Classification and Prediction of Gastric Cancer from Saliva Diagnosis using Artificial Neural Network

Muhammad Aqeel Aslam, Cuili Xue, Manhua Liu, Kan Wang, Daxiang Cui

Abstract—In medical research, non-invasive diagnostic tools have become an emerging technique for the diagnosis of fatal disease in the last few years. Saliva analysis for the detection of Gastric cancer (GC) also belongs to this powerful new research field. According to the WHO, cancer is heterogeneous disease with different subtypes. Early prognosis and diagnosis are key to improve the survival rate. It has become necessary in cancer research to facilitate the subsequent clinical management of patients. In this study, we have found 10 Amino acid biomarkers in saliva and extracted 19 fingerprint Raman bands produced by these biomarkers, that can be used to distinguish cancer patients from healthy persons. These Amino acid biomarkers vary according to the health condition of the patient. Computer-Aided Diagnostic (CAD) techniques allow us to learn the common and hidden patterns from the input datasets and predict the cancer status most accurately and efficiently. We have developed a multilayer feedforward neural network using a scaled conjugate gradient backpropagation technique. The proposed method produces an accuracy of 92.27%, sensitivity of 94.8 %, and specificity of 90.2%. In conclusion, our approach using the saliva analysis and Amino acid biomarkers in saliva has enabled us to reliably detect gastric cancer at very high accuracy.

Index Terms—Amino acid biomarkers, Artificial Neural Network, Gastric Cancer Classification, Machine Learning, Surface Enhanced Raman Scattering

I. Introduction

Cells are the most sophisticated molecular assemblies of the human body that can respond to the surrounding environment. According to the National Central Cancer Registry (NCCR), in China, the morbidity and mortality rates have increased to a large extent [1].

Manuscript received October 30, 2019; revised Oct 26 2020. This work was supported by 973 project (2017FYA 0205304), National Natural Science Foundation of China (No. 81225010, 81028009 and 31170961), and the Research Fund of Yantai Information Technology Research Institute of Shanghai Jiao Tong University.

Muhammad Aqeel Aslam is a PhD candidate in the School of Electronic Informational and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.

Daxiang Cui is working as Distinguished Professor in the department of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. (Corresponding Author Email: maqeelaslam@hotmail.com / dxcui@sjtu.edu.cn),

Cuiki Xue, Manhua Liu and Kan Wang are with School of Electronic Informational and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China (email: cuili.xue@sjtu.edu.cn; mhliu@sjtu.edu.cn; xk_wa@163.com.

Since the last decade, researchers are trying to develop new methods for the prognosis of cancer disease. A continuous evolution related to cancer research has been formed since the last few years [2]. Scientists have already developed different methods, which also include screening cancer at an early stage [3]. A wide range of diseases can be diagnosed and monitored by current medical technologies, but researchers are developing more sophisticated methods for the early diagnosis of fatal diseases [4]. Current research lines are trying to develop new methods, which can predict cancer at an early stage even before the symptoms occur. This is not only important for cancer, because most diseases have a much higher survival rate if they are diagnosed at an early stage.

Moreover, new methods for the diagnosis of gastric cancer have been developed in the last decades. Most of the disease can be diagnosed and treated clinically but patients suffer heavily from the time-consuming processes. Cancer is actually a heterogeneous disease with different subtypes. Quite a lot of studies have been presented in the literature based on diverse approaches that permit premature cancer investigation and prediction [5], [6], [7], [8]. Explicitly, these studies described methods associated with the profiling of circulating miRNAs that have been established as promising classes for cancer detection and identification. However, the application of these methods in screening cancer at an early phases and differentiating between benign and malignant tumors is limited due to their low sensitivity. Several features regarding the prediction ability of cancer based on gene expression signatures are discussed in references [9], [10]. In these studies, researchers have thoroughly discussed advantages as well as weaknesses of microarrays based cancer prediction methodologies. While gene signatures might expressively improve our ability for prognosis of cancer patients, but very few of them are used clinically. Therefore, more in-depth researches and experimentations along-with larger data sets and more satisfactory validation are needed. There are more than one hundred types of cancer that affect sixty parts of the human body [11]. Metastasis of the primary tumor is responsible for more than 90% of cancer deaths. As Fig. 1 shows, the top ten most common cancers in China according to the new cases reported in the 2018 [12].

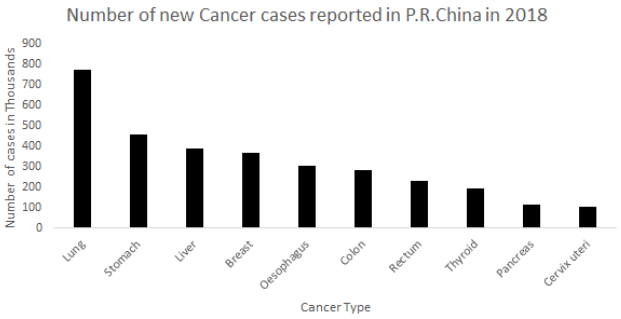

Fig. 1. Ten top most cancers in China

Among all cancer types, gastric cancer (GC) is second-leading cause of cancer-related deaths in China [13], [14]. GC is very common all over the world and China is heavily affected by Gastric Cancer with 42% of worldwide cases [15]. According to the previous studies, the cure rate is up to 90 % if gastric cancer is diagnosed at an early stage, while it is only 24% in the case of late diagnosis.Therefore, early diagnosis is essential to reduce cancer mortality. However, GC patients only present symptoms at an advanced stage [16]. Only Japan has a 90% five-year survival rate due to strict and comprehensive screening [17]. In addition, European countries have a very low survival rate of 10% to 30% [18]. GC is associated with several factors, such as lifestyle, environment, genetic health issues, etc. [19]. Nowadays, the living standards of the individuals have been improved, and the individuals are more aware of their health issues, which results in a decrease in the number of casualties all over the world. However, this cancer is still among the leading cause of deaths around the globe.

GC is divided into two categories, that is Early Gastric Cancer (EGC) (stage I and II), and Advanced Gastric Cancer (AGC) (stage III and IV) [20], [21]. Endoscopy and biopsy are widely used tools for the diagnosis of the Gastric Cancer in the clinic. However, these methods not only bring great discomfort and pain to the patients but also easily miss diagnosis due to the vague symptoms of EGC [22]. In fact, GC patients miss the best time of treatment due to the lack of diagnosis at the early stages [23]. This does not only affect the patients' health but also pushes them under a lot of psychological stress upon GC diagnosis. In recent years, researchers have been focusing on the development of the non-invasive tools for the diagnosis of fatal diseases including GC. Our research focuses on the development of a reliable and non-invasive method for EGC and AGC diagnosis.

In recent years, different Volatile Organic Compounds (VOCs) and mRNA have been extensively studied to diagnose cancer at early stages [24]. Small metabolites have been used for the diagnosis of AGC and EGC, and they are gaining increasing attraction for their more stable and reliable characteristics [25]. Several small metabolite biomarkers have been found in the urine, blood as well as saliva [26].

Recently developed salivary diagnostics and blood biomarker-based screening methods are non-invasive techniques and require less time to diagnose the disease. According to the previous studies by our group, amino acids can be selected as small biomarkers to screen and detect GC among the population. Saliva is a human fluid that can be collected accurately without any medical consideration [27] and is considered to have less interfering physiological chemicals [28]. Therefore, saliva diagnostic is a revolutionary approach for cancer detection [29]. In addition, Surface-Enhanced Raman Sensors are the best candidate for such methodologies as they have the capability of single-molecule detection with signature fingerprint spectra [30].

Machine Learning (ML) belongs to the Artificial Intelligence branch and relates the problem of learning from data samples to the general concept of inference [31], [32], [33]. In recent years, machine learning has shown tremendous attraction to the researchers. ML is playing a vital role in the development of technological advances in many fields. Particularly, medical researchers pay great attention to machine learning which has been used as a significant tool in the medical field. These algorithms have the capability of detecting complex features from the datasets, which are very useful in predicting the type of cancer. The applications of ML techniques have shown that the accuracy of cancer prediction has been increased up to 15 -20 % in the last few years [34]. Machine Learning can be divided into two main categories, (i) supervised learning (ii) unsupervised learning. ML has also been proven to be an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an n-dimensional space for a given set of biological samples, using different techniques and algorithms [35]. The ML process consists of two steps. Firstly, estimate the unknown dependencies from the original dataset, secondly, predict the output from the new data. Every sample is described with the help of several features, and these features may or may not be related. If we already know about the specification of the data, it then becomes easier for us to select the right tools and techniques for the particular dataset. Improvement of data quality and preprocessing steps are very important to make the analysis more stable and reliable. The quality of the

dataset may be affected due to noise, missing data, duplicate data, or biased data. The preprocessing steps lead us to improve the quality of the raw data. ML algorithms perform better when the dimensionality is reduced [36]. In this sense, ML algorithms serve as a set of tools utilized to facilitate prediction, pattern-recognition, and classification in cancer diagnosis. Machine Learning comprises of four steps, including collecting data, selecting the best model, training the selected model, and testing the model [37]. The ML algorithms are able to predict the sequence of genes which are responsible for cancer induction, and determine the prognostic [38], [39]. The use of machine learning algorithms to classify tumors is a very recent development in the medical field. The confusion matrix composed of training set, validation set, and test set. A good detection scheme will produce high positive rate and high negative rate. If the system is not properly trained it will produce high false-positive rate and high false-negative rate [40]. In the last decade, different ML techniques and feature selection algorithms have been adopted by the researchers for the disease prediction, detection, and prognosis [41], [42], [43], [44], [45], [46]. Due to advances in the medical field, most of the diseases can be treated, but the treatment process is quite large, and people suffer a lot during these procedures. Fatal diseases, like cancer, HIV still need to be cured at early stages. Researchers are still focusing to provide solution for their early detection. If we detect fatal diseases in early stages, the high cure rates can be achieved.

In our study, we used the classification method to distinguish gastric cancer patients from healthy persons. In this classification, we put each observation into the category by using a scaled conjugate gradient back propagation classifier. The purpose of this study is to develop a classifier that determines whether the person has gastric cancer or the person is healthy by analyzing saliva samples. We developed an effective machine learning-based classifiers approach for the detection of gastric cancer.

In results section, we stated the performance of each classifier. The performance of each classifier is stated in terms of confusion matrix and Receiver Operating Characteristics (ROC) curve.

II. Materials and Methods

In this section, we will discuss the classification of gastric cancer and our proposed machine learning algorithm

A. Surface-Enhanced Raman Scattering (SERS)

Surface-Enhanced Raman Scattering (SERS) is the modified version of Raman technology, which enhances the Raman signal significantly to realize single-molecule detection. SERS sensors have some advantages over traditional Raman technology, which include high sensitivity, high efficiency and high volumes. Due to all these advantages, SERS technology has been widely used in laboratory material researches, chemical properties of the sample, life sciences, and industrial process control [47], [48], [49], [50], [51]. However, the Raman signal analysis is very difficult when the spectra containing contaminants including (1) sample and background fluorescence (2) cosmic rays, which drift the baseline. We need some preprocessing steps to sort out these problems.

B. Dominant Features

There were hundreds of biomarkers present in the human fluid. According to the previous studies, we found ten amino acids as biomarkers in the saliva which can distinguish the gastric cancer patients from the healthy person. In that study, we used the Mann Whitney U test to compare the metabolic level between GC patients and healthy person, and logistics regression to classify the data. All nineteen Raman bands and

Table 1. The relation between the nineteen bands in Raman spectra as fingerprints and corresponding biomarkers. [21]

Band No.	Band position (cm^{-1})	Biomarkers	Band No.	Band position (cm^{-1})	Biomarkers
1	435	Gln, Hyl, Pro, Tyr	**11**	961	His, Glu, Pro, Tyr
2	488	Tau, Gly, EtN, Hyl, Tyr	**12**	1037	Tau, EtN, Ala, Pro, Tyr
3	530	Tau, Gln, His, Ala, Glu	**13**	1053	Tau, Gln, EtN, Hyl
4	642	His, Ala, Pro, Tyr	**14**	1109	Tau, Gln, EtN, His, Ala
5	725	Tau, Gln, His, Glu	**15**	1197	His, Hyl, Pro, Tyr
6	781	Gln, Glu, Pro, Tyr	**16**	1222	Hyl, Pro, Tyr
7	843	Tau, EtN, His, Ala, Hyl, Pro, Tyr	**17**	1450	Tau, Gly Gln EtN, Ala, Glu, Hyl, Pro
8	869	Gly, Gln, EtN, Glu, Hyl	**18**	1500	His
9	917	Gln, Ala, Glu, Pro	**19**	1710	Gln
10	933	His, Glu, Pro			

corresponding ten amino acids as attributes are listed in Table I.

C. Naïve Bayes Classifier

The Bayesian method uses probabilities and statistics. All variables in NBC are conditionally independent of each other [52]. A model was determined by using the dataset. From the training data we have malignant (M) and benign (B), given representation A, and the probability of occurring event 1 and event 2 is carried out by P(A/B), P(A/M) respectively. P(A/M) is the probability that the sample is malignant type, and P(A/B) is the probability that the sample belongs to the benign class. The Bayesian classifier uses the Bayesian formula to calculate the probability of the specific event.

$$p(b|m) = \frac{p(m|b)p(b)}{p(m)} \quad (1)$$

In equation (1), p(m) is the probability of the training data and p(b) is the probability of the event b occurring, whereas p (m | b) is the conditional probability of m, when b is given, p (b | m) is the conditional probability of b, when m is given. The Bayesian theorem was used to determine whether the input x_i from the dataset X, belongs to the class s_a or s_m. The s_a and s_b represent two different classes.

$$p(x_i|s_a)P(s_a) > P(x_i|s_b)p(s_b) \quad (2)$$

In equation (2) a and b are not equal and are representing two different positive integers.

D. K-NN Classifier

In classification, a non-parametric technique is used by the K-NN algorithm [53]. We have bi-dimensional feature space where A and C are the training vectors. By using vectors, we want to classify the data c, which is a feature vector. The classification of data c depends upon the k nearest neighbors, k is a positive integer, and generally the value of k<5. The data was classified with the most number of votes presenting in the neighbors. The class of data c is very close to an element if k =1. We used the Euclidian method to calculate the distance between the neighboring vectors.

The Euclidian distance was calculated by the following formula shown in (3).

$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \quad (3)$$

E. Support Vector Machines (SVM) Classifier

Support Vector Machines (SVM) are the supervised machine learning algorithm. After the analysis, it classifies the data. SVM classifier is the best candidate for the binary classification due to its robustness and rapidness. SVM does not be affected by noisy data [54]. Hyperplane was found out in the first step of the SVM. A hyperplane is a line that divides the data into two separate classes. The hyperplane has an optimal linear distance between the classes so that the probability of making the wrong decision is very less. The support vectors are used for the creation of hyperplane.

In this study, to avoid the possible problems associated with SVM with linear kernel, we have used four different types of SVM, including SVM, SVM with linear kernel, SVM with Polynomial Kernel and SVM with Sigmoid Kernel. SVM can predict unseen data accurately. For the given training data (x_i, y_i) for i = 1,….. N such that $x_i \epsilon \mathbb{R}$, and $y_i \epsilon$ {-1,1}. We find out the following result.

$$f(x_i) = \begin{cases} \geq 0, when\ y_i = +1 \\ < 0. when\ y_i = -1 \end{cases} \quad (4)$$

Equation (5) was used to find the linear classification of the dataset.

$$f(x_i) = \omega^T x_i + b \quad (5)$$

Here, ω is the weight vector, and b is the bias. Table II represents the parameters of the different SVM used in this study. These parameters include the SVM type, SVM kernel, cost function, number of support vectors, and specific parameters of each type. Along-with these parameters, SVM with polynomial kernel with 3rd order degree polynomial was used to get the best performance, and two numbers of classes in each type of SVM.

Table II. Parameters of four Support Vector Machine Methods

SVM Type	C-Classification	C-Classification	C-Classification	C-Classification
SVM Kernel	Radial	Linear	Polynomial	Sigmoid
Cost value	1	1	1	1
No. of support vectors	89	55	126	56

F. Proposed Artificial Neural Network

A basic structure of feed-forward neural network with three layers has been shown in Fig. 2 with sigmoid hidden and Softmax output neurons. Neural Network consists of three layers, which are the input layer, hidden layer, and output layer. The input layer brings the data into the system which only deals with the inputs. The input layer forwards the inputs to the hidden layer. The performance of the neural network depends upon the number of neurons present in the hidden layers. In machine learning, one or more layers, known as *hidden layers, are added between input and output so that neurons can learn more complicated features.*

Gastric Cancer classification is a complex system. It was realized by a robust estimation methodology. Neural Network algorithm was used to solve this cancer classification problem. In this study, we have used the pattern classification method for the prediction and classification of gastric cancer patients. Scaled conjugate gradient back propagation was used for the training of the artificial network classifier. This algorithm comprises of two parts. Firstly, it uses a scaled conjugate gradient method to search the optimal distance. Secondly, it uses back propagation for error reduction. This algorithm is computationally fast as it does not perform line search at each iteration.

This algorithm requires the network response to all training inputs be computed several times for each search. The error value is checked at each iteration and the weight value is then updated [55]. It was designed in such a way so that it does not consume time while searching the line. Our proposed ANN algorithm comprises of two networks, back propagation and scaled conjugate gradient.

The back propagation was used to minimize the error of the network. It uses an interaction to reduce the error. At each iteration, the output compares its updated weights with the previous updates. In this way, the error was reduced to a minimal value. The working diagram of the back propagation Neural Network is shown in Fig. 3.

The input vector is $\{x_1, x_2, \ldots\ldots, x_p\}$, the hidden layer vector consists of h_1 and h_2, the output layer vector is composed of y_1 and y_2. Equations (6) and (7) shows the output of each hidden layer. h_1 is the output of the first hidden layer, whereas the h_2 represents the output of the second hidden layer. The weight vector $\{w_1, w_2, \ldots\ldots, w_p\}$ represents the weights, which are multiplied by the input vector to produce the output of the hidden layer.

$$h_1 = f(x_1 * \omega_1 + x_2 * \omega_3 + b_1) \quad (6)$$

$$h_2 = f(x_2 * \omega_4 + x_1 * \omega_2 + b_1) \quad (7)$$

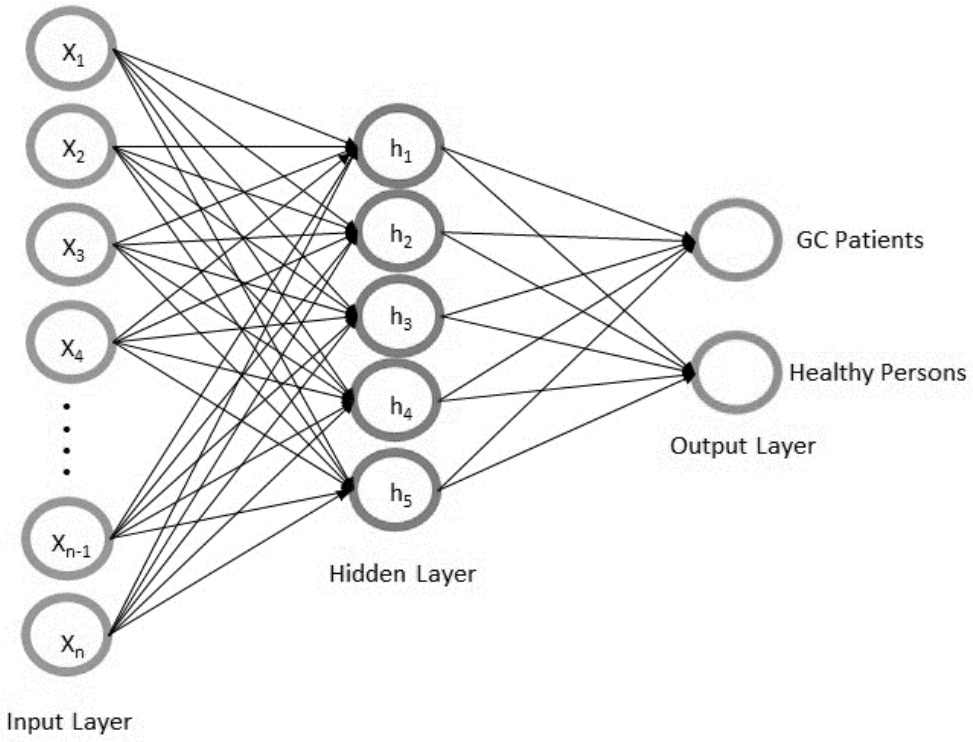

Fig. 2. Basic Structure of Artificial Neural Network

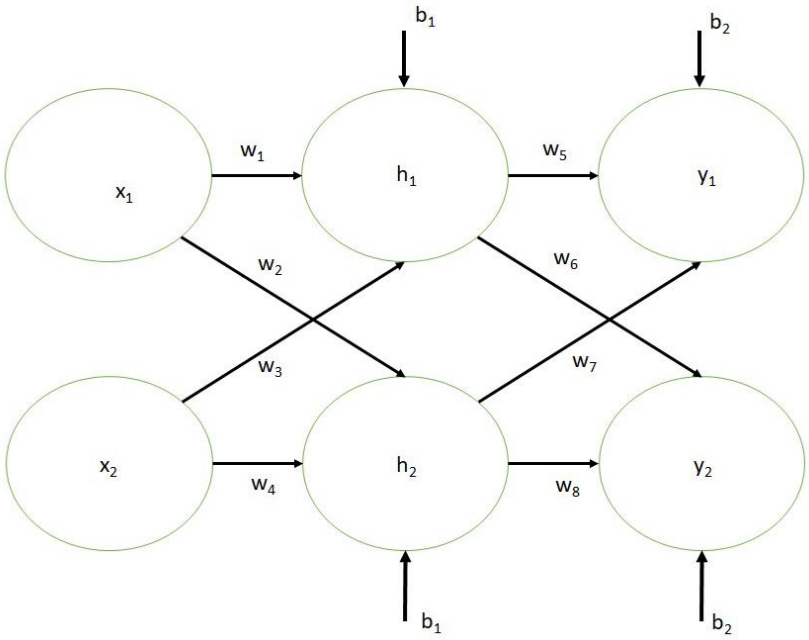

Fig. 3. Working diagram of back propagation Neural Network

The output of the hidden layer acts as an input of the output layer. Equations (8) and (9) represents the output of the 1st output layer and 2nd output layer, respectively.

$$y_1 = f(h_1 * \omega_5 + h_2 * \omega_7 + b_2) \quad (8)$$

$$y_2 = f(h_1 * \omega_6 + h_2 * \omega_8 + b_2) \quad (9)$$

y_1 is the output of the first output layer, whereas the y_2 is the output of the second output layer. *f(.)* is the sigmoid function, which was chosen for the activation function. The sigmoid function was selected for feedforward neural networks [50]. The output produced by this function is only positive values. The sigmoid function is shown in (10).

$$\emptyset(x) = \frac{1}{1 + e^{-x}} \quad (10)$$

The back propagation minimizes the error of the neural network which is defined as in (11):

$$E = \frac{1}{2} \sum_p (t_p - y_p)^T (t_p - y_p) \quad (11)$$

Here, E is the error induced in the network, t_p is the targeted output, y_p is the output of the network, T shows the transpose of the matrix.

The second part of the algorithm is a scaled conjugate gradient. We developed the Scaled Conjugate Gradient (SCG) network, which trains the multilayer feedforward neural network. It has a faster convergence rate. In this technique, a search process was carried out along conjugate directions. New steepest descent direction was combined with previous search direction, these two directions are conjugate to each other. The new steepest descent direction was calculated by these two directions in each iteration. Equations (12) and (13) were used to find out the weight changes in successive steps.

$$\omega_{t+1} = \omega_t + \alpha_t d_t \quad (12)$$

$$d_t = g_t + \beta_t d_{t-1} \quad (13)$$

Here, d_t and d_{t-1} are the conjugate directions in succeeding iterations. The coefficient α_t is used to find the step size, and β_t is used to determine the search direction, g_t and g_{t-1} are the corresponding gradient directions. We have used a scaled conjugate gradient back propagation algorithm for the training of the ANN in this study. The steps of the algorithm are shown in Table III [55].

Table III. Scaled Conjugate Gradient Back propagation steps for Gastric Cancer Classification

Step 1. Split data into two sets, one contains cancer features, second contains cancer targets
Step 2. Initialize the model with Random weights
Step 3. Calculate the output of the network
Step 4. Calculate the error between actual output and targeted output
Step 5. Calculate the network back propagates error
Step 6. Minimize error by adjusting weights
Step 7. Repeat steps 3 to 6, until error is acceptable

This neural network uses a global and multivariate function, which minimizes the error. This function depending on the weights in the network is equivalent to an optimization point of view learning in a neural network. The training of the algorithm stopped if the network reached the maximum number of epochs or execution time exceeded the maximum limit. The output layer is the last layer of the network, producing the result of this classifier.

G. Classifier Performance Testing

Once the classification model is prepared, several parameters are used to evaluate the performance of the trained model. Accuracy, AUC (Area Under Curve), Cross-Entropy, Receiver Operating Characteristics (ROC), Specificity, and Selectivity are the most broadly used parameters. The overall performance of the classifier was assessed by the quantitative metrics of AUC and accuracy. Accuracy shows how many predictions are correctly presented on the targeted data. Cross entropy is used to evaluate the performance of the proposed ANN. It calculates the performance in the form of targets and outputs. If the value of the cross-entropy is smaller than the classification results, the performance of the classifier is good. The cross-entropy is calculated for a pair of target output elements. The classification outcome is represented by the ROC curve. These curves have been used to evaluate the predictive ability of the classifier. The performance of the system is measured with the following parameters. In our study, we have two outputs, which are healthy person or the cancer patient.

True Positive (TP): When the person is a cancer patient, and the neural network also recognizes it as cancer patients.

True Negative (TN): When the person is healthy, and the neural network also calculates it as healthy person.

False Positive (FP): When a person was labeled as a healthy person, but the neural network classifies it as a cancer patient.

False Negative (FN): When a person was labeled as a cancer patient, but the neural network predicts it as a healthy person.

True positive is also known as sensitivity, which presents the ability of the classifier to identify the disease correctly. The true negative rate also called specificity, which is the ability of the classifier to identify those who do not possess the disease correctly. TN rate or specificity is a measure of the classifier to detect cancer patients. Selectivity measures the classifier's ability to reject the false detection of a healthy person. The detection rate is defined as an average of sensitivity and specificity. These parameters are calculated from equations as shown in (14) - (17).

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

$$Selectivity = \frac{TP}{TP + FP} \tag{16}$$

$$Detection\ Rate = \frac{Sensitivity + Specificity}{2} \tag{17}$$

The performance of each classifier was examined by ROC curves. Furthermore, ANN architecture was developed, trained and tested using routines written in MATLAB toolbox 10.2 (The MathWorks, Natick MA). However, the routines of K-NN classifier, SVM classifier and NB classifier were written in the R-studio package.

III. Experimental Work

A. Preparation of Dataset

We prepared the dataset, which has to be in a suitable form for the machine learning algorithm. There are two hundred twenty (220) volunteers, who participated in this research work. We collected their saliva and prepared the dataset. There were twenty-two attributes in each saliva sample. Out of the twenty-two, nineteen were volatile biomarkers. We excluded the patient's name, age, and gender, as this information is not to be feed in the classification algorithm. We found ten Amino biomarkers corresponding to nineteen SERS spectra fingerprint bands which were responsible for the distinguishing the cancer patients from the non-cancerous persons. Table IV shows the number of the person from each class.

Table IV. Clinical Characteristics of Volunteers

Group	Number	Age (Year)	Gender (M : F)
GC	104	53 ± 9	63 : 41
Controls	116	35± 10	67: 49

B. Data Preprocessing

After collecting the saliva samples, several preprocessing steps were carried out for the preparation of the dataset. The data obtained from the Surface Enhanced Raman Spectroscopy, the data needs to be processed before further used. Pre-processing of Raman spectra is necessary for further steps. We reduced and eliminate the irrelevant, random, and systemic variations from the obtained data. As Fig. 4 describes, the complete flowchart of data pre-processing steps.

When the saliva hit the detector, spikes were originated, which is a single event. Spikes have positive peaks with narrow bandwidth. Spikes are random in time produced due to random positions on the sensors. As we know, the bandwidth of spike is much smaller with the comparison to the Raman spectra, we use this assumption and remove the spikes from the Raman data. After removing the spikes from the data, we smoothened the data. As noise is random and it has a very high frequency than the Raman data. We wanted to remove this high-frequency signal which has degrade the information. Several methods can be used for denoising the data including average filter, median filter. We used a median filter for the denoising purpose, which belongs to the non-linear filter.

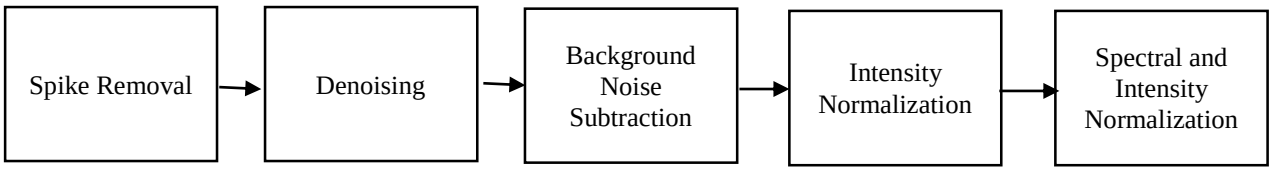

Fig. 4. Preprocessing Steps of Raman Data

The median filter preserves edges while eliminating the noise effectively. By using the median filter, we have reduced the noise, and in consequence, the SNR has been improved to a great extent. We have used a second-order median filter which has defined by (18).

$$M = \begin{cases} \frac{x(N-1)}{2} & N\ is\ odd \\ \frac{1}{2}\left[x\left(\frac{N}{2}\right) + x\left(\frac{N}{2}+1\right)\right] & N\ is\ even \end{cases} \quad (18)$$

Baseline correction is an essential part of the preprocessing of the Raman spectra to avoid overfitting problem, which produces negative values in the data or left-over background. Baseline correction does not cut down Raman band signal strength [56]. All the preprocessing steps were carried out using LabSpec software. As Fig. 5 shows, the Raman spectrum before and after the baseline correction.

C. Parameters of proposed ANN

We developed three different models during this study, and each of them has a different number of neurons in the hidden layer. In the second experiment, we developed an ANN using 10 neurons in the hidden layer. For the third experiment, the number of neurons set to 10 for the first hidden layer and 20 number of neurons for the second hidden layer. The parameters for single layer architecture and multilayer architecture based ANN is shown in Table V.

The dataset consists of 220 samples, which was divided into three parts, training data, validation data, and test data. After the completion of the training of each classifier, the test data was used to evaluate the performance of each stated method. In last, we compared the performance of the proposed multilayer feedforward classifier with NB, SVM and K-NN classifiers. All the methods studied are presented in Table VI.

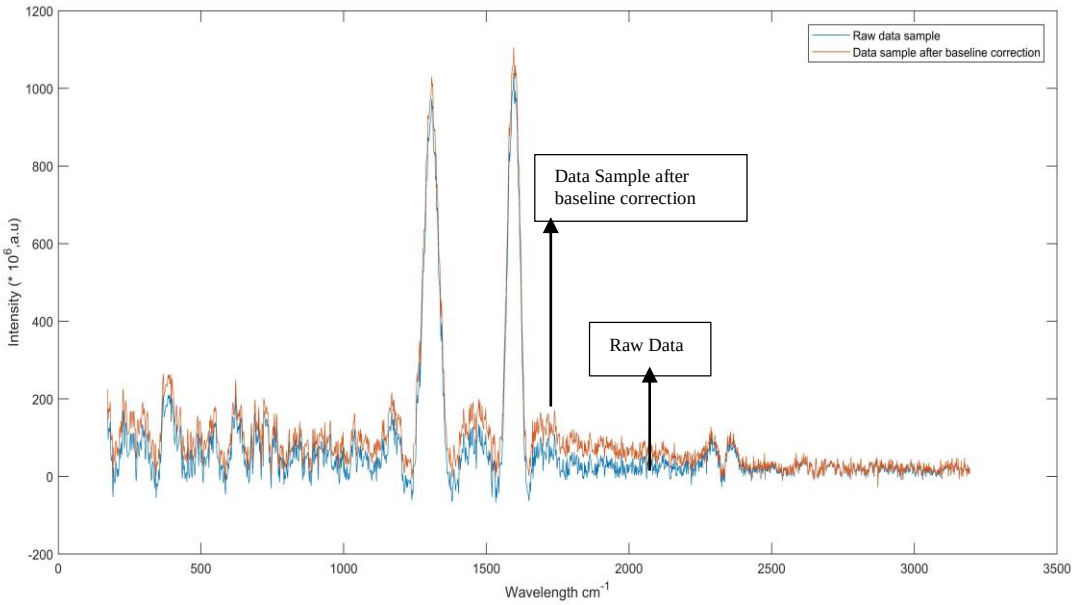

Fig. 5. Before and After Baseline Correction of the Input Sample

Table V. Parameters for Single Layer and Multilayer Feedforward Neural Network

Experiment No	Number of Neurons in 1st Hidden Layer	Number of Neurons in 2nd Hidden Layer	Number of Samples Training Data	Number of Samples Validation Data	Number of Samples Test Data
1.	10	-	154	33	33
2.	10	20	154	33	33

Table VI. Neural Network Models for this study

Acronyms	Description
NBC	Naïve Bayes Classifier
SVM	Support Vector Machine
K-NN	K- Nearest Neighbor Classifier
ANN	Artificial Neural Network
ML-ANN	Multi-Layer Artificial Neural Network

IV. Experimental Results and Discussions

The pattern produced by amino acid biomarkers is an important parameter for the classification of gastric cancer patients. These SERS spectra bands of Amino acid biomarkers were used to distinguish gastric cancer patients from healthy persons. The raw data was collected from SERS and then converted into useful information. Two columns were selected from the complete data of the sensors. Software LabSpec was used to collect the useful information in the form of text files. These text files are then used in MATLAB as input files. Fig. 6 shows, the result of one saliva sample. On the abscissa and ordinate, we have the wavelength and the intensity respectively. The nineteen dominant features were extracted from each saliva sample. After preprocessing, the proposed model was applied to the dataset to discriminate the GC patients from healthy persons.

In this study, the aim was to develop a good architecture for the GC classification from saliva samples using ANN. The performance of the proposed neural network was investigated under the impact of the number of neurons in the hidden layer. To justify the performance of the proposed methods, we compared the results of all the methods mentioned in Table VI. We used 187 samples to train and validate, and 33 samples to test the neural network.

In first experiment, we developed some common techniques of data mining, which include Naive Bayes classifier, K-NN classifier and SVM with four different kernels. The SVM classifier with linear kernel based neural network performed well for the dataset and produced an accuracy of >89%, whereas the NB classifier produced an overall accuracy of 65.45%. As Fig. 7 shows, the ROC curve for six different data mining schemes.

From Fig. 7, we can see that NBC and SVM with sigmoid kernel have the lowest area under the curve, and SVM with linear kernel has the maximum area under the curve, which is followed by K-NN classifier and SVM with polynomial kernel. We can conclude that among these data mining methods, SVM classifier with linear kernel can predicted malignant and healthy samples more accurately and effectively. Although, SVM classifier with polynomial based kernel has 100% sensitivity, but this model has misclassified 27 malignant samples.

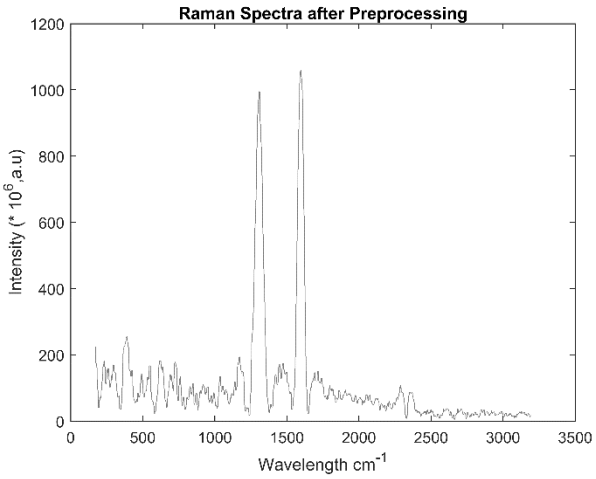

Fig. 6. Raman spectrum of Saliva Sample after data preprocessing

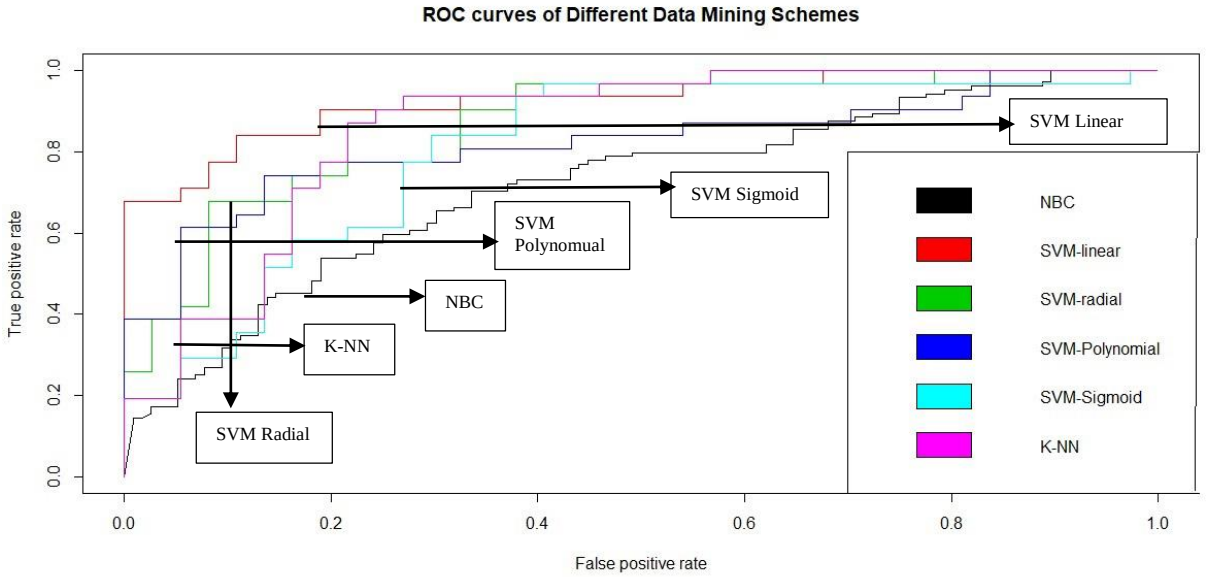

Fig. 7. ROC Curves of Data Mining Schemes for this study

In the second experiment, we developed a single layer ANN-based model to distinguish the GC patients from healthy persons. Once we completed the training of the model, we tested the developed model with the test data. The single hidden layer consists of 10 neurons. The single-layer ANN-based model has distinguished 89 GC samples and 101 healthy samples accurately. However, the model has misclassified 15 samples from each class, producing an accuracy of 86.4%. In particular, the model has produced 87.9% and 84.8% accuracy for training and test data, respectively. The sensitivity is 85.6%, specificity is 87.4% and detection rate is 86.5%. As Fig. 8 represents, the ROC curves for single layer based neural network.

In the third experiment, we developed multilayer feedforward neural network. The multilayer feedforward neural network has outperformed the other methods. As Fig. 9 shows, that this model has produced an overall accuracy of 92.3% and an accuracy of 90.9% on test data.

This method produced a specificity of 90.2%, sensitivity of 94.8%, selectivity of 88.5% and detection rate of 92.5%. This model has misclassified 17 samples, including 5 from the healthy persons and 12 from the GC patients. The multilayer feedforward neural network has outperformed the other state of art methods.

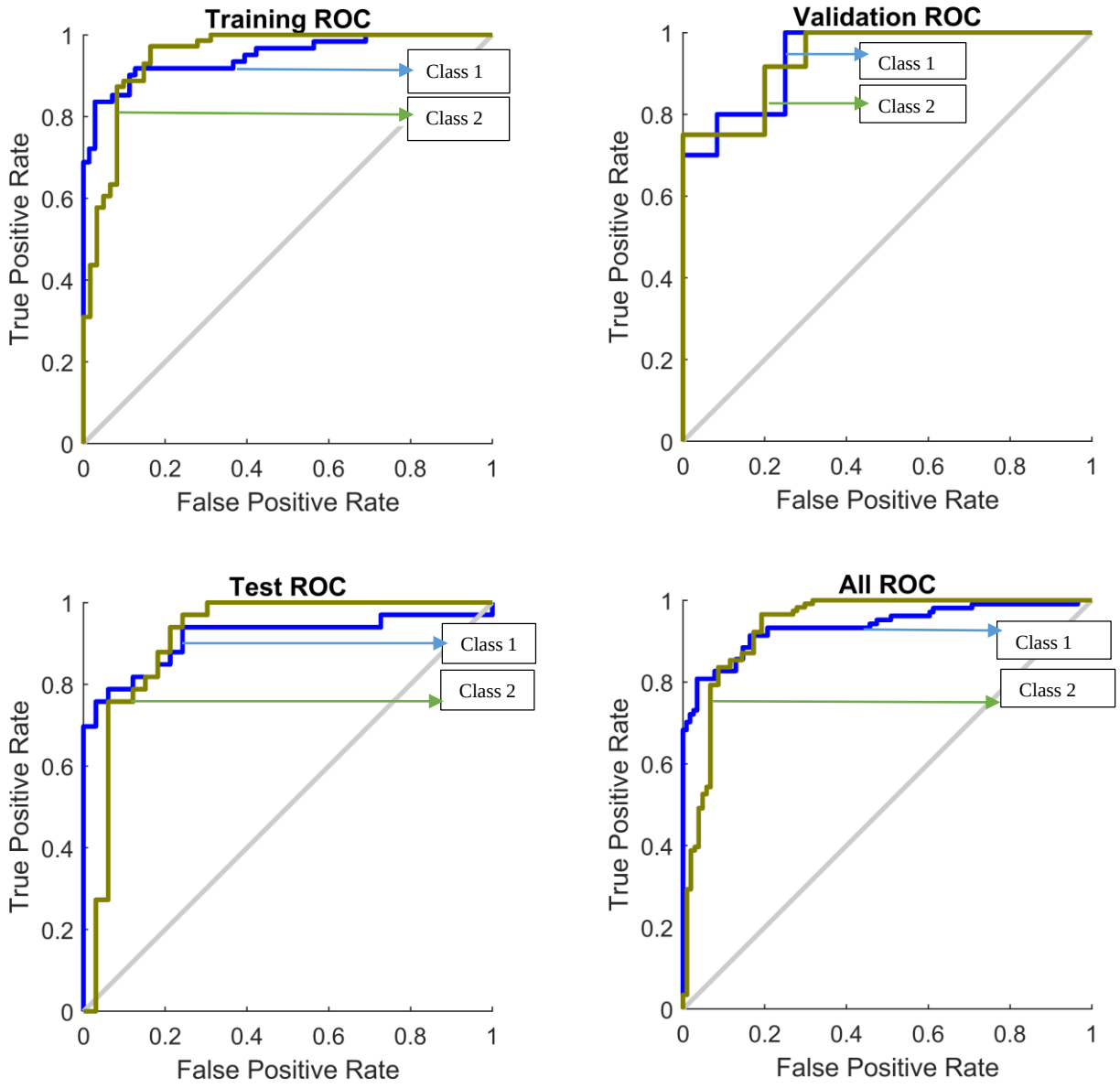

Fig. 8. ROC Curves for single Layer Artificial Neural Network

Fig. 9. Confusiom Matrix for Multilayer Feedforward Neural NEtwork

Table VII. Comparison of different state of art Neural Networks with Proposed Model for Gastric Cancer Classification

Model Name	Accuracy(%)	Selectivity(%)	Specificity(%)	Sensitivity(%)	Detection Rate(%)
Logistic Regression [21]	-	-	87.7	80	-
Nave Bayes Classifier	65.45	41.3	87.01	74.1	80.56
K-NN Classifier	87.7	82.6	85.6	90.5	88.05
SVM Classifier (Linear Kernel)	89.54	88.4	89.7	89.3	89.5
SVM Classifier (Radial Kernel)	85	79.8	87.3	83.2	85.25
SVM Classifier (Polynomial Kernel)	87.7	74.03	81.1	100	90.5
SVM Classifier (Sigmoid Kernel)	69.6	61.5	70.3	68.9	69.6
Single Layer ANN	86.4	85.6	87.4	85.6	86.5
Multilayer ANN	92.27	88.5	90.2	94.8	92.5

Table VII summarized the results of all the methods that have been discussed in this study. We compare the results of this study with the previous results published for this dataset. The proposed architecture has performed exceptionally better than NB, SVM, K-NN, and Logistic Regression schemes.

V. Conclusion

In conclusion, ten Amino acid biomarkers were identified. The SRES spectra bands of these biomarkers were used to differentiate the GC patients from healthy persons using ANN. After extracting the dominant features from the amino acid biomarkers, an artificial neural network was developed. In classification, the sample data was classified into two predefined classes. Based on the above results, our proposed algorithm has produced an accuracy of 92.7%, specificity of 90.2%, and selectivity of 88.5% using a feedforward neural network. The performance of the algorithm depends upon the Softmax and activation function.

Furthermore, it was also observed that the performance of the proposed model depends upon the number of neurons and the number of hidden layers. We have classified Gastric Cancer using the saliva samples with high accuracy by developing a neural network architecture. This model was used to perform classification and prediction. Our trained model can also be used to classify the new data, which has not been seen yet. In this way, we can classify whether a new person is affected by cancer, or the person is healthy.

In the coming years, we need to develop a procedure and system that widely acceptable to saliva sample analysis around the globe. By developing such new emerging methods, we can reduce the number of cancer-related deaths. Not only the number of causalities will decrease, but also we can reduce the mortality rate which is currently very high.

Acknowledgment

The discussion with Prof. Mark I. Ogden in the Department of Chemistry of Curtin University is also gratefully acknowledged.

Contributions: Daxiang Cui conceived and designed the research project. Cuili Xue collected samples and performed the experiments, data acquisition and characterization. Muhammad Aqeel Aslam finished Salivary Diagnostics based Gastric Cancer Classification using Artificial Neural Network and Data mining techniques. Manhua Liu, and Kan Wang discussed about the data analysis strategy, all the authors contributed to the data analysis and the writing of this manuscript, and all authors reviewed the manuscript, and given approval to the final version of the manuscript.

Compliance with Ethical Standards

Ethical Approval: All investigational protocols were evaluated and approved by the Ethical Committee of Shanghai Jiao Tong University.

Informed Consent: Informed consent was obtained from patients in accordance with the guidelines for the conduction of clinical research.

Conflict of Interest: The authors declare they have no conflict of interest, competing interest, financial or otherwise.

References

[1]. Chen W, Zheng R, Zeng H, Zhang S, He J, "Annual Report on Status of Cancer in China, 2011," **Chinese Journal of Cancer Research, vol. 27, no. 1**, pp. 2-12, 2015.

[2]. Hanahan D, Weinberg RA, "Hallmarks of Cancer: The Next Generation," **Cell, vol. 144, no. 5**, pp. 646-674, 2011.

[3]. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI, "Machine Learning Applications in Cancer Prognosis and Prediction," **Computational and Structural Biotechnology Journal, vol. 13**, pp. 8-17, 2015.

[4]. Zheng Y, Wang K, Zhang J, Qin W, Yan X, Shen G, Gao G, Pan F, Cui D, "Simultaneous Quantitative Detection of Helicobacter Pylori Based on a Rapid and Sensitive Testing Platform Using Quantum Dots-Labeled Immunochromatiographic Test Strips," **Nanoscale Research Letters, vol. 11, no. 1**, pp. 1-11, 2016.

[5]. Fortunato O, Boeri M, Verri C, Conte D, Mensah M, Suatoni P, Pastorino U, Sozzi G, "Assessment of Circulating microRNAs in Plasma of Lung Cancer Patients," **Molecules, vol. 19, no. 3,** pp. 3038-3054, 2014.

[6]. Heneghan HM, Miller N, Kerin MJ, "MiRNAs as Biomarkers and Therapeutic Targets in Cancer," **Current Opinion in Pharmacology, vol. 10, no. 5**, pp. 543-550, 2010.

[7]. Madhavan D, Cuk K, Burwinkel B, Yang R, "Cancer Diagnosis and Prognosis Decoded by Blood-Based Circulating microRNA Signatures," **Frontiers in Genetics, vol. 4**, pp. 1-13, 2013.

[8]. Zen K, Zhang CY, "Circulating microRNAs: A Novel Class of Biomarkers to Diagnose and Monitor Human Cancers," **Medicinal Research Reviews**, **vol. 32, no. 2**, pp. 326-348, 2012.

[9]. Koscielny S, "Why Most Gene Expression Signatures of Tumors have not been Useful in the Clinic," **Science Translational Medicine, vol 2, no. 14**, pp. 14ps2, 2010.

[10]. Michiels S, Koscielny S, Hill C, "Prediction of Cancer Outcome with Microarrays: A Multiple Random Validation Strategy," **The Lancet, vol. 365, no. 9458**, pp. 488-492, 2005.

[11]. Qin W, Wang K, Xiao K, Hou Y, Lu W, Xu H, Wo Y, Feng S, Cui D, "Carcinoembryonic Antigen Detection with "Handing"-controlled Fluorescence Spectroscopy Using a Color Matrix for Point-of-Care Applications," **Biosensors and Bioelectronics, vol. 90**, pp. 508-515, 2017.

[12]. The Global Cancer Observatory, China, International Agency for Research on Cancer, 2018, https://gco.iarc.fr/today/data/factsheets/populations/160-china-fact-sheets.pdf. Accessed: 06 June 2019.

[13]. Hartgrink HH, Jansen EP, van Grieken NC, van de Velde CJ, "Gastric Cancer". **The Lancet, vol. 374, no. 9688**, pp. 477-490, 2009.

[14]. Yang L, Zhu H, Wei B, Yao L, Su C, Mu Y, "Construction, Structural Modeling of a Novel ScFv Against Human Gastric Cancer from Phage-Display Library," **Nano Biomed Eng, vol. 3, no. 1**, pp. 29-33, 2011.

[15]. Washington K, "7th Edition of The AJCC Cancer Staging Manual: Stomach," **Annals of Surgical Oncology, vol. 17, no. 12**, pp. 3077-3079, 2010.

[16]. Sitarz R, Skierucha M, Mielko J, Offerhaus GJ, Maciejewski R, Polkowski WP, "Gastric Cancer: Epidemiology, Prevention, Classification, and Treatment," **Cancer Management and Research, vol. 10**, pp. 239-248, 2018.

[17]. Stock M, Otto F, "Gene Deregulation in Gastric Cancer," **Gene, vol. 360, no. 1**, pp. 1-19, 2005.

[18]. Parkin DM, Bray F, Ferlay J, Pisani P, "Global Cancer Statistics, 2002," **CA: A Cancer Journal for Clinicians, vol. 55, no. 2**, pp. 74-108, 2005.

[19]. Wadhwa R, Song S, Lee JS, Yao Y, Wei Q, Ajani JA, "Gastric Cancer—Molecular and Clinical Dimensions," **Nature Reviews Clinical Oncology, vol. 10, no. 11**, pp. 643-655, 2013.

[20]. Ooki A, Yamashita K, Kikuchi S, Sakuramoto S, Katada N, Watanabe M, "Phosphatase of Regenerating Liver-3 As a Prognostic Biomarker in Histologically Node-Negative Gastric Cancer," **Oncology Reports, vol. 21, no. 6**, pp. 1467-1475, 2009.

[21]. Chen Y, Cheng S, Zhang A, Song J, Chang J, Wang K, Zhang Y, Li S, Liu H, Alfranca G, Aslam MA, "Salivary Analysis Based on Surface Enhanced Raman Scattering Sensors Distinguishes Early and Advanced Gastric Cancer Patients from Healthy Persons," **Journal of Biomedical Nanotechnology, vol. 14, no. 10**, pp. 1773-1784, 2018.

[22]. Axon A, "Symptoms and Diagnosis of Gastric Cancer at Early Curable Stage," **Best Practice & Research Clinical Gastroenterology, vol. 20, no. 4**, pp. 697-708, 2006.

[23]. Yazici O, Sendur MA, Ozdemir N, Aksoy S, "Targeted Therapies in Gastric Cancer and Future Perspectives," **World Journal of Gastroenterology, vol. 22, no. 2**, pp. 471-489, 2016.

[24]. Chen Y, Zhang Y, Pan F, Liu J, Wang K, Zhang C, Cheng S, Lu L, Zhang W, Zhang Z, Zhi X, "Breath Analysis Based on Surface-Enhanced Raman Scattering Sensors Distinguishes Early and Advanced Gastric Cancer Patients from Healthy Persons," **ACS Nano, vol. 10, no. 9**, pp. 8169-8179, 2016.

[25]. Abate-Shen C, Shen MM, "The Prostate-Cancer Metabolome," **Nature, vol. 457, no. 7231**, pp. 799-800, 2009.

[26]. Chen Y, Zhang J, Guo L, Liu L, Wen J, Xu L, Yan M, Li Z, Zhang X, Nan P, Jiang J, "A Characteristic Biosignature for Discrimination of Gastric Cancer from Healthy Population by High Throughput GC-MS Analysis," **Oncotarget, vol. 7, no. 52**, pp. 87496-87510, 2016.

[27]. Hofman LF, "Human Saliva as a Diagnostic Specimen," **The Journal of Nutrition, vol. 131, no. 5**, pp. 1621S-1625S, 2001.

[28]. Chamberlain J, "The Analysis of Drugs in Biological Fluids 2nd Edition," **CRC Press,** pp. 35-66, 1995.

[29]. Frederich M, Pirotte B, Fillet M, De Tullio P, "Metabolomics as a Challenging Approach for Medicinal Chemistry and Personalized Medicine," **Journal of Medicinal Chemistry, vol. 59, no. 19**, pp. 8649-8666, 2016.

[30]. Koo KM, McNamara B, Wee EJ, Wang Y, Trau M, "Rapid and Sensitive Fusion Gene Detection in Prostate Cancer Urinary Specimens by Label-Free Surface-Enhanced Raman Scattering," **Journal of Biomedical Nanotechnology, vol. 12, no. 9**, pp. 1798-1805, 2016.

[31]. Bishop CM, "Pattern Recognition and Machine Learning," **Springer**, New York, USA, 2006.

[32]. Mitchell TM, "The Discipline of Machine Learning," **Pittsburgh: Carnegie Mellon University,**

School of Computer Science, Machine Learning Department 2006.

[33]. Witten IH, Frank E, Hall MA, "Practical Machine Learning Tools and Techniques," **Morgan Kaufmann**, p. 578, 2005.

[34]. Cruz JA, Wishart DS, "Applications of Machine Learning in Cancer Prediction and Prognosis," **Cancer Informatics, vol. 2**, pp. 59-78, 2006.

[35]. Niknejad A, Petrovic D, "Introduction to Computational Intelligence Techniques and Areas of Their Applications in Medicine," **Med Appl Artif Intell,** p. 51-63, 2013.

[36]. Tan PN, Steinbach M, Kumar V, "Introduction to Data Mining," **Pearson Education India 2016**.

[37]. Gokhale S, "Ultrasound Characterization of Breast Masses," **The Indian Journal of Radiology & Imaging, vol. 19, no. 3**, pp. 242-247, 2009.

[38]. Tang J, Rangayyan RM, Xu J, El Naqa I, Yang Y, "Computer-Aided Detection and Diagnosis of Breast Cancer with Mammography: Recent Advances," **IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 2**, pp. 236-251, 2009.

[39]. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ, "Cancer Statistics, 2008," **CA: A Cancer Journal for Clinicians, vol. 58, no. 2**, pp. 71-96, 2008.

[40]. Guyon I, Weston J, Barnhill S, Vapnik V, "Gene Selection for Cancer Classification using Support Vector Machines," **Machine Learning, vol. 46**, pp. 389-422, 2002.

[41]. Cicchetti DV, "Neural networks and diagnosis in the clinical laboratory: state of the art," **Clinical Chemistry, vol. 38, no. 1,** pp. 9-10, 1992.

[42]. Cochran AJ, "Prediction of outcome for patients with cutaneous melanoma," **Pigment Cell Research, vol. 10, no. 3,** pp. 162-167, 1997.

[43]. Exarchos KP, Goletsis Y, Fotiadis DI, "Multiparametric decision support system for the prediction of oral cancer reoccurrence," **IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 6,** pp. 1127-1134, 2011.

[44]. Kononenko Igor, "Machine learning for medical diagnosis: history, state of the art and perspective," **Artificial Intelligence in medicine, vol. 23, no. 1,** pp. 89-109, 2001.

[45]. Park K, Ali A, Kim D, An Y, Kim M, Shin H, "Robust predictive model for evaluating breast cancer survivability," **Engineering Applications of Artificial Intelligence, vol. 26, no. 9,** pp. 2194-2205, 2013.

[46]. Sun Y, Goodison S, Li J, Liu L, Farmerie W, "Improved breast cancer prognosis through the combination of clinical and genetic markers," **Bioinformatics, vol. 23, no. 1,** pp. 30-37, 2007.

[47]. Tian F, Conde J, Bao C, Chen Y, Curtin J, Cui D, "Gold Nanostars for Efficient in Vitro and in Vivo Real-Time SERS Detection and Drug Delivery via Plasmonic-Tunable Raman / FTIR Imaging," **Biomaterials, vol. 106**, pp. 87-97, 2016.

[48]. Su X, Wang Y, Wang W, Sun K, Chen L, "Phospholipid Encapsulated AuNR@ Ag/Au nanosphere SERS Tags with Environmental Stimulus Responsive Signal Property," **ACS Applied Materials & Interfaces, vol. 8, no. 16**, pp. 10201-10211, 2016.

[49]. Butler HJ, Ashton L, Bird B, Cinque G, Curtis K, Dorney J, Esmonde-White K, Fullwood NJ, Gardner B, Martin-Hirsch PL, Walsh MJ, "Using Raman Spectroscopy to Characterize Biological Materials," **Nature Protocols, vol. 11, no. 4**, pp. 664-687, 2016.

[50]. Haruna K, Saleh TA, Al Thagfi J, Al-Saadi AA, "Structural Properties, Vibrational Spectra and Surface-Enhanced Raman Scattering of 2, 4, 6-trichloro-and tribromoanilines: A Comparative Study," **Journal of Molecular Structure, vol. 1121**, pp. 7-15, 2016.

[51]. Haruna K, Saleh TA, Hossain MK, Al-Saadi AA, "Hydroxylamine Reduced Silver Colloid for naphthalene and phenanthrene Detection using Surface-Enhanced Raman Spectroscopy," **Chemical Engineering Journal, vol. 304**, pp. 141-148, 2016.

[52]. Amrane M, Oukid S, Gagaoua I, Ensarİ T, "Breast Cancer Classification using Machine Learning," **Proceedings of IEEE 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT),** 18-19 April, 2018, Istanbul, Turkey, pp. 1-4.

[53]. Bhuvaneswari P, Therese AB, "Detection of Cancer in Lung with k-NN Classification using Genetic Algorithm," **Procedia Materials Science, vol. 10**, pp. 433-440, 2015.

[54]. Yan W, Wang K, Xu H, Huo X, Jin Q, Cui D, "Machine Learning Approach to Enhance the Performance of MNP-Labeled Lateral Flow Immunoassay," **Nano-Micro Letters, vol. 11, no. 1**, pp. 1-15, 2019.

[55]. Freeman JA, Skapura DM, "Algorithms, Applications, and Programming Techniques. Neural Networks," **Addison Wesley Longman Publishing Company**, New York, USA, 1991.

[56]. Chen YS, Hsu YC, "Effective and Efficient Baseline Correction Algorithm for Raman Spectra," Lecture Notes in Engineeering and Computer Science: **Proceedings of International MultiConference of Engineers and Computer Scientists 2019, IMECS 2019,** 13-15 March, 2019, Hong Kong, pp. 295-298.

Daxiang Cui is currently working as Professor in the Department of Instrument Science & Engineering, School of Electronic Information and Electrical Engineering of Shanghai Jiao Tong University. He obtained his MD and PhD at the Fourth Military Medical University in 1998. He was a post-doctoral fellow in Max Planck Institute for Metals Research from 2001 to 2004, and was a visiting professor in Waseda University from 2007 to 2008. So far he is the Distinguished Professor, Changjiang Scholar at Shanghai Jiao Tong University. He has published over 300 papers in international peer-reviewed journals, with a high-index of 45. His research interests include controlled synthesis and biosafety evaluation of nanomaterials, nanoparticles-labeling and Nano-effects-based tumor theranostic technologies, high efficient drug delivery system and RNA Nano-drug.

Muhammad Aqeel Aslam became a member in 2020. He was born on 16th Jan, 1985 in Pakistan. He did his MS in Electrical Engineering from National University of Science and Technology, Islamabad, Pakistan in 2012. Currently, he is doing PhD from the Department of Instrument Science & Engineering, School of Electronic Information and Electrical Engineering of Shanghai Jiao Tong University. His current interests are biosensors based medical clinical applications and Artificial Intelligence for medical diagnosis.

Cuili Xue is a PhD student in the School of Electronic Information and Electrical Enginering, Shanghai Jiao Tong University.

Manhua Liu is working as Asscoiate Professpr in rhe School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong Unoversity.

Kan Wang is working as an Associate Professor in the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University.