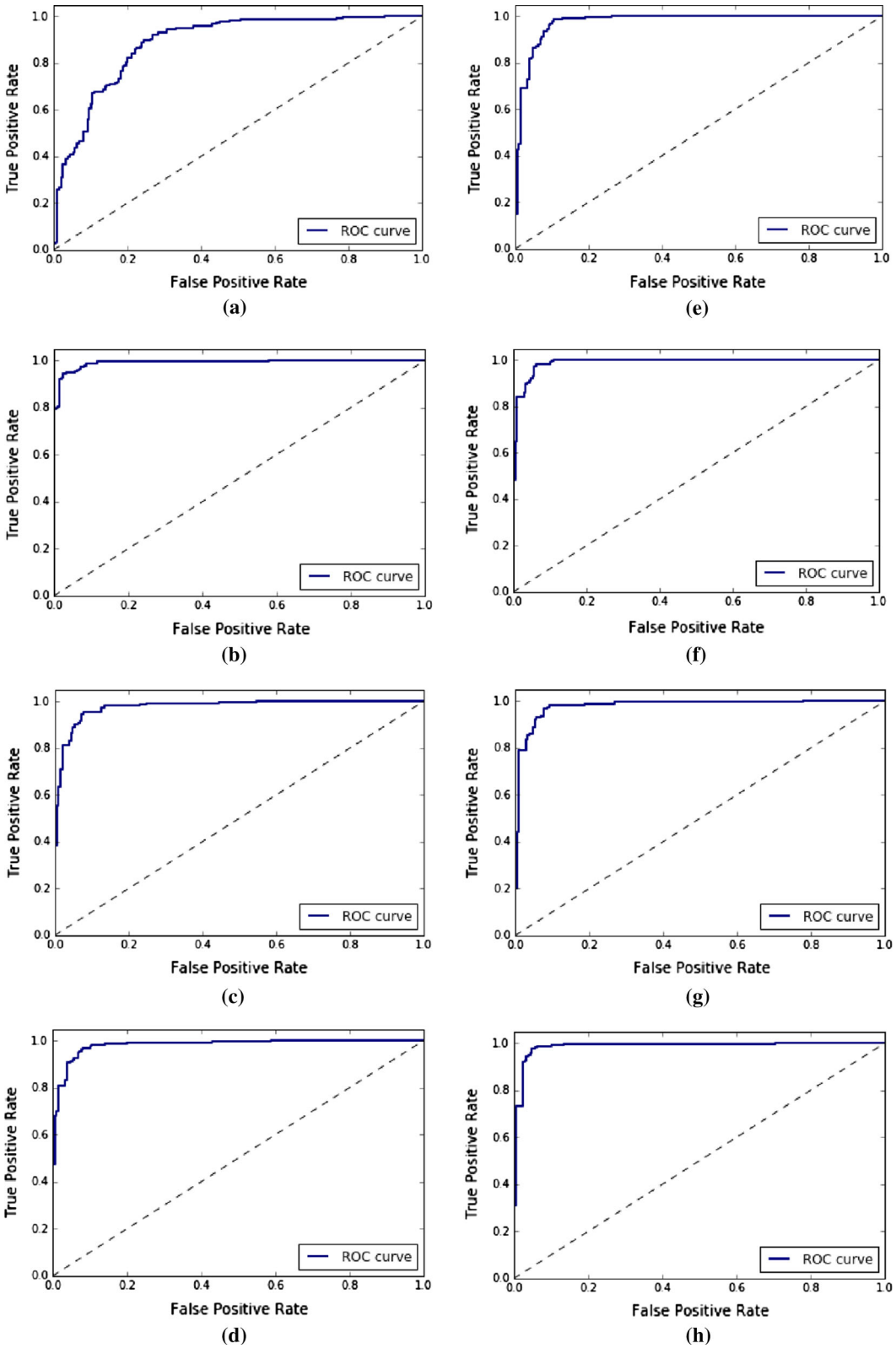

Fig. 7 ROC performance curves of the best performing SVM models in different feature spaces. **a** ROC curve with SVM-Poly using AAC, **b** ROC curve with SVM-Poly using SAAC, **c** ROC curve with SVM-Poly using PseAAC-P, **d** ROC curve with SVM-Poly using PseAAC-S, **e** ROC curve with SVM-RBF using AAC, **f** ROC curve with SVM-RBF using SAAC, **g** ROC curve with SVM-RBF using PseAAC-P, **h** ROC curve with SVM-RBF using PseAAC-S

SAAC, PseAAC-P, and PseAAC-S feature spaces, respectively. Due to more generalized data modeling capability of RBF kernel than polynomial kernel, SVM-RBF has better AUC of ROC curve.

6.2 Temporal Cost Comparison

The time complexity (cost) of learning algorithms is computed empirically. The execution time of learning algorithms depends on the software environment and the computing machine. We reported the results using DELL Optiplex-7040 computing machine with Processor: Intel(R) Core(TM), i7-6700 CPU @3.40 GHz (4 CPUs), 3.4 GHz, Memory: 32,768 MB RAM machine in Windows 10 operating system. Table 5 depicts the training and testing time of various learners in different feature spaces. The table shows that NB algorithm consumed the least training time. However, RF algorithm taken maximum training time. This is because RF learner is based on the construction of a large number of random trees. The final decision is combined using the individual decisions of various constructed trees. After model development, the temporal cost of the models is evaluated for testing examples. NB algorithm consumed the lowest average execution time 0.158 s on testing examples. KNN obtained the highest average time 1.2131 s. However, RF and NB algorithms have consumed approximately the same average execution times 0.164 and 0.158 s, respectively. On the other hand, in testing phase, SVM-Poly and SVM-RBF algorithms consumed average time 0.596 and 0.713 s, respectively. Although RF obtained the maximum training time, but once model is trained it took least time as compared to other learners. In general, learning algorithms consumed relatively large training time. However, after successful model development, they consumed less time during testing phase for lung cancer classification.

Table 5 Execution time of proposed classification models

Proposed models	Training time/s	Testing time/s
SVM-Poly		
SVM_{AAC}	10.608	0.1950
SVM_{SAAC}	4.1307	0.1206
SVM_{DC}	43.465	2.3967
$SVM_{PseAAC\text{-}P}$	14.308	0.1103
$SVM_{PseAAC\text{-}S}$	6.4872	0.1604
SVM-RBF		
SVM_{AAC}	6.5374	0.2004
SVM_{SAAC}	6.3268	0.2907
SVM_{DC}	57.332	2.4264
$SVM_{PseAAC\text{-}P}$	8.3020	0.2606
$SVM_{PseAAC\text{-}S}$	11.794	0.3910
RF		
RF_{AAC}	67.980	0.1504
RF_{SAAC}	82.218	0.1503
RF_{DC}	99.556	0.1602
$RF_{PseAAC\text{-}P}$	80.213	0.1905
$RF_{PseAAC\text{-}S}$	96.847	0.1503
KNN		
KNN_{AAC}	2.6871	0.1804
KNN_{SAAC}	4.3014	0.5715
KNN_{DC}	14.987	4.4217
$KNN_{PseAAC\text{-}P}$	3.2587	0.3509
$KNN_{PseAAC\text{-}S}$	3.7900	0.5414
NB		
NB_{AAC}	1.5942	0.1807
NB_{SAAC}	1.0731	0.1701
NB_{DC}	1.9752	0.1804
$NB_{PseAAC\text{-}P}$	1.7296	0.2104
$NB_{PseAAC\text{-}S}$	1.1129	0.05013

6.3 Performance Comparison with Previous Approaches

Researchers have employed different data modalities and feature selection approaches for lung cancer classification [18–23,43,44]. Each approach has its own merits and demerits depending upon complexity of data modalities and feature selection. We compared the performance of proposed models with previous conventional approaches for lung cancer classification. This comparison would be informative to analyze the useful information related to lung cancer classification problem.

Table 6 depicts the performance of proposed classification models with previous models. Different classification models of Bayesian Network, random forest, nearest neighbor, SVM, and random committee have reported different accuracy values of 0.85, 0.79, 0.70, 0.76, and 0.77 respectively [18]. In another study [19], SVM algorithm-based models of SVM-Linear, SVM-Evolutionary, SVM-Lib, SVM-POS, SVM-Fast Large Margin, and SVM-Hyper have been developed. These models obtained values of accuracy in range of 0.51–0.82, 0.51–0.82, 0.47–0.65, 0.52–0.67, 0.43–0.56, 0.34–0.78, and 0.28–0.42, respectively. Further, in another study [20], Gaussian SVM, Linear SVM, Logistic regression, Naïve Bayes, and RF models have achieved accuracy values in range of 0.83–0.85, 0.80–0.82, 0.80–0.82, 0.75–0.77, and 0.86 respectively. In [23], reported model has yielded AUC (0.8600) for the classification of cancer and non-cancer. On the other hand, our proposed classification models SVM-Poly, SVM-RBF, RF, KNN, and NB have obtained the highest accuracy values of 0.9938, 0.9943, 0.9850, 0.9871, and 0.9684 respectively. From this analysis, we summarized that proposed models have outperformed previous models

Table 6 Performance comparison of proposed models with previous approaches

Feature extraction strategy	Methods	Accuracy
Structural and physicochemical properties [18]	Bayesian Network	0.8500
	Random forest	0.72–0.79
	Nearest neighbor	0.69–0.70
	SVM	0.7600
	Random committee	0.69–0.77
Structural and physicochemical attributes with attribute weighting models [19]	SVM	0.51–0.82
	SVM-Linear	0.51–0.82
	SVM-Evolutionary	0.47–0.65
	SVM-Lib	0.52–0.67
	SVM-POS	0.43–0.56
	SVM-Fast Large Margin	0.34–0.78
	SVM-Hyper	0.28–0.42
CpG methylation, histone H3 methylation modification, nucleotide composition, and conservation [20]	Gaussian SVM	0.83–0.85
	Linear SVM	0.80–0.82
	Logistic regression	0.80–0.82
	Naïve Bayes	0.75–0.77
	Random forest	0.8600
Radiographic signatures with clinical data [23]	Radiographic and clinical model	0.8600
Composition and physiochemical(Proposed models) properties	SVM-Poly, SVM-RBF	0.9938, 0.9943
	Random forest	0.9850
	K-nearest neighbor	0.9871
	Naïve Bayes	0.9684

for lung cancer classification. This is because our models have exploited effectively discriminant molecular descriptors using physiochemical properties of protein amino acids. In future, we intend to classify other types of cancers using influential features of differentially expressed genes. This information related to differentially expressed genes can be retrieved from the literature [45,46].

7 Conclusion

In this study, for early lung cancer classification, we explored a new dimension of using discriminant information of mutated genes revealed through the statistical analysis of protein amino acid sequences. From this analysis, we found twenty most frequent mutated genes TP53, EGFR, KMT2D, PDE4DIP, ATM, ZNF521, DICER1, CTNNB1 RUNX1T1, SMARCA4, FBXW7, NF1, PIK3CA, STK11, NTRk3, APC, PTPRB, BRCA2, MYH11 and AMER1. The abnormal mutation in these genes is major cause of lung cancer. We have developed several lung cancer classification models using discriminant information of mutated genes expressed in different feature spaces. Our results highlight that SVM and RBF models have the best performance in SAAC feature space. The proposed models have demonstrated improved performance as compared to other approaches. This is because the proposed models have effectively exploited the discriminant information related to cancer and non-cancer protein amino acid sequences. The proposed approach would be effective to increase the survival rate of lung cancer patients. It is anticipated that proposed model would be useful for academia, researchers, and practitioners in decision making for cancer diagnosis, prognosis, precision medicine, and drug discovery.

Acknowledgements This work is supported by HEC, Government of Pakistan under Indigenous Ph.D. Fellowship for 5000 scholars, Phase-II (PIN #. 213-59474-2PS2-056). Authors are also thankful to Pakistan Institute of Engineering and Applied Sciences (PIEAS) for providing resources.

Compliance with Ethical Standards

Conflict of interest The authors, Mohsin Sattar and Abdul Majid, declare that they have no conflict of interest.

Human and Animal Rights This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of

1975, as revised in 2008 (5). Additional informed consent was obtained from all patients for which identifying information is included in this article.

References

1. Torre, L.A.; Siegel, R.L.; Ward, E.M.; Jemal, A.: Global cancer incidence and mortality rates and trends: an update. Cancer Epidemiol. Biomark. Prev. **25**(1), 16–27 (2016)
2. Stoppler, M.C.: Lung cancer facts. https://www.medicinenet.com/lung_cancer/article.htm#lung_cancer_facts. Accessed 10 Jan 2018
3. Stoppler, M.C.: Causes of lung cancer in non-smokers. https://www.medicinenet.com/script/main/art.asp?articlekey=53012. Accessed 11 Jan. 2018
4. Siegel, R.L.; Miller, K.D.; Jemal, A.: Cancer statistics, 2018. CA Cancer J. Clin. **68**(1), 7–30 (2018)
5. Luqman, M.; Javed, M.M.; Daud, S.; Raheem, N.; Ahmad, J.; Khan, A.-U.-H.: Risk factors for lung cancer in the Pakistani population. Asia Pac. J. Cancer Prev. **15**(7), 3035–3039 (2014)
6. Gilad, S.; Lithwick-Yanai, G.; Barshack, I.; Benjamin, S.; Krivitsky, I.; Edmonston, T.B.; Bibbo, M.; Thurm, C.; Horowitz, L.; Huang, Y.; Feinmesser, M.; Steve Hou, J.; Cyr, B.; Burnstein, I.; Gibori, H.; Dromi, N.; Sanden, M.; Kushnir, M.; Aharonov, R.: Classification of the four main types of lung cancer using a microRNA-based diagnostic assay. J. Mol. Diagn. **14**(5), 510–517 (2012)
7. Lee, K.J.; Lee, J.H.; Chung, H.K.; Choi, J.; Park, J.; Park, S.S.; Ju, E.J.; Park, J.; Shin, S.H.; Park, H.J.; Ko, E.J.; Suh, N.; Kim, I.; Hwang, J.J.; Song, S.Y.; Jeong, S.-Y.; Choi, E.K.: Novel peptides functionally targeting in vivo human lung cancer discovered by in vivo peptide displayed phage screening. Amino Acids **47**(2), 281–289 (2015)
8. Cheung, C.H.Y.; Juan, H.: Quantitative proteomics in lung cancer. J. Biomed. Sci. **24**(1), 37–47 (2017)
9. Detterbeck, F.C.; Boffa, D.J.; Kim, A.W.; Tanoue, L.T.: The eighth edition lung cancer stage classification. Chest **151**(1), 193–203 (2017)
10. Consortium, T.U.: UniProt: the universal protein knowledgebase. Nucleic Acids Res. **45**(D1), D158–D169 (2017)
11. Fraser, A.: Essential human genes. Cell Syst. **1**(6), 381–382 (2015)
12. Dela-Cruz, C.S.; Tanoue, L.T.; Matthay, R.A.: Lung cancer: epidemiology, etiology, and prevention. Clin. Chest Med. **32**(4), 605–644 (2011)
13. Ho, V.; Parent, M.-E.; Pintos, J.; Abrahamowicz, M.; Danieli, C.; Richardson, L.; Bourbonnais, R.; Gauvin, L.; Siemiatycki, J.; Koushik, A.: Physical activity and lung cancer risk in men and women. Cancer Causes Control **28**(4), 309–318 (2017)
14. Halvorsen, A.R.; Silwal-Pandit, L.; Meza-Zepeda, L.A.; Vodak, D.; Vu, P.; Sagerup, C.; Hovig, E.; Myklebost, O.; Børresen-Dale, A.-L.; Brustugun, O.T.; Helland, Å.: TP53 mutation spectrum in smokers and never smoking lung cancer patients. Front. Genet. **7**, 85 (2016). https://doi.org/10.3389/fgene.2016.00085
15. Forbes, S.A.; Beare, D.; Boutselakis, H.; Bamford, S.; Bindal, N.; Tate, J.; Cole, C.G.; Ward, S.; Dawson, E.; Ponting, L.; Stefancsik, R.; Harsha, B.; Kok, C.Y.; Jia, M.; Jubb, H.; Sondka, Z.; Thompson, S.; De, T.; Campbell, P.J.: COSMIC: somatic cancer genetics at high-resolution (2017). https://doi.org/10.1093/nar/gkw1121
16. NIH: TCGA: The Cancer Genome Atalas. https://cancergenome.nih.gov. Accesses 25 Sept. 2017
17. Augert, A.; Zhang, Q.; Bates, B.; Cui, M.; Wang, X.; Wildey, G.; Dowlati, A.; MacPherson, D.: Small cell lung cancer exhibits frequent inactivating mutations in the histone methyltransferase KMT2D/MLL2: CALGB 151111 (Alliance). J. Thorac. Oncol. **12**(4), 704–713 (2017)
18. Ramani, R.G.; Jacob, S.G.: Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models. PLoS ONE **8**(3), e58772 (2013). https://doi.org/10.1371/journal.pone.0058772
19. Hosseinzadeh, F.; KayvanJoo, A.H.; Ebrahimi, M.; Goliaei, B.: Prediction of lung tumor types based on protein attributes by machine learning algorithms. SpringerPlus **2**, 238 (2013). https://doi.org/10.1186/2193-1801-2-238
20. Li, J.; Ching, T.; Huang, S.; Garmire, L.X.: Using epigenomics data to predict gene expression in lung cancer. BMC Bioinform. **16**(5), 5–10 (2015)
21. Zhang, Y.; Elgizouli, M.; Schöttker, B.; Holleczek, B.; Nieters, A.; Brenner, H.: Smoking-associated DNA methylation markers predict lung cancer incidence. Clin. Epigenetics **8**, 127 (2016). https://doi.org/10.1186/s13148-016-0292-4
22. Salim, A.; Amjesh, R.; Vinod, C.S.S.: SVM based lung cancer prediction using microRNA expression profiling from NGS data. Paper Presented at the Asian Conference on Intelligent Information and Database Systems, vol. 38, pp. 599–609 (2016)
23. Velazquez, E.R.; Parmar, C.; Liu, Y.; Coroller, T.P.; Cruz, G.; Stringfield, O.; Ye, Z.; Makrigiorgos, M.; Fennessy, F.; Mak, R.H.; Gillies, R.; Quackenbush, J.; Aerts, H.J.W.L.: Somatic mutations drive distinct imaging phenotypes in lung cancer. Cancer Res. **77**(14), 3922–3930 (2017)
24. Ji-Yeon, Y.; Yoshihara, K.; Tanaka, K.; Hatae, M.; Masuzaki, H.; Itamochi, H.; Takano, M.; Ushijima, K.; Tanyi, J.L.; Coukos, G.; Lu, Y.; Mills, G.B.; Verhaak, R.G.W.: Predicting time to ovarian carcinoma recurrence using protein markers. J. Clin. Invest. **123**(9), 3740–3750 (2013)
25. Ali, S.; Majid, A.: Can-Evo-Ens: classifier stacking based evolutionary ensemble system for prediction of human breast cancer using amino acid sequences. J. Biomed. Inform. **54**, 256–269 (2015)
26. Munteanu, C.R.; Magalhães, A.L.; Uriarte, E.; González-Díaz, H.: Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J. Theor. Biol. **257**, 303–311 (2009)
27. Ali, S.; Majid, A.; Khan, A.: IDM-PhyChm-Ens: intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids. Amino Acids **46**(4), 977–993 (2014)
28. Robertson, W.W.; Steliga, M.A.; Siegel, E.R.; Arnaoutakis, K.: Accuracy of fine needle aspiration and core lung biopsies to predict histology in patients with non-small cell lung cancer. Med. Oncol. **31**(6), 967 (2014). https://doi.org/10.1007/s12032-014-0967-7
29. Online Mendelian Inheritance in Man (OMIM). Johns Hopkins University, Baltimore. https://www.omim.org/. Accessed October 10 (2017)
30. Smedley, D.; Haider, S.; Ballester, B.; Holland, R.; London, D.; Thorisson, G.; Kasprzyk, A.: BioMart: biological queries made easy. BMC Genom. **10**(1), 22 (2009). https://doi.org/10.1186/1471-2164-10-22
31. Zerbino, D.R.; Achuthan, P.; Akanni, W.; Amode, M.R.; Barrell, D.; Bhai, J.; Billis, K.; Cummins, C.; Gall, A.; Girón, C.G.; Gil, L.; Gordon, L.; Haggerty, L.; Haskell, E.; Hourlier, T.; Izuogu, O.G.; Janacek, S.H.; Juettemann, T.; To, J.K.; Laird, M.R.; Lavidas, I.; Liu, Z.; Loveland, J.E.; Maurel, T.; McLaren, W.; Moore, B.; Mudge, J.; Murphy, D.N.; Newman, V.; Nuhn, M.; Ogeh, D.; Ong, C.K.; Parker, A.; Patricio, M.; Riat, H.S.; Schuilenburg, H.; Sheppard, D.; Sparrow, H.; Taylor, K.; Thormann, A.; Vullo, A.; Walts, B.; Zadissa, A.; Frankish, A.; Hunt, S.E.; Kostadima, M.; Langridge, N.; Martin, F.J.; Muffato, M.; Perry, E.; Ruffier, M.; Staines, D.M.; Trevanion, S.J.; Aken, B.L.; Cunningham, F.; Yates, A.; Flicek, P.: Ensembl 2018. Nucleic Acids Res. **46**(D1), D754–D761 (2018). https://doi.org/10.1093/nar/gkx1098

32. Mirza, M.T.; Khan, A.; Tahir, M.; Lee, Y.S.: MitProt-Pred: predicting mitochondrial proteins of plasmodium falciparum parasite using diverse physiochemical properties and ensemble classification. Comput. Biol. Med. **43**(10), 1502–1511 (2013)
33. Chen, C.; Zhou, X.; Tian, Y.; Zou, X.; Cai, P.: Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal. Biochem. **357**, 116–121 (2006)
34. Limongelli, I.; Marini, S.; Bellazzi, R.: PaPI: pseudo amino acid composition to score human protein-coding variants. BMC Bioinform. **16**, 123 (2015). https://doi.org/10.1186/s12859-015-0554-8
35. Chou, K.C.; Zhang, C.T.: Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. **30**(4), 275–349 (1995)
36. Sugiyama, M.: Introduction to Statistical Machine Learning, pp. 237–244. Morgan Kaufmann, Boston (2016)
37. Theodoridis, S.: Machine Learning: A Bayesian and Optimization Prospective. Elsevier, Hoboken (2015)
38. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Berlin (1999)
39. Duda, R.O.; Hart, P.E.; Stork, D.G.: Pattern Classification, 2nd edn. Wiley, Hoboken (2000)
40. Python Software Foundation. https://www.python.org/. Accessed June 2017
41. Jiao, Y.; Du, P.: Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant. Biol. **4**(4), 320–330 (2016)
42. Tom, F.: ROC graphs: notes and practical considerations for researchers. Mach. Learn. **31**, 1–38 (2004)
43. Kuijjer, M.L.; Paulson, J.N.; Salzman, P.; Ding, W.; Quackenbush, J.: Cancer subtype identification using somatic mutation data. Br. J. Cancer **118**, 1492–1501 (2018)
44. Weng, T.-Y.; Wang, C.-Y.; Hung, Y.-H.; Chen, W.-C.; Chen, Y.-L.; Lai, M.-D.: Differential expression pattern of THBS1 and THBS2 in lung cancer: clinical outcome and a systematic-analysis of microarray databases. PLoS ONE **11**(8), e0161007 (2016). https://doi.org/10.1371/journal.pone.0161007
45. Liu, J.X.; Gao, Y.L.; Xu, Y.; Zheng, C.H.; You, J.: Differential expression analysis on RNA-seq count data based on penalized matrix decomposition. IEEE Trans. Nanobiosci. **13**(1), 12–18 (2014)
46. Liu, J.-X.; Wang, Y.-T.; Zheng, C.-H.; Sha, W.; Mi, J.-X.; Xu, Y.: Robust PCA based method for discovering differentially expressed genes. BMC Bioinform. **14**(8), S3 (2013). https://doi.org/10.1186/1471-2105-14-s8-s3