Computer Methods and Programs in Biomedicine 245 (2024) 108019

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

A novel method for detection of pancreatic Ductal Adenocarcinoma using explainable machine learning

Murtaza Aslam [a], Fozia Rajbdad [a], Shoaib Azmat [b], Zheng Li [a], J. Philip Boudreaux [c], Ramcharan Thiagarajan [c], Shaomian Yao [d], Jian Xu [a,*]

[a] Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803, USA
[b] Department of Electrical and Computer Engineering, COMSATS University Islamabad, Pakistan
[c] Department of Surgery, School of Medicine, Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA
[d] Department of Comparative Biomedical Sciences, School of Veterinary Medicine, Louisiana State University, Baton Rouge, LA 70803, USA

ARTICLE INFO

Keywords:
Pancreatic ductal adenocarcinoma
Raman spectroscopy
Mutation
Explainable features
Support vector machine-recursive feature elimination

ABSTRACT

Background and Objective: Pancreatic Ductal Adenocarcinoma (PDAC) is a form of pancreatic cancer that is one of the primary causes of cancer-related deaths globally, with less than 10 % of the five years survival rate. The prognosis of pancreatic cancer has remained poor in the last four decades, mainly due to the lack of early diagnostic mechanisms. This study proposes a novel method for detecting PDAC using explainable and supervised machine learning from Raman spectroscopic signals.

Methods: An insightful feature set consisting of statistical, peak, and extended empirical mode decomposition features is selected using the support vector machine recursive feature elimination method integrated with a correlation bias reduction. Explicable features successfully identified mutations in Kirsten rat sarcoma viral oncogene homolog (KRAS) and tumor suppressor protein53 (TP53) in the fingerprint region for the first time in the literature. PDAC and normal pancreas are classified using K-nearest neighbor, linear discriminant analysis, and support vector machine classifiers.

Results: This study achieved a classification accuracy of 98.5% using a nonlinear support vector machine. Our proposed method reduced test time by 28.5 % and saved 85.6 % memory utilization, which reduces complexity significantly and is more accurate than the state-of-the-art method. The generalization of the proposed method is assessed by fifteen-fold cross-validation, and its performance is evaluated using accuracy, specificity, sensitivity, and receiver operating characteristic curves.

Conclusions: In this study, we proposed a method to detect and define the fingerprint region for PDAC using explainable machine learning. This simple, accurate, and efficient method for PDAC detection in mice could be generalized to examine human pancreatic cancer and provide a basis for precise chemotherapy for early cancer treatment.

1. Introduction

Pancreatic Ductal Adenocarcinoma (PDAC), the fourteenth most common malignancy, is one of the prominent causes of death worldwide. According to the GLOBOCAN survey, pancreatic cancer is the seventh most typical cause of cancer-related deaths worldwide and the fourth in the United States [1,2]. The one-year survival rate of patients with PDAC is 24 %, and the five-year survival rate is < 10 % [3,4]. Only 10–20 % of PDAC can be removed with surgery; however, partial removal can lead to local recurrence of pancreatic cancer [5]. The causes of PDAC are due to different types of mutations at the gene level; KRAS is mutated in 80.56 % of PDAC cases. It belongs to the Ras gene family of the oncogene class, which provides instructions for making the K-Ras protein part of the RAS/MAPK pathways [6]. It is activated upon receiving signals from the outside of the cell. It can switch on downstream pathways to initiate cell growth, division, and self-destruction (apoptosis). Different proteins are involved in activating and signaling pathways that keep Ras inactive to prevent persistent activation. However, these proteins become inactive due to mutations, which keep Ras activated, leading to malignant transformation. KRAS mutants are

* Corresponding author.
E-mail address: jianxu1@lsu.edu (J. Xu).

https://doi.org/10.1016/j.cmpb.2024.108019
Received 23 April 2023; Received in revised form 9 January 2024; Accepted 10 January 2024
Available online 13 January 2024

divided into three categories based on codon and G12 mutation at codon 12, with subtype G12V being one of them [7], representing 25.0 % of nearly all PDAC [8].

Similarly, TP53 is central in adjusting the safe microenvironment but is frequently mutated in PDAC [9]. This tumor suppressor gene is recruited when the DNA is damaged [10]. Any damage to the p53 protein makes it unable to bind the DNA; hence, no decision can be made to repair the DNA or destroy the cell [11]. Dysfunctional TP53 causes the accumulation of damaged DNA in cells [12]. These damaged cells are more likely to grow out of control, forming tumors [13]. TP53 alterations are present in nearly half of all human cancers, making them the most common source of cancer [9]. Moreover, the Suppressor of Mothers against Decapentaplegic (SMAD4) is another tumor suppressor protein that prevents cells from developing and dividing rapidly [14]. It is deactivated in 50–60 % of PDAC cases due to homozygous omission or mutation [15]. Alterations in the SMAD4 gene appear in the late phases when the carcinoma is histologically identifiable [16].

Early detection of PDAC is the best way to cure this disease; hence, a suitable device for early detection is essential. Histopathology is an invasive cancer evaluation method and is currently used for the diagnosis of PDAC. However, it is challenging to evaluate PDAC with intraoperative histopathology because of impurities in the blood and digestive tract cells [17,18]. Similarly, different non-invasive imaging technologies used for detection include computed tomography (CT) [19], magnetic resonance imaging (MRI) [20], positron emission tomography (PET) [21], and endoscopic ultrasound (EUS) [22]. However, these imaging techniques are expensive, bulky, time-consuming, and have portability and availability challenges, and some of these methods have radiation exposure issues if repetition is required. In addition, the detection accuracy of PDAC is usually compromised because it is a retroperitoneal organ, and the operator's expertise is also needed in some cases [23].

Among optical imaging techniques, Raman spectroscopy (RS) is an evolving diagnostic tool for analyzing chemical components in many fields. RS has distinctive advantages, including a non-ionization nature and high specificity. RS is label-free; hence, no sample preparation is required [24]. In many machine learning (ML)-based studies, RS is used to detect different types of cancers, such as brain, breast, cervical, and skin cancers, with accuracies of 96 %, 90 %, 85.7 %, and 91–92 %, respectively [25–29]. Deep learning is one of the most popular ML techniques because of its high prediction performance [30]. For PDAC, Li et al. used convolutional neural networks (CNN) and RS signals and obtained a significantly high detection accuracy of greater than 97 % [31]. The medical domain requires reason and insight to understand beyond standard quantitative performance evaluation [32]. Moreover, explaining the ability of the analysis is important for clinical diagnosis. Therefore, methods that only detect cancer, such as deep learning, usually cannot assist in treating cancer. In addition, patients must always understand chemotherapy for their cancer treatment to reduce inaccurate beliefs about chemotherapy, which can increase their life expectancy [33]. In addition, deep learning requires large training datasets, has high computational costs, and lacks generality in new data results [34].

Relevant and explainable feature extraction is, therefore, a critical factor for improving the classification accuracy and understanding of data patterns. Statistical primary features, such as the mean, mode, median, standard deviation, and variance, are the simplest features used for data analysis [35]. With increased variability and randomness of the data, derived statistical features such as skewness, kurtosis, shape factor, impulse factor, crest factor, and clearance factor can help to extract more explainable features [36,37]. Similarly, empirical mode decomposition (EMD) is suitable for nonlinear and nonstationary signals such as EEG, optical, and chemical [38–40]. In addition, feature selection methods to remove redundant features can significantly improve detection accuracy. In some studies, feature selection methods are also embedded with classification algorithms to achieve better classification accuracy, such as support vector machine recursive feature elimination (SVM-RFE), proposed by [41]. This method is less susceptible to overfitting and is highly efficient when the feature set is large [42].

The extracted features are fed into the classification algorithm to separate the classes in the data. Classification algorithms, such as linear discriminant analysis (LDA), linear support vector machine (LSVM), k-nearest neighbor (kNN), and nonlinear support vector machine (NLSVM), have been integrated with RS for the detection of various types of cancers such as esophageal, breast, prostate, colon, and liver [27,43–46]. Similarly, ML classifiers have also been integrated with RS for diagnosing diseases such as the classification of kidney stones and renin hypertension [47,48].

This study proposes a novel method for PDAC detection using statistical, peak, and EMD extended explicable features obtained from the RS signals. Redundancy in features is removed by the SVM-RFE feature selection method integrated with the correlation bias reduction (CBR) method. To the best of our knowledge, this is the first study to define a fingerprint region (600–1800 cm^{-1}) using obtained features. Distinct novel regions for normal and mutated KRAS and TP53, for a better understanding of PDAC, are also defined in this work. Subsequently, the classification of PDAC from the normal pancreas is performed using LDA, kNN, LSVM, and NLSVM. The NLSVM classification algorithm reduced 85.6 % memory utilization, 28.5 % testing time, 80.5 % training time and has 1.1 % higher classification accuracy than the state-of-the-art algorithm [31]. The accuracy remained above 95 % when other simpler machine learning algorithms, such as LDA, LSVM, and kNN, were used, compared to 98.5 % for NLSVM. Performance evaluation parameters such as accuracy, specificity, sensitivity, and receiver operating characteristic (ROC) curves are used to evaluate the PDAC detection method [49]. The contributions of the proposed method for PDAC detection are summarized as follows.

This study proposes a novel method for detecting PDAC, a form of pancreatic cancer with less than a 10 % of the five years survival rate.

To the best of our knowledge, this is the first study to define a fingerprint region that can be crucial in determining a standard pattern for early PDAC detection in humans using explainable features that successfully identified mutations in Kirsten rat sarcoma viral oncogene homolog (KRAS) and tumor suppressor protein53 (TP53).

PDAC and normal pancreas are classified using multiple machine-learning classifiers and achieved a classification accuracy of 98.5 % using NLSVM, and reduced test time by 28.5 % and saved 85.6 % memory utilization compared to the state-of-the-art method.

2. Materials and methods

The proposed method for detecting PDAC using RS signals involves data acquisition, feature extraction, selection, and classification. The data used in the proposed study are described in the Materials section. The Methods section discusses the feature extraction, selection, and classification techniques.

2.1. Materials

In this subsection, details about data and data acquisition are discussed. Materials include a description of the cell line, animal model, data acquisition, and system used to process the data.

i) *Cell Line:*

The human CFPAC-1 cell line (ATCCR CRL 1918TM, pancreatic ductal adenocarcinoma) was used in this study. Tumor cells were cultured in Iscove's Modified Dulbecco's Medium (ATCCR 30-2005TM) with 10 % fetal bovine serum (Neuromics, Edina, Minnesota) at 37 °C and 5 % CO_2 in a humidified environment.

i) *Animal Model:*

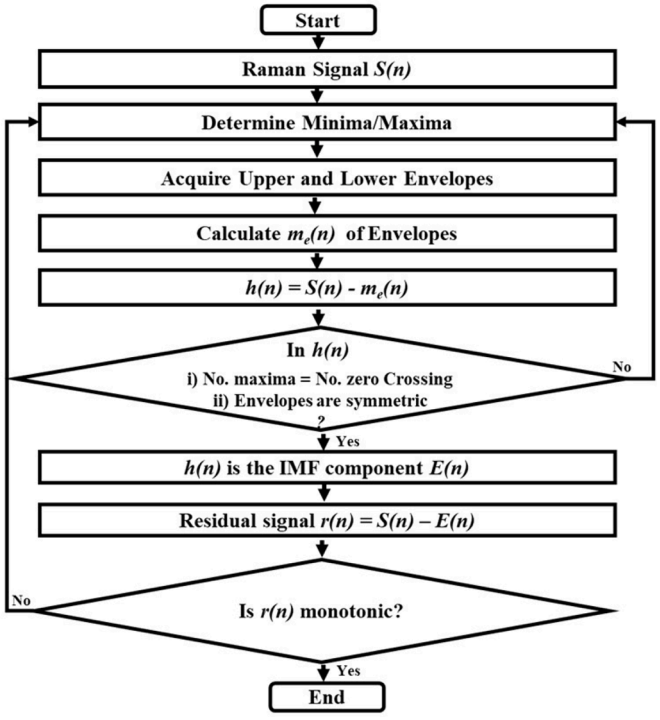


Fig. 1. Flowchart of EMD.

Data of 6–8-week-old female immunocompetent athymic nude Nu/J mice (Jackson Laboratories, Bar Harbor, Maine, USA) animal model is used for this study. After the CFPAC-1 cells were developed in the media, nearly all the cells were injected into the back of the mice by subcutaneous infusion. When the tumor size was around 1 cm, the mice were euthanized, and the whole tumor and normal pancreas were extracted. This study was approved by the Institutional Animal Care and Use Committee of Louisiana State University (IACUC#20–046), and all procedures followed the regulations on animal investigation.

i) *Data Acquisition*:

RS data acquisition system consists of a laser diode (Turnkey Raman Lasers-785 Series, Ocean Optics Inc., Dunedin, Florida, United States), QE Pro-spectrometer (Ocean Optics, Inc), a Raman probe (RPS785, InPhotonics Inc., Norwood, Massachusetts, United States), and Ocean Wave hardware to computer interface software. The RS data used in this paper was obtained from 20 mice. 2529 RS signals were collected from a mouse: 1305 signals from the tumor and 1224 signals from the normal pancreas.

i) *System Specifications:*

All the methods for signal analyses were implemented in Python and MATLAB programming languages using an Intel(R) Core (TM) i7-1165G7 computer system with a 2.80 GHz processor and 16.0 GB memory.

2.2. Methods

This subsection discusses in detail the methods used in this paper. These methods include statistical, EMD, and peak feature extraction, feature selection using SVM-RFE integrated with CBR, and classification using LDA, kNN, LSVM, and NLSVM.

i) *Feature Extraction*

Several distinct features from the PDAC data have been extracted. The features that can contribute more to drawing decision boundaries between cancer and normal RS signals are used. These features include basic statistical features such as mean, median, mode, root mean square, standard deviation, variance, and derived features such as shape factor, skewness, coefficient of variation, and kurtosis. Similarly, impulse metrics, such as impulse factor, crest factor, clearance factor, and Shannon entropy, are also obtained as potential features for classifying cancer tissue. Moreover, signal processing metrics, such as energy, power, and nonlinear energy, are also obtained. In addition, EMD features that are extracted from intrinsic functions are employed. Finally, the peak features are also extracted. In each of the 2529 RS signals (trials), N is the total number of discrete time samples in a single trial S, $\overline{S}$ is the mean value of the trial, S_{pmax} is the highest peak, and S_i is the i^{th} sample. All features are mathematically described as follows.

a) Empirical Mode Decomposition: It is based on the representation of signals into a set of functions called intrinsic mode functions (IMFs) through the sifting process [38] described in Fig. 1.

Subsequently, the original signal can be obtained by adding all IMFs (E(n)) acquired during the sifting process, and the residual signal (r(n)) given in Eq. (1).

$$S(n) = \sum_{i=1}^{I} E_i(n) + r(n) \tag{1}$$

In this work, the average signals of the PDAC and normal pancreas are obtained by taking the means of all trials of training signals of each class as described in Eq (2).

$$A_{np}(n) = \frac{1}{M_{np}} \sum_{i=1}^{M_{np}} S_{np,i}(n),\ A_{pd}(n) = \frac{1}{M_{pd}} \sum_{i=1}^{M_{pd}} S_{pd,i}(n) \tag{2}$$

$A_{np}(n)$ is the average RS signal for the normal pancreas, and $A_{pd}(n)$ is the average RS signal of PDAC. M_{np} and M_{pd} are the total number of trials for normal pancreas and PDAC respectively. $S_{np}(n)$ is the normal pancreas trial and $S_{pd}(n)$ is the PDAC trial. EMD is applied to $A_{np}(n)$ and $A_{pd}(n)$ to obtained IMFs given in Eq. (3).

$$A_{np}(n) = \sum_{i=1}^{I_{np}} E_{np,i}(n),\ A_{pd}(n) = \sum_{i=1}^{I_{pd}} E_{pd,i}(n) \tag{3}$$

Where $E_{np,i}(n)$and $E_{pd,i}(n)$ are the IMFs of $A_{np}(n)$ and $A_{pd}(n)$ respectively, I_{np} and I_{pd} are the respective IMF counts, and the residual signal $r(n)$ is discarded. Hence, any trial signal $S(n)$ can be approximated using IMFs of the average signals.

$$S_{np}(n) \cong \sum_{i=1}^{I_{np}} a_{np,i}\ E_{np,i}(n) = \widehat{S}_{np}(n) \tag{4}$$

$$S_{pd}(n) \cong \sum_{i=1}^{I_{pd}} a_{pd,i}\ E_{pd,i}(n) = \widehat{S}_{pd}(n) \tag{5}$$

Eqs. (4) and 5, $a_{np,i}$ and $a_{pd,i}$ are extension coefficients obtained from the IMFs of the average signals A_{np} and A_{pd} The extended coefficients in this paper are calculated the same as[38], by using the pseudoinverse problem, constrained with the least squared error as given in Eqs. (6), (7), and (8).

$$B_{np}a_{np} = \widehat{S}_{np},\ B_{pd}a_{pd} = \widehat{S}_{pd} \tag{6}$$

$$B_{np} = \begin{bmatrix} E_{np,1}(0) & E_{np,2}(0) & \cdots & E_{np,I_{np}}(0) \\ E_{np,1}(1) & E_{np,2}(1) & \cdots & E_{np,I_{np}}(1) \\ \vdots & \vdots & & \vdots \\ E_{np,1}(N-1) & E_{np,2}(N-1) & & E_{np,I_{np}}(N-1) \end{bmatrix} \tag{7}$$

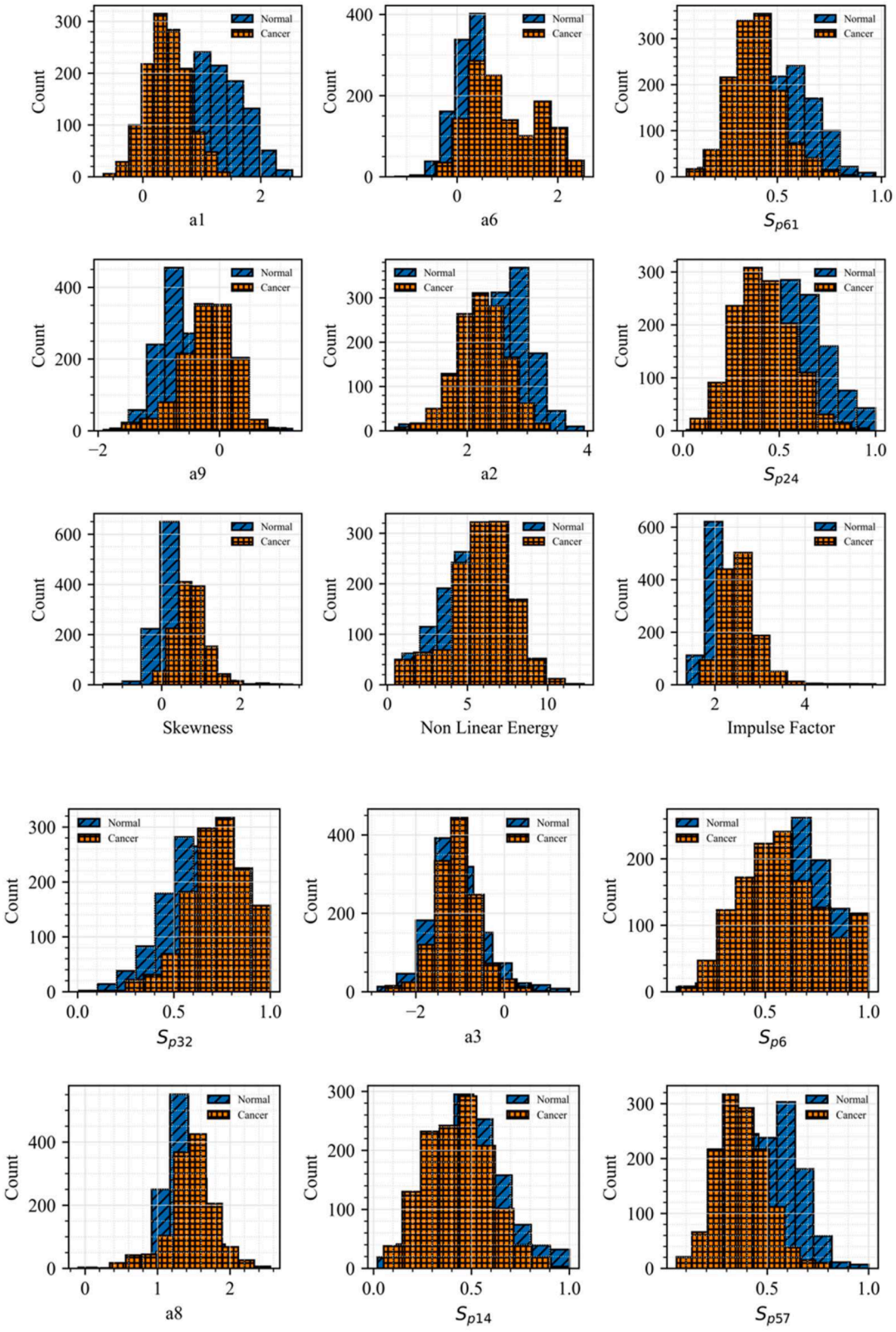


Fig. 2. Distributions of highest-ranked statistical, peak, and EMD extended features.

$$B_{pd} = \begin{bmatrix} E_{pd,1}(0) & E_{pd,2}(0) & \dots & E_{pd,I_{pd}}(0) \\ E_{pd,1}(1) & E_{pd,2}(1) & \dots & E_{pd,I_{pd}}(1) \\ \vdots & \vdots & & \vdots \\ E_{pd,1}(N-1) & E_{pd,2}(N-1) & & E_{pd,I_{pd}}(N-1) \end{bmatrix} \quad (8)$$

a_{np} and a_{pd} are calculated based on the least squared error constraint, which results in B_{np} and B_{pd} as non-square matrices. These obtained coefficients a_{np} and a_{pd}, of the trial signal S of both classes given in Eqs. (9) and (10) below, are further used as potential features for classification.

Table 1
Classification Performance of ML Classifiers with Statistical, EMD Extended Features, and Peak Features.

Feature Type	Selected Features	Classifier	Average Accuracy (%) ± SD	Average Sensitivity (%) ± SD	Average Specificity (%) ± SD
Statistical Features	All	LDA	81.89 ± 2.09	82.60 ± 2.08	81.12 ± 4.04
		kNN	81.33 ± 3.09	82.76 ± 3.53	79.80 ± 5.76
		LSVM	83.35 ± 2.29	81.36 ± 3.95	85.21 ± 2.81
		NLSVM	84.10 ± 2.79	85.20 ± 4.09	83.06 ± 3.70
	Skewness, Non-Linear Energy, Clearance Factor, Impulse Factor, Crest Factor, Kurtosis	LDA	82.60 ± 2.42	86.51 ± 3.80	78.44 ± 5.38
		kNN	87.19 ± 2.84	84.80 ± 4.79	89.42 ± 3.34
		LSVM	83.19 ± 2.19	84.67 ± 3.44	81.62 ± 4.93
		NLSVM	**87.34 ± 2.40**	88.89 ± 3.44	85.78 ± 3.71
EMD Extended Features	All	LDA	91.70 ± 1.80	95.17 ± 2.00	87.98 ± 4.02
		kNN	93.12 ± 1.58	96.62 ± 1.26	89.37 ± 3.26
		LSVM	92.10 ± 1.75	94.71 ± 1.61	89.21 ± 3.80
		NLSVM	**95.13 ± 1.76**	93.30 ± 2.93	96.85 ± 1.97
Statistical and EMD Extended Features	All	LDA	93.33 ± 1.74	97.37 ± 1.99	89.30 ± 2.08
		kNN	85.79 ± 2.33	82.93 ± 3.19	88.66 ± 3.25
		LSVM	94.33 ± 1.67	97.08 ± 2.70	91.58 ± 2.07
		NLSVM	78.85 ± 2.58	98.00 ± 1.05	56.30 ± 5.17
	a1, a6, Skewness, a9, a2, a3, a8, a4, Impulse Factor, Clearance Factor, Non-Linear Energy	LDA	92.00 ± 2.38	95.94 ± 1.56	87.80 ± 4.30
		kNN	93.63 ± 1.66	96.78 ± 2.27	90.27 ± 2.13
		LSVM	92.72 ± 2.48	94.55 ± 2.32	90.76 ± 3.64
		NLSVM	**95.21 ± 1.48**	97.31 ± 1.83	92.97 ± 1.89
Peak Features	All	LDA	94.98 ± 1.53	96.16 ± 2.20	93.71 ± 2.34
		kNN	96.40 ± 1.39	97.16 ± 1.43	95.60 ± 2.21
		LSVM	94.94 ± 1.69	93.80 ± 1.93	96.01 ± 2.41
		NLSVM	97.03 ± 1.08	97.62 ± 1.34	96.40 ± 1.82
	70	LDA	94.66 ± 1.60	96.55 ± 2.22	92.64 ± 2.39
		kNN	95.45 ± 1.71	96.47 ± 1.97	94.35 ± 3.37
		LSVM	94.22 ± 1.85	95.63 ± 2.00	92.72 ± 2.63
		NLSVM	**97.30 ± 1.47**	98.40 ± 1.36	96.15 ± 2.95
Statistical, EMD Extended Features and Peak Features	All	LDA	97.27 ± 1.28	95.67 ± 2.27	98.77 ± 1.34
		kNN	86.63 ± 2.92	84.88 ± 4.22	88.38 ± 3.20
		LSVM	97.23 ± 1.53	98.00 ± 1.19	96.40 ± 2.68
		NLSVM	68.17 ± 2.60	98.77 ± 0.64	34.48 ± 5.30
	50	LDA	96.60 ± 1.35	98.90 ± 1.24	93.96 ± 2.66
		kNN	96.40 ± 1.48	98.78 ± 1.14	93.62 ± 2.56
		LSVM	96.84 ± 1.13	97.85 ± 1.37	95.75 ± 2.35
		NLSVM	**98.50 ± 1.11**	99.00 ± 0.97	97.99 ± 1.95

$$a_{np} = \begin{bmatrix} a_{np,1} & a_{np,2} & \dots & a_{np,I_{np}} \end{bmatrix}^T = \left(B_{np}^T B_{np}\right)^{-1} B_{np}^T S_{np} \quad (9)$$

$$a_{pd} = \begin{bmatrix} a_{pd,1} & a_{pd,2} & \dots & a_{pd,I_{pd}} \end{bmatrix}^T = \left(B_{pd}^T B_{pd}\right)^{-1} B_{pd}^T S_{pd} \quad (10)$$

a) Peak Features (S_p): Each data trial from wave number 600 cm^{-1} to 3975 cm^{-1} has been divided into 135 windows, consisting of 25 wave numbers. Peak values from every window have been extracted, and window's maximum peak value is obtained as a feature as described in Eq. (11). If no prominent peaks are obtained from any window, then the average value from that window is extracted as a feature.

$$S_p = \left[S_{p1}, S_{p2}, S_{p3} \dots S_{pn}\right], \; n = 1, 2, 3 \dots 135 \quad (11)$$

b) Shape factor (S_{SF}): It is independent of the dimensions of the signal, and it depends only on the signal's shape, given in Eq. (12).

$$S_{SF} = \frac{S_{rms}}{\bar{S}} \quad (12)$$

Where S_{rms} is the root mean square of each trial.

a) Skewness (S_{skew}): Asymmetry of the signal distribution is described with the help of skewness, given in Eq. (13).

$$S_{skew} = \frac{\frac{1}{N}\sum_{i=1}^{N}(S_i - \bar{S})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(S_i - \bar{S})^2\right]^{3/2}} \quad (13)$$

b) Coefficient of variation (S_{CV}): It is the ratio of the standard deviation to the sample mean and describes the data distribution relative to the trial mean given in Eq. (14).

$$S_{CV} = \frac{S_{SD}}{\bar{S}} \quad (14)$$

Where S_{SD} is the standard deviation of a trial.

a) Kurtosis (S_{kurt}): Outliers can be analyzed with the help of kurtosis. If data is prone to outliers, the value of kurtosis will increase, and Eq. (15) below describes kurtosis.

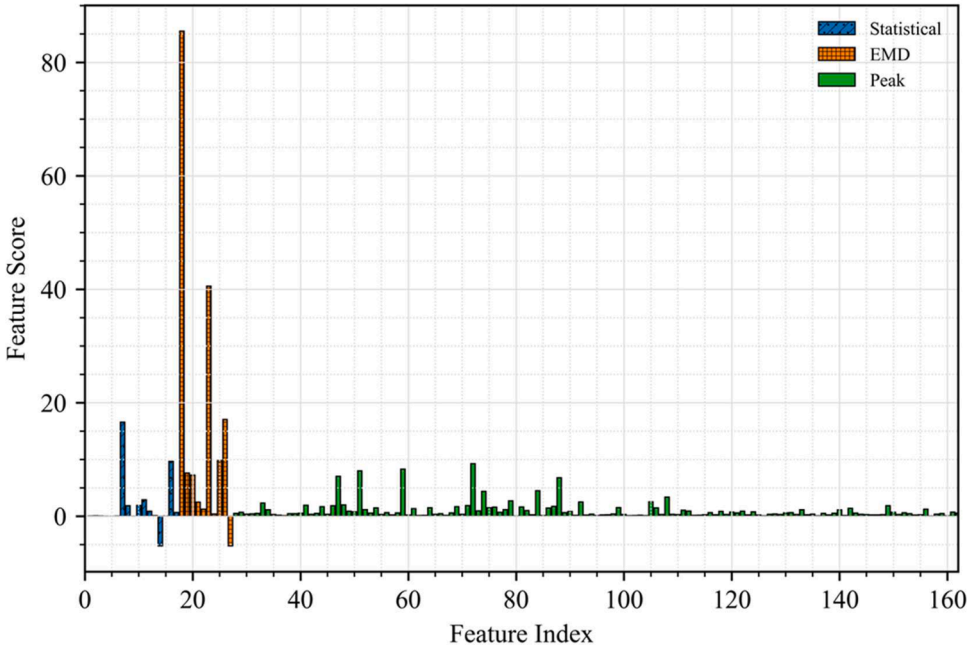


Fig. 3. Feature scores assigned to statistical, EMD extended features, and peak features by the SVM-RFE incorporated with CBR feature selection method.

$$S_{kurt} = \frac{\frac{1}{N}\sum_{i=1}^{N}(S_i - \overline{S})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(S_i - \overline{S})^2\right]^2} \tag{15}$$

b) Impulse factor (S_{IF}): Comparison of the peak height to the signal's mean level can be analyzed using the impulse factor in Eq. (16).

$$S_{IF} = \frac{S_{pmax}}{\overline{S}} \tag{16}$$

c) *Crest factor (S_{crest}):* Noise usually appears in the peaks of the signals before it appears in the energy of the signals. Hence, signal noise can be detected using the crest factor in Eq. (17).

$$S_{crest} = \frac{S_{pmax}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}|S_i^2|}} \tag{17}$$

d) *Clearance Factor (S_{clear}):* This feature can give higher separability between classes where data has higher overlap. The clearance factor is defined mathematically with Eq. (18).

$$S_{clear} = \frac{S_{pmax}}{\left(\frac{1}{N}\sum_{i=1}^{N}|\sqrt{|S_i|}|^2\right)} \tag{18}$$

e) *Shannon Entropy(ShEn):* It reflects the randomness of our data which helps in observing the relevant information between the trials and is defined with Eq. (19). Here p_i is the probability of each sample.

$$ShEn(S) = \sum_{i=1}^{N} p_i \, logp_i \tag{19}$$

f) *Nonlinear Energy (NE):* It is a derived energy concept in the form of nonlinear energy for the non-stationary signal used in this paper, specified by Eq. (20).

$$NE(S) = \sum_{i=1}^{N-2}\left(S^2[i] - S[i+1]S[i-1]\right) \tag{20}$$

i) *Feature selection*

In this study, the feature selection is performed using SVM-RFE to obtain an optimal feature set. In SVM-RFE, the search for the best feature subset and model construction is combined by integrating feature selection as part of the classifier algorithm [50]. NLSVM training data are mapped to a higher dimensional space R^h, when it is non-linearly separable in the input space R^l given in Eq. (21).

$$f \epsilon R^l \rightarrow \Phi(f) R^h \tag{21}$$

This approach is based on solving the quadratic optimization problem in the feature space. The dual Lagrangian that needs to be minimized is defined in Eq. (22).

$$L_d(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \Phi(f_i)\Phi(f_j) \tag{22}$$

Where, α_i and α_j are the Lagrangian multipliers, y_i, and y_j are the class labels, $K(f_i,f_j) = \Phi(f_i)\Phi(f_j)$ is the kernel function, f_i, f_j are the features set of both classes, and n is the number of trials. Kernels must be calculated in the input space to obtain the required scalar product in the feature space; hence, the mapping can be avoided. Gaussian kernel functions are used excessively in such cases, as defined in Eq. (23) for our proposed work.

$$K\left(f_i,f_j\right) = e^{-\gamma f_i - f_j^2} \tag{23}$$

In ranking of features, if a feature's elimination produces slight variations in the objective function given in Eq. (24), those features can be removed. This directs us to the subsequent ranking condition for feature k.

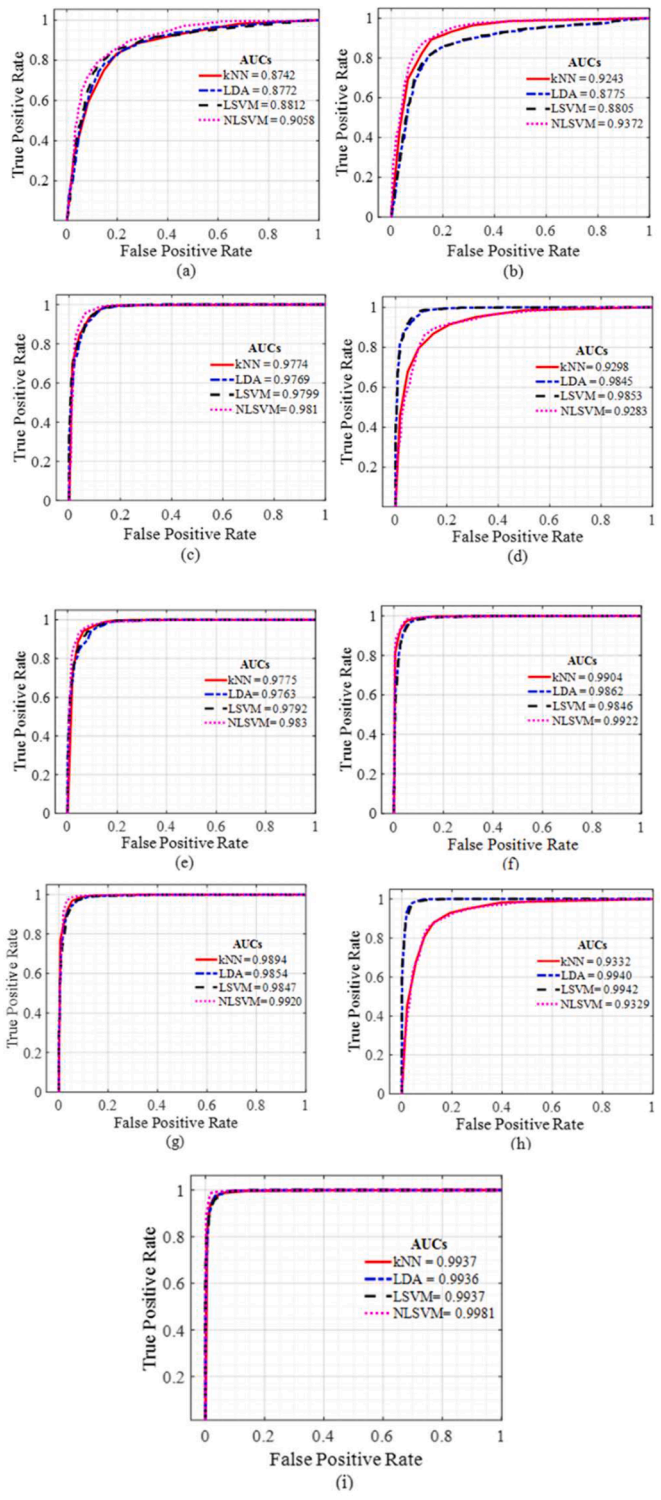


Fig. 4. AUC values of kNN, LDA, LSVM, and NLSVM classifiers represented using different features combinations (a) AUC using statistical features (b) AUC using selected statistical features: skewness, nonlinear energy, clearance factor, impulse factor, crest factor, and kurtosis (c) AUC using EMD extended features (d) AUC using all statistical and EMD features (e) AUC using selected features of EMD: a1-a4, a6, a8, a9, and selected statistical features: skewness, impulse factor, clearance factor, and nonlinear energy (f) AUC using all peak features (g) AUC using selected 70 peak features (h) AUC using all 162 features (i) AUC using selected 50 features.

Table 2
Comparison of Proposed ML Classification Methods and State of the Art.

Classifier	Average Accuracy (%) ± SD	Average Sensitivity (%) ± SD	Average Specificity (%) ± SD	Train/ Test Time (*sec*)	Memory (MB)
CNN [31]	97.39 ± 1.44	97.80 ± 0.90	96.85 ± 2.52	24.35/ 6.63	2338
LDA	97.27 ± 1.28	95.67 ± 2.27	98.77 ± 1.34	4.77/ 4.77	335
LSVM	97.23 ± 1.53	98.00 ± 1.19	96.40 ± 2.68	4.91 /4.91	336.2
kNN	96 0.40 ± 1.48	98.78 ± 1.14	93.62 ± 2.56	5.60/ 5.60	335.8
NLSVM	98.50 ± 1.10	99.00 ± 0.97	97.99 ± 1.95	4.74/ 4.74	337.5

$$j(k) = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(f_i, f_j) - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K\left(f_i^{-k}, f_j^{-k}\right) \quad (24)$$

Whereas *(-k)* means that the k^{th} feature has been removed without changing α'*s*, all those features that have small *j*'*s* are eliminated in each iteration. Furthermore, highly correlated features obtained from SVM-RFE can bring wrong estimation and inaccurate prediction, which is solved using the CBR method proposed by [50].

i) *Classification*

The proposed work classified RS signals into the normal pancreas and PDAC classes using ML algorithms. The paper employed linear and nonlinear ML classifier algorithms such as LDA, LSVM, kNN, and NLSVM. Fisher's criteria in LDA determine the ratio between the normal/tumor class variance and within-class variance [27]. kNN is a nonlinear but superficial classifier; it does not require training, and Euclidean distance is used as a distance metric to obtain the decision boundaries, partitioning the features into two regions. In the case of SVM, discussed in detail in the feature selection subsection above, the kernel function obtains the decision boundary.

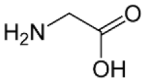
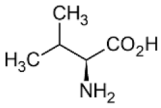
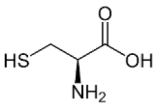
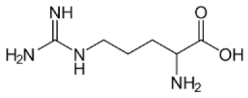
3. Results and discussion

This paper proposes a new method for detecting PDAC from RS signals using supervised ML. In total, 162 features have been extracted from the data. Later, redundant and irrelevant features are removed using SVM-RFE integrated with the CBR feature selection method. In addition, k-fold cross-validation is used in this paper; 2529 (normal=1224, PDAC = 1305) trials are split into training and cross-validation. 10 % of the data is used for the testing, and the remaining data is used for the 15-fold cross-validation. Each time a fold is used for validation, the remaining folds are used for training [45].

3.1. Results

The highest-ranked 15 features out of 162 for detecting PDAC from the normal pancreas are shown in Fig. 2. EMD features a_1 and a_2 show substantial separability between the two classes. Similarly, other features, including skewness, impulse factor, S_{p57}, S_{p61}, and a_9, have also shown considerable separability between the two classes. The distribution of some features, such as nonlinear energy, a_3, S_{p6}, and S_{p14}, demonstrated less separability in Fig. 2. However, when these features and other features are mapped to higher dimensions, a significant increase in classification accuracy is observed, as shown in Table 1. Therefore, these features can accurately classify PDAC and normal tissues. The feature selection is performed using the SVM-RFE and CBR algorithms. The features are ranked based on the feature scores achieved using one-by-one elimination of features in each iteration. Furthermore, the highest score is assigned to the EMD coefficient feature a_1, as shown

Table 3
Characteristics of Mutation Types in PDAC.

Characteristics	KRAS [62]	TP53[63]	SMAD4[63]
Gene Sequence	c.35G>*T*	c.724T>C	c.1_1659del1659
Mutated Protein Sequence	G12V	C242R	p.0?
Wild Amino Acid	Glycine	Cysteine [62]	Probably no protein is produced
Chemical Structure			
Wild Amino Acid Side Chain	H-(Aliphatic)	$HSCH_2$ – (Sulphur)	
Significant Peaks	508, 898, 1038, 1336, 1414, 1448[59]	644–686, 905–920, 1016–1055,1376–1389[60]	
Mutated Amino Acid	Valine	Arginine	
Mutated Amino Acid Side Chain	(CH3)2CH–(Aliphatic)	HN═C(NH2) NH(CH2)3– (Cationic)	
Chemical Structure			
Significant Peaks	757,829,847,948,967,1064,1270, 1328, 1336,1362,1411,1450,1473 [61]	857,894,930,980,1086,1176, 1317,1365,1408,1446[62]	

in Fig. 3. It can be observed from the feature scores that significant features appear in all three domains, i.e., statistical, EMD coefficients, and peak features.

The average accuracy, sensitivity, and specificity of test data for LDA, kNN, LSVM, and NLSVM classifiers using different combinations of statistical, EMD extended features, and peak features are reported in Table 1. For the statistical features, the six highest-ranked features achieved an average accuracy of 87.34 ± 2.40 % using the NLSVM classifier. The accuracy is improved to 95.13 ± 1.76 % when all EMD extended features are used as a feature set with the NLSVM classifier. Furthermore, when all 27 statistical and EMD extended features are combined, LSVM produced the highest average accuracy of 94.33 ± 1.67 %, whereas the average accuracy of NLSVM dropped to 78.85 ± 2.58 %. However, feature selection improved the average accuracy of NLSVM, which is 95.21 ± 1.48 % using 11 selected features. For peak features, an average accuracy of 97.30 ± 1.47 % is achieved by the NLSVM classifier for 70 selected peak features. Finally, when all the peak features are combined with all the statistical and EMD coefficients features, LDA attained an average accuracy of 97.27 ± 1.28 %, whereas NLSVM's average accuracy dropped to 68.17 ± 2.60 %. However, after applying feature selection, NLSVM achieved the overall highest average accuracy of 98.50 ± 1.11 % using 50 selected features. The maximum average accuracies for kNN and LSVM are 96.40 ± 1.48 % and 97.23 ± 1.53 % using 50 selected features and all the features, respectively. In addition, performance evaluation parameters, such as sensitivity and specificity, are assessed for classifiers. Like average accuracy, NLSVM produced the highest average sensitivity and specificity compared to LDA, kNN, and LSVM for most feature combinations, as shown in Table 1. NLSVM achieved the highest average sensitivity and specificity of 99.00 ± 0.97 % and 97.99 ± 1.95 %, respectively, using 50 selected features from 162 features. It can also be observed from Table 1 that the peak features achieved the highest test accuracy among the accuracies of all the individual features.

ROC curve is a method to measure the performance of the classification problem at various threshold values. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) for each possible cutoff. The area under the curve (AUC) is calculated from ROC and represents class separability [51]. Fig. 4 shows the average ROC curves and AUC values of LDA, kNN, LSVM, and NLSVM classifiers for feature set combinations shown in Table 1. The highest average AUC value using 50 selected features for the kNN is 0.9937, and for NLSVM, it is 0.9981. At the same time, LDA and LSVM achieved 0.9940 and 0.9942 respectively for all features. Table 2 summarizes the performance of the proposed method for PDAC detection and compares it with state-of-the-art [31]. The parameters such as accuracy, memory, and processing time are compared. In this work, NLSVM achieved the highest detection accuracy of 98.5 %, 1.1 % more than the state-of-the-art. In addition, the sensitivity of the proposed technique is also 1.2 % more. Similarly, training and testing time of the proposed method for detecting PDAC is reduced by 80.5 % and 28.5 % [31], respectively. Also, the training and the testing time summarized in Table 2 remained the same for all the classifiers except CNN. Furthermore, the proposed method used 337.5 MB of memory, whereas [31] used 2338 MB, which is also 85.6 % more memory utilization. Likewise, the processing time and memory usage of LDA, LSVM, and kNN are significantly less than the [31], but these methods are also less accurate.

3.2. Discussion

KRAS, TP53, and SMAD4 mutation in human cell line have been used in the mouse model for PDAC detection in this study and tumor xenograft mouse model for understanding the biology and mechanism underlying PDAC. Several studies have used similar mouse models because these murine cancer models show high degrees of molecular homology with their human equivalents, which is very informative for preclinical studies; however, variations exist regarding histologic development [52–54]. In addition, murine cancer models have short reproductive cycles, large litter sizes, low cost, and are easy to handle [55]. Each mutation type is shown in detail in Fig. 5. In KRAS G12V, the mutation occurs due to a single nucleotide substitution (thymine replaces guanine in gene sequences) at glycine-encoding codon-12, in which glycine at position 12 is replaced with valine (G12V), as shown in Fig. 5(a–c). In our study, the cause of PDAC is also caused by a mutation in SMAD4. Fig. 5(a) illustrates the SMAD4 signaling pathway, in which SMAD4 is switched off, which uplifts the cell cycle. Similarly, the third type of mutation identified in this paper is TP53, which causes PDAC. TP53 is located on chromosome 17, a phosphoprotein comprising 393 amino acids [56]. When there is damage to the DNA of the cells, the TP53 gene expresses and produces more TP53 tumor suppressor protein.

The major functions of increased TP53 expression and gene regulation are (i) cell cycle arrest, (ii) DNA repair, and (iii) apoptosis (cell death) [57]. Cancer caused by TP53 is either a mutation in the TP53

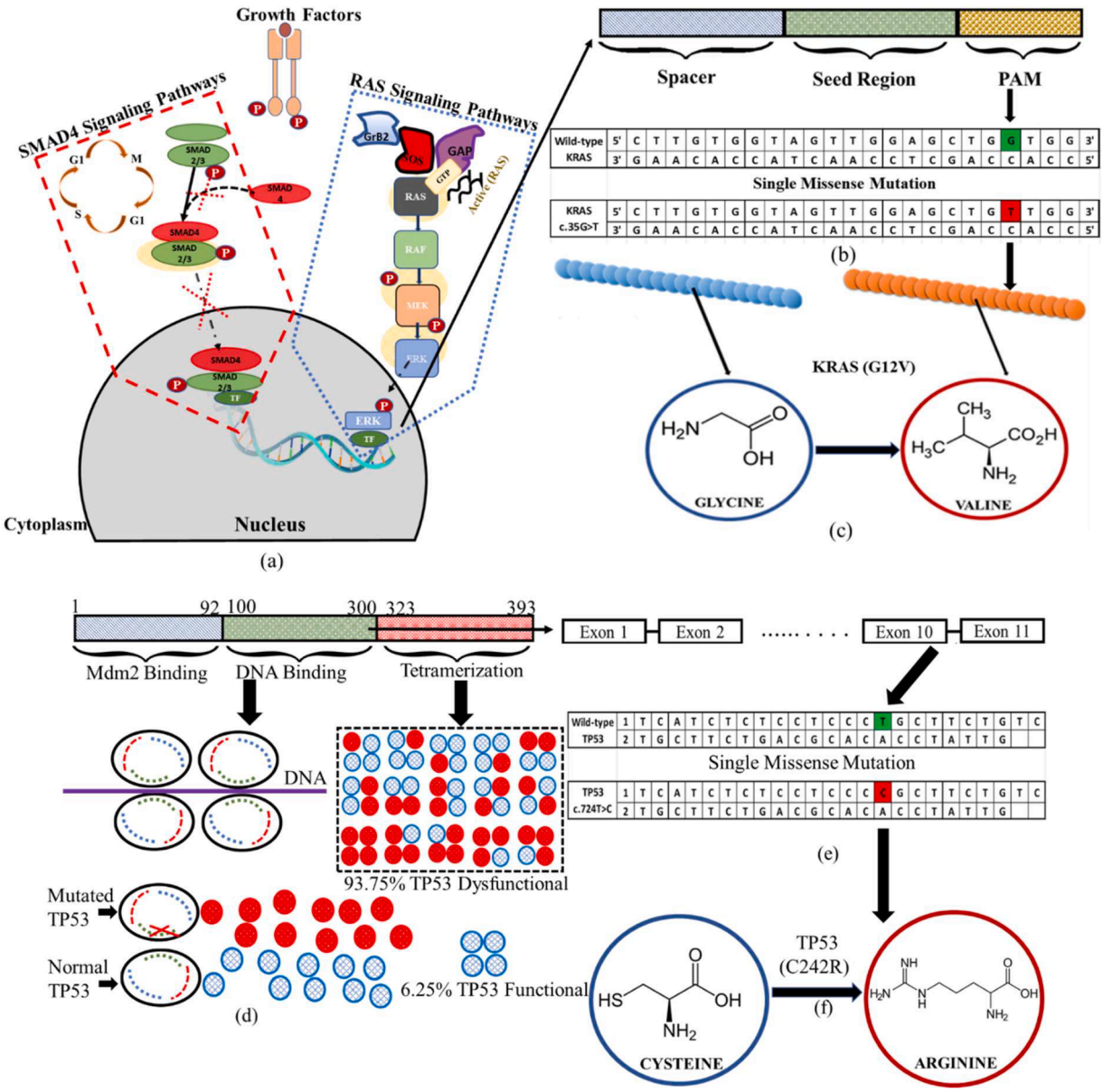


Fig. 5. KRAS, TP53, and SMAD4 (a). RAS and SMAD4 signaling pathways for cell proliferation include active RAS, which sends the signal for cell division to the nucleus. (b) In KRAS, gene mutation at codon 35 from guanine to thymine. (c) Amino acid glycine at position 12 is replaced with valine amino acid. (d) TP53 mutation and cell proliferation: the mutation in the DNA binding region caused a mutation in TP53, which resulted in the development of 93.75 % dysfunctional and 6.25 % functional TP53 protein. (e) In TP53, gene mutation at codon 724 replaces thymine with cytosine. (f) Amino acid cysteine at position 242 is replaced with arginine amino acid.

gene or a protein that regulates TP53 [58]. In this paper, mutated TP53 has c.724T>C gene sequence and p.C242R protein sequence, as shown in Fig. 5(d–f).

This paper defines the fingerprint regions for PDAC detection using peak features that are biologically relevant to the chemical composition of normal pancreas and PDAC, as shown in Fig. 6. The Wild TP53 is composed of cysteine amino acids that contain sulphur, and it has a Raman shift from 605 to 660 cm^{-1}, as shown in Fig. 6. The Peak feature S_{p3} represents the high concentration of sulphur in wild-type TP53. Other bonds in cysteine are represented by the features S_{p17} and S_{p18}, as shown in Fig. 6. Sulphur is replaced with a cationic side chain in mutated TP53, which has a 1075 to 1100 cm^{-1} Raman shift. The S_{p20} feature represents a mutated side chain, as shown in Fig. 6.

Similarly, in wild-type KRAS, the concentration of the amino acid glycine with an H-aliphatic side chain is high. This group has a 1336 cm^{-1} Raman shift, represented by the S_{p30} feature. In mutated KRAS, the glycine amino acid is replaced with valine, which has (CH3) CH-aliphatic side chain. Mutated KRAS is represented by feature S_{p33}, which has a 1365–1451 cm^{-1} Raman shift. Moreover, due to these mutations and SMAD4, uncontrolled cell proliferation occurred, resulting in increased lipid-protein and amide I lipids, which can also be observed from 1400 to 1800 cm^{-1} in Fig. 6. This is a prominent sign of PDAC. Hence, changes in the gene sequence in each mutation type

3.3. Limitation and future work

The current study was limited to murine cancer models. We are interested in using this method to detect other cancers because the proposed method is simple, accurate, and efficient for PDAC detection in mice and could be extrapolated to evaluate human pancreatic cancer and other cancer types. The explainable features extracted from PDAC can be helpful for other Raman spectrum cancer studies. Therefore, the

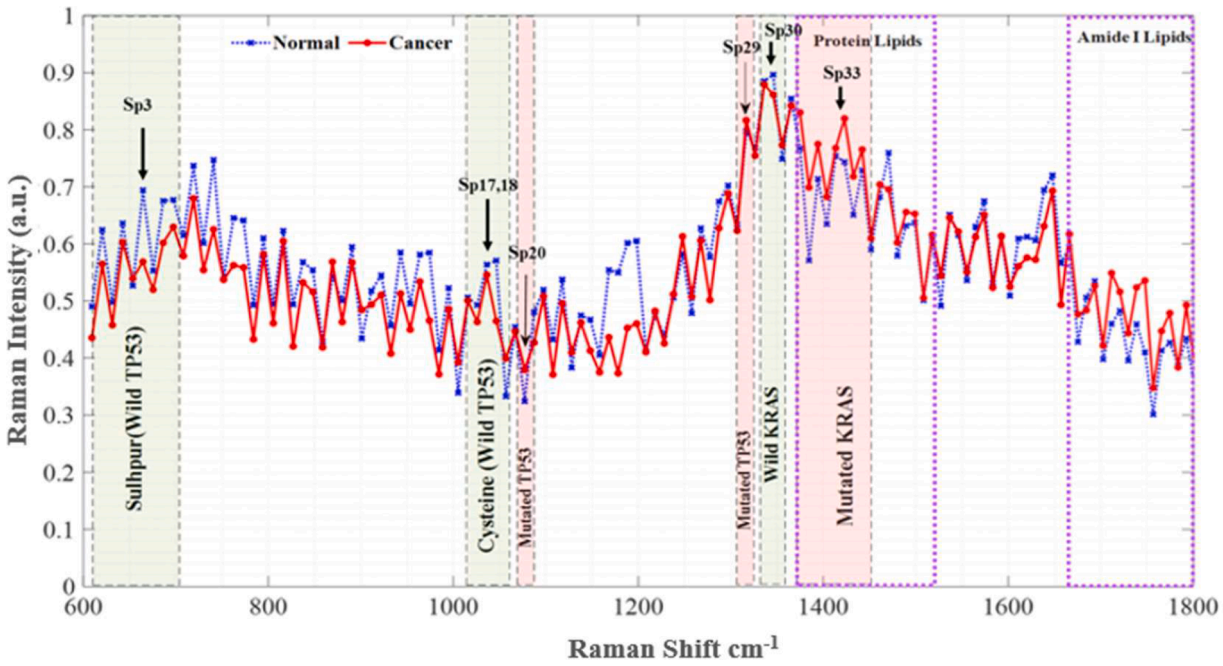


Fig. 6. Biological relevance of average peak features with TP53 and KRAS mutation types result in different amino acid attachment at a particular position in the protein molecule formation process. Table 3 summarizes the mutation types used in the development of PDAC in mice and the changes due to the mutations in the corresponding protein sequences, side groups, chemical structure, and Raman shift in significant peaks.

findings will be evaluated further in conjunction with studies on human pancreatic cancer and other cancer types in the near future.

4. Conclusion

Pancreatic Ductal Adenocarcinoma (PDAC) is a deadly disease, with a less than 10 % survival rate worldwide. It is a progressively more common cause of tumor death and can be cured if detected earlier. This study proposes a novel method for PDAC detection using distinctive explainable features acquired from Raman spectroscopic signals. Our study is the first to define fingerprint regions for PDAC in the literature. This work also reports unique mutation regions in the fingerprint region of PDAC for wild-type and mutated proteins, such as KRAS, TP53, and SMAD4. This study also shows the relationship between the obtained features and Raman shifts in the fingerprint regions. This novel relevance can significantly enhance the accuracy of chemotherapy for early stage PDAC. The dimensions of the feature set are reduced with SVM-RFE integrated with the CBR feature selection method. Subsequently, supervised machine-learning algorithms are used to classify the test samples into cancer and normal classes. The best average classification accuracy of 98.5 % is achieved by NLSVM, which is 1.1 % higher than the state-of-the-art classification accuracy. The method used in this study reduces the test time by 28.5 % and saves 85.6 % memory utilization compared to the existing method. The current study was limited to murine cancer models. The findings will be evaluated further in conjunction with human pancreatic cancer studies in the near future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Statement of Ethical approval

The dataset used in this research was acquired using protocol approved by Institutional Animal Care and Use Committee (IACUC) Louisiana State University.

Acknowledgments

This research is supported by Louisiana State University Faculty Research Grants (009875, 010215), LSU Collaborative Cancer Research Initiative (010163), LSU Leveraging Innovation for Technology Transfer (LIFT2) Grants (LSU-2021-LIFT-009 and LSU-2020-LIFT-008), Louisiana Board of Regents Grant (LEQSF (2018-21)-RD-A-09), the National Science Foundation (NSF) CAREER award (2046929), and Higher Education Commission of Pakistan.

References

[1] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, 2021, CA Cancer J. Clin. 71 (1) (2021) 7–33.
[2] Q. Chen, J. Li, P. Shen, H. Yuan, J. Yin, W. Ge, K. Jiang, Biological functions, mechanisms, and clinical significance of circular RNA in pancreatic cancer: a promising rising star, Cell Biosci. 12 (1) (2022) 1–26.
[3] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (3) (2021) 209–249.
[4] J.D. Mizrahi, R. Surana, J.W. Valle, R.T. Shroff, Pancreatic cancer, Lancet 395 (10242) (2020) 2008–2020.
[5] Z. Li, Z. Li, A. Ramos, J.P. Boudreaux, R. Thiagarajan, Y.B. Mattison, J. Xu, Detection of pancreatic cancer by indocyanine green-assisted fluorescence imaging in the first and second near-infrared windows, Cancer Commun. 41 (12) (2021) 1431.
[6] J. Luo, KRAS mutation in pancreatic cancer, in: Seminars in Oncology, 48, WB Saunders, 2021, pp. 10–18.
[7] D. Uprety, A.A. Adjei, KRAS: from undruggable to a druggable cancer target, Cancer Treat. Rev. 89 (2020) 102070.
[8] AACR Project Genie Consortium, AACR Project GENIE Consortium, F. André, M. Arnedos, A.S. Baras, J. Baselga, H Zhang, AACR Project GENIE: powering precision medicine through an international consortium, Cancer Discov. 7 (8) (2017) 818–831.
[9] A.J. Levine, M. Oren, The first 30 years of p53: growing ever more complex, Nat. Rev. Cancer 9 (10) (2009) 749–758.
[10] W. Feroz, A.M.A. Sheikh, Exploring the multiple roles of guardian of the genome: P53, Egypt. J. Med. Human Genet. 21 (1) (2020) 1–23.
[11] S. Sengupta, C.C. Harris, p53: traffic cop at the crossroads of DNA repair and recombination, Nat. Rev. Mol. Cell Biol. 6 (1) (2005) 44–55.

[12] H. Offer, N. Erez, I. Zurer, X. Tang, M. Milyavsky, N. Goldfinger, V. Rotter, The onset of p53-dependent DNA repair or apoptosis is determined by the level of accumulated damaged DNA, Carcinogenesis 23 (6) (2002) 1025–1032.
[13] U.M. Moll, N. Slade, p63 and p73: roles in development and tumor formation, Mol. Cancer Res. 2 (7) (2004) 371–386.
[14] M.P. de Caestecker, E. Piek, A.B. Roberts, Role of transforming growth factor-β signaling in cancer, J. Natl. Cancer Inst. 92 (17) (2000) 1388–1402.
[15] Z. Krška, J. Šváb, D. Hoskovec, J. Ulrych, Pancreatic cancer diagnostics and treatment–current state, Prague Med. Rep. 116 (4) (2015) 253–267.
[16] L. Kubiczkova, L. Sedlarikova, R. Hajek, S. Sevcikova, TGF-β–an excellent servant but a bad master, J. Transl. Med. 10 (2012) 1–24.
[17] J. Xu, D.A. Kooby, S. Nie, Nanofluorophore assisted fluorescence image-guided cancer surgery, J. Med. Clin. Res. Rev 2 (2018) 1–3.
[18] J. Xu, D. Kooby, B. Kairdolf, S. Nie, New horizons in intraoperative diagnostics of cancer in image and spectroscopy guided pancreatic cancer surgery, New Horizons Clin. Case Rep. 1 (2017) 2.
[19] D.P. Singh, S. Sheedy, A.H. Goenka, M. Wells, N.J. Lee, J. Barlow, S.T. Chari, Computerized tomography scan in pre-diagnostic pancreatic ductal adenocarcinoma: stages of progression and potential benefits of early intervention: a retrospective study, Pancreatology 20 (7) (2020) 1495–1501.
[20] C.G. Guo, S. Ren, X. Chen, Q.D. Wang, W.B. Xiao, J.F. Zhang, Z.Q. Wang, Pancreatic neuroendocrine tumor: prediction of the tumor grade using magnetic resonance imaging findings and texture analysis with 3-T magnetic resonance, Cancer Manage. Res. 11 (2019) 1933.
[21] X. Zhang, L. Detering, D. Sultan, H. Luehmann, L. Li, G.S. Heo, Y. Liu, CC chemokine receptor 2-targeting copper nanoparticles for positron emission tomography-guided delivery of gemcitabine for pancreatic ductal adenocarcinoma, ACS Nano 15 (1) (2021) 1186–1198.
[22] K. Kurihara, K. Hanada, A. Shimizu, Endoscopic ultrasonography diagnosis of early pancreatic cancer, Diagnostics 10 (12) (2020) 1086.
[23] L. Zhang, S. Sanagapalli, A. Stoita, Challenges in diagnosis of pancreatic cancer, World J. Gastroenterol. 24 (19) (2018) 2047.
[24] Z. Li, Z. Li, Q. Chen, J. Zhang, M.E. Dunham, A.J. McWhorter, J. Xu, Machine-learning-assisted spontaneous Raman spectroscopy classification and feature extraction for the diagnosis of human laryngeal cancer, Comput. Biol. Med. 146 (2022) 105617.
[25] I.P. Santos, E.M. Barroso, T.C.B. Schut, P.J. Caspers, C.G. van Lanschot, D.H. Choi, S. Koljenović, Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics, Analyst 142 (17) (2017) 3025–3047.
[26] M. Jermyn, K. Mok, J. Mercier, J. Desroches, J. Pichette, K. Saint-Arnaud, F. Leblond, Intraoperative brain cancer detection with Raman spectroscopy in humans, Sci. Transl. Med. 7 (274) (2015), 274ra19-274ra19.
[27] C.H. Liu, Y. Zhou, Y. Sun, J.Y. Li, L.X. Zhou, S. Boydston-White, R.R. Alfano, Resonance Raman and Raman spectroscopy for breast cancer detection, Technol. Cancer Res. Treat. 12 (4) (2013) 371–382.
[28] J.L. González-Solís, J.C. Martínez-Espinosa, L.A. Torres-González, A. Aguilar-Lemarroy, L.F. Jave-Suárez, P. Palomares-Anda, Cervical cancer detection based on serum sample Raman spectroscopy, Lasers Med. Sci. 29 (3) (2014) 979–985.
[29] J. Zhao, H. Lui, S. Kalia, H. Zeng, Real-time Raman spectroscopy for automatic in vivo skin cancer detection: an independent validation, Anal. Bioanal. Chem. 407 (27) (2015) 8373–8379.
[30] A. Mathew, P. Amudha, S. Sivakumari, Deep learning techniques: an overview. International Conference On Advanced Machine Learning Technologies and Applications, Springer, Singapore, 2020, pp. 599–608.
[31] Z. Li, Z. Li, Q. Chen, A. Ramos, J. Zhang, J.P. Boudreaux, J. Xu, Detection of pancreatic cancer by convolutional-neural-network-assisted spontaneous Raman spectroscopy with critical feature visualization, Neural Networks 144 (2021) 455–464.
[32] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, A. Binder, Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, Sci. Rep. 10 (1) (2020) 1–12.
[33] J.W. Mack, A. Walling, S. Dy, A.L.M. Antonio, J. Adams, N.L. Keating, D. Tisnado, Patient beliefs that chemotherapy may be curative, and care received at the end of life among patients with metastatic lung and colorectal cancer, Cancer 121 (11) (2015) 1891–1897.
[34] L. Bote-Curiel, S. Munoz-Romero, A. Gerrero-Curieses, J.L. Rojo-Álvarez, Deep learning and big data in healthcare: a double review for critical beginners, Appl. Sci. 9 (11) (2019) 2331.
[35] M. Faal, F. Almasganj, ECG Signal modeling using volatility properties: its application in sleep apnea syndrome, J. Healthc Eng. (2021).
[36] R.R. Rajanna, N. Sriraam, P.R. Vittal, U. Arun, Performance evaluation of woven conductive dry textile electrodes for continuous ECG signals acquisition, IEEE Sens. J. 20 (3) (2019) 1573–1581.
[37] M.M. Shidore, S.S. Athreya, S. Deshpande, R. Jalnekar, Screening of knee-joint vibroarthrographic signals using time and spectral domain features, Biomed. Signal Process. Control 68 (2021) 102808.
[38] A. Arasteh, M.H. Moradi, A. Janghorbani, A novel method based on empirical mode decomposition for P300-based detection of deception, IEEE Trans. Inf. Forens. Sec. 11 (11) (2016) 2584–2593.
[39] C. Li, X. Wang, Z. Tao, Q. Wang, S. Du, Extraction of time varying information from noisy signals: an approach based on the empirical mode decomposition, Mech. Syst. Signal Process. 25 (3) (2011) 812–820.
[40] L. Yang, P. Wei, C. Zhong, Z. Meng, P. Wang, Y.Y. Tang, A fractal dimension and empirical mode decomposition-based method for protein sequence analysis, Int. J. Patt. Recognit. Artif. Intell. 33 (11) (2019) 1940020.
[41] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[42] P.A. Mundra, J.C. Rajapakse, SVM-RFE with MRMR filter for gene selection, IEEE Trans. Nanobiosci. 9 (1) (2009) 31–37.
[43] S.X. Li, Q.Y. Zeng, L.F. Li, Y.J. Zhang, M.M. Wan, Z.M. Liu, S.H. Liu, Study of support vector machine and serum surface-enhanced Raman spectroscopy for noninvasive esophageal cancer detection, J. Biomed. Opt. 18 (2) (2013) 027008.
[44] S. Li, Y. Zhang, J. Xu, L. Li, Q. Zeng, L. Lin, S. Liu, Noninvasive prostate cancer screening based on serum surface-enhanced Raman spectroscopy and support vector machine, Appl. Phys. Lett. 105 (9) (2014) 091104.
[45] Y. Yu, Y. Lin, C. Xu, K. Lin, Q. Ye, X. Wang, J. Lin, Label-free detection of nasopharyngeal and liver cancer using surface-enhanced Raman spectroscopy and partial lease squares combined with support vector machine, Biomed. Opt. Express 9 (12) (2018) 6053–6066.
[46] X. Li, T. Yang, S. Li, D. Wang, Y. Song, S. Zhang, Raman spectroscopy combined with principal component analysis and k nearest neighbour analysis for non-invasive detection of colon cancer, Laser Phys. 26 (3) (2016) 035702.
[47] X. Cui, Z. Zhao, G. Zhang, S. Chen, Y. Zhao, J. Lu, Analysis and classification of kidney stones based on Raman spectroscopy, Biomed. Opt. Express 9 (9) (2018) 4175–4183.
[48] X. Zheng, G. Lv, Y. Zhang, X. Lv, Z. Gao, J. Tang, J. Mo, Rapid and non-invasive screening of high renin hypertension using Raman spectroscopy and different classification algorithms, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 215 (2019) 244–248.
[49] C.M. Florkowski, Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests, Clin. Biochem. Rev. 29 (Suppl 1) (2008) S83.
[50] K. Yan, D. Zhang, Feature selection and analysis on correlated gas sensor data with recursive feature elimination, Sens. Actuators B: Chem. 212 (2015) 353–363.
[51] L. Hamel, Model assessment with ROC curves. Encyclopedia of Data Warehousing and Mining, IGI Global, 2009, pp. 1316–1323. Second Edition.
[52] M. Cekanova, K. Rathore, Animal models and therapeutic molecular targets of cancer: utility and limitations, Drug Des. Dev. Ther. (2014) 1911–1922.
[53] J.X. Miao, J.Y. Wang, H.Z. Li, H.R. Guo, L.S.C. Dunmall, Z.X. Zhang, Y.H. Wang, Promising xenograft animal model recapitulating the features of human pancreatic cancer, World J. Gastroenterol. 26 (32) (2020) 4802.
[54] C.I. Hwang, S.F. Boj, H. Clevers, D.A. Tuveson, Preclinical models of pancreatic ductal adenocarcinoma, J. Pathol. 238 (2) (2016) 197–204.
[55] N.P. Lee, C.M. Chan, L.N. Tung, H.K. Wang, S. Law, Tumor xenograft animal models for esophageal squamous cell carcinoma, J. Biomed. Sci. 25 (1) (2018) 1–8.
[56] G.K. Maximov, K.G. Maximov, The role of p53 tumor-suppressor protein in apoptosis and cancerogenesis, Biotechnol. Biotechnol.Equip. 22 (2) (2008) 664–668.
[57] T. Riley, E. Sontag, P. Chen, A. Levine, Transcriptional control of human p53-regulated genes, Nat. Rev. mol. Cell Biol. 9 (5) (2008) 402–412.
[58] H.E. Marei, A. Althani, N. Afifi, A. Hasan, T. Caceci, G. Pozzoli, C. Cenciarelli, p53 signaling in cancer progression and therapy, Cancer Cell Int. 21 (1) (2021) 1–15.
[59] E. Podstawka, Y. Ozaki, L.M. Proniewicz, Part I: surface-enhanced Raman spectroscopy investigation of amino acids and their homodipeptides adsorbed on colloidal silver, Appl. Spectrosc. 58 (5) (2004) 570–580.
[60] H. Lee, M.S. Kim, S.W. Suh, Raman spectroscopy of sulfur-containing amino acids and their derivatives adsorbed on silver, J. Raman Spectrosc. 22 (2) (1991) 91–96.
[61] G. Zhu, X. Zhu, Q. Fan, X. Wan, Raman spectra of amino acids and their aqueous solutions, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 78 (3) (2011) 1187–1195.
[62] M. Masetti, G. Acquaviva, M. Visani, G. Tallini, A. Fornelli, M. Ragazzi, D. de Biase, Long-term survivors of pancreatic adenocarcinoma show low rates of genetic alterations in KRAS, TP53 and SMAD4, Cancer Biomarkers 21 (2) (2018) 323–334.