https://doi.org/10.48048/tis.2023.5425

Biomarker Based Detection and Staging of Breast Cancer from Blood Using Raman Spectroscopy and Deep Learning Technique

Renjith Vijayakumar Selvarani* **and Paul Subha Hency Jose**

Department of Biomedical Engineering, Karunya Institute of Technology and Sciences,
Tamil Nadu 641114, India

(*Corresponding author's e-mail: notify_renjithvs@yahoo.com)

Received: 28 July 2022, Revised: 29 August 2022, Accepted: 5 September 2022, Published: 16 March 2023

Abstract

Breast cancer (BC) or breast neoplasm is causing major menace to the life of women around the world. The significance of early detection and staging of BC has been substantial in diagnosing protocol. This work aims to develop an automated system that combines multivariate data analysis (PCA - principal components analysis) with ensemble recurrent neural network models (stacked OGRU-LSTM) to identify Raman spectral characteristics that can be used as spectral cancer markers for the detection of BC progression and staging. Features of blood plasma from histopathologically diagnosed BC candidates were compared to healthy ones in this study. The same is performed on different leading classification models as the stacked basic RNN, the stacked-RNN-LSTM, and RNN-GRU models. A total of 2,340 Raman spectra generated is evaluated in this study. It is found from the study that stage 3 and stage 2 are structurally identical, but with PCA-Factorial Discriminant Analysis (FDA) they can be distinguished from each other, hence the Raman spectrum pertaining to blood plasma samples of the BC candidates is classified efficiently, yielding potentially high values of specificity and sensitivity for all the BC stages. Comparative classification results show that the stacked OGRU-LSTM model outperforms well for BC detection, and better differentiates various stages of BC by employing the multivariate data analysis technique. The stacked OGRU-LSTM model achieved the highest classification accuracy (97.89 %), Cohen-kappa score (0.928), F1-score (0.957), and the lowermost number of test loss and MSE (0.037), indicating that the model outperforms other baseline classifiers.

Keywords: ANN, Breast cancer detection, LSTM, OGRU, OGRU-LSTM, Staging of breast neoplasm, Stacked RNN

Introduction

Text Breast cancer (BC) is one of the most common cancers that affect a large percentage of females around the globe. BC starts in the cells that line the milk ducts and extents into lymph and tumor over time [1]. 50 % of BC patients are discovered in the later stages of the disease, such as stages III and IV, especially in India. When it comes to developed countries, the diagnosis rate reaches 12 % [2]. According to a study, around 7 % of women are diagnosed with BC before the age of 40, and for more than 40 % of all cancers in this age group that women face is BC [3].

In accordance with the National Cancer Institute, United States, BC treatment cost around $20.5 billion in 2020 [4]. However, the rate of survival is highly advanced in developed countries [5]. The role of hereditary factors in BC is estimated to be 5 - 10 % [6,7]. Late pregnancy, breastfeeding failure, nulliparity, post-menopausal weight gain, alcohol intake, and lack of exercise are all potential risk factors [8]. BC cells originate in mammary gland tissue and have a 7-year evolution time. 100 to 300 days are needed to double the BC cells and require 30 doublings for breast neoplasm [9].

Cancer is sometimes referred to as a fatal disease since it develops without causing any symptoms. This highpoint the need for early detection of BC, as well as advanced screening procedures that can detect early-stage anomalies at a micro-level. Breast cancer can be treated using a multimodal strategy that combines surgery, hormone therapy, chemotherapy, targeted therapy, and radiation therapy. Hence, detection in the early phase will aid clinicians in recommending possible and effective treatments.

The proposed method in this research work processes spectral data from the blood plasma of BC patients with Raman spectroscopy and determines whether or not the patient is affected by BC. The various stages of BC are a critical aspect of the diagnosis process, especially in terms of survival chances. The classification of BC is based on lymph node involvement (N), tumor size (T), and metastasis or no metastases (M) [5]. In stage 0, BC has not progressed beyond the initiation point, the tumor has moved to certain fatty tissues in the breast and expanded in size in stage 1, while in stage 2 it has migrated to 3 lymph nodes and it has extended to the chest wall in stage 3. It has progressed from the lymph nodes or breast to the bones, lungs, liver, and brains at stage 4, commonly known as the metastatic phase, after that there is no higher stage classification.

Considering the limits of current cancer detection methods, particularly early-stage cancer detection, there is an immediate necessity to develop novel methods for rapidly diagnosing cancer at a very early stage. The novelty of the proposed system is that it is able to detect BC at early stages in asymptomatic candidates with high accuracy in real-time with a minimal quantity of blood samples. This would potentially lead to the identification of BC at an early stage, thus allowing timely diagnosis and treatment to save lives. The Raman spectroscopy can identify label-free breast cancer and analyze the variation in chemical composition that exists among the different samples based on the morphology. It can identify variations in molecular structure and composition in the constituents of tissue and cell such as carbohydrates, proteins, lipids, and nucleic acids that occur during tumor growth. The metabolic products of tumors are carried by the blood and can be tracked during circulation. Thereby analyzing the structural differences and molecular composition of the biomarkers circulated within blood samples effectively by employing the deep learning classifiers at the molecular level. The ability of Raman spectroscopy to assist the diagnosis of early-stage breast cancer [11] is potent since these biochemical variations manifest [10] before the appearance of clinical signs frequently recognized by imaging protocols.

To help healthcare professionals, deep learning (DL) based data discovery algorithms that work on spectral patterns among large amounts of spectral data are being utilized. DL techniques, which are a subset of machine learning, are frequently beneficial due to their self-adaptive structure, which can analyze data with less computational processing. Instead of doing it manually by an expert clinician, the computer-aided DL process is employed, allowing clinicians to participate in the analysis of BC. Deep learning outperforms traditional artificial neural networks (ANN) as it allows for the creation of networks with more than 2 layers [17]. In this paper, a DL architecture-based system with multivariate data analysis is used to improve the accuracy of BC prediction and classification of stages using Raman spectral data. A Recurrent Neural Network (RNN) [6] is a DL model with a feedback response loop structure that is frequently used for prediction. By enhancing the learning process of the typical gating unit called the GRU model, this research work put forward an Optimized Gated Recurrent Unit (OGRU) model, OGRU increases the prediction accuracy and learning efficiency of the deep GRU model. To eliminate interference from inessential information, the output of the gate is utilized to filter the current input data. The reset gate modifies the output of the update gate, which can improve learning efficiency by speeding up convergence and suppressing the attenuation problem of the gradient. In this research work, an ensemble stacked OGRU-LSTM model is proposed which can detect the result of the diagnosis. Two RNN variants are employed for prognosis, the Long-short Term Memory (LSTM) and OGRU. Quantitative, qualitative, comparative, and complexity measurements are all used in the analysis of the algorithms proposed. The spectral characteristics obtained from blood plasma samples from patients diagnosed with stage 2, 3, or 4 BCs are analyzed using Raman spectroscopy and multivariate analyses along with the OGRU-LSTM classification model.

The paper is strongly motivated for early BC prediction and staging as;

1) To generate a balanced spectral dataset, pre-processing is performed;

2) For improved accuracy, OGRU and LSTM (Hybrid/ensemble deep learning models)-based classification models are used;

3) The multivariate technique is used to validate the distinctive spectral signatures of blood plasma samples with stages 2, 3, or 4 BC.

4) The model is compared to existing techniques to show that it outperforms the most recent models sin all aspects.

The rest part of this paper is structured as follows. Section 2 summarizes several existing methods. Section 3 particulars the sample preparation, acquisition of Raman spectral signal, preprocessing of the spectra, and finally data augmentation pertaining to this research under materials and methods. Section 4 describes the methodology and implementation, where it starts with the scope of using GRU as the base architecture of deep learning for time series data pertaining to this research work, afterward describing 4 deep learning architectures along with the proposed OGRU-LSTM followed by RNN-LSTM, GRU-RNN, and RNN. Section 5 is the results and discussion pertaining to this research, where the tested results of the classification methods have been discussed, and the results of each deep learning model are compared to conclude the optimum deep learning method pertaining to this research work. The conclusion is enumerated in Section 6.

Related works

To distinguish among benign and malignant BCs, researchers used neural networks and ultrasound imaging with multi-fractal dimension characteristics. The majority of the work was done on imaging. The study reported the most precise classification results, with an accuracy of 82.04 % [19]. In the Weka tool for the identification of BCs, a comparative analysis of clustering approaches, such as LVQ, hierarchical clustering, DBSCAN, and canopy, was undertaken in [20]. The first clustering technique was found to have the highest prediction accuracy of 72 % [20], according to the published results. For BC classification, CNN and Multiple Instance Learning (MIL) were coupled. The studies used the BreaKHis database, which included 8,000 malignant and benign BCs biopsy images. With a magnification factor of 40x, the classification rate was found to be 92.1 % [22].

RNN is a sort of neural network (NN) model that can process both parallel and sequential data. By adding memory cells into the neural layer, similar operations to those of the human brain can be replicated. Bidirectional Recurrent Neural Network (BiRNN) is another RNN model that is modeled to process input sequences with known start and end in advance. Subsequently, RNN uses data from the previous session, and BiRNN might be used to make even more improvements. Data from 2 different sources can be handled by Bi-RNN. The sequence is processed by RNN from start to finish, while the other processes it backward from beginning to end [31], taking into account both the start and end context of each sequence element. LSTM-RNN and GRU-RNN are 2 RNN alternatives that differ in their gating units. The LSTM is a type of RNN that incorporates prediction based on context, which is not taken into account by regular RNN. By training RNN, the vanishing gradients problem can be eliminated by LSTM.

When it comes to Raman spectroscopy and deep learning models for early-stage breast cancer identification and classification of stages from blood plasma, not many works are being conducted effectively to the best of our knowledge, Raman spectroscopy can capture the structural and chemical features of blood plasma from spectral characteristics such as the position of peaks, and wavelength [37]. Krishna *et al.* [37] and Bergholt *et al.* [21]. employed Raman spectroscopy to analyze the oral sample in an *in vivo* environment. The findings of the experiment indicate that the features of the spectral patterns are used to classify normal tissue specifically in high wavenumber. Singh *et al.* put forwarded the implication of Raman spectroscopy in oral cavity lesions detection [23]. Raman spectroscopy has significant benefits in the detection of oral cancer tissue. The tumor cells alter the morphology and structure of tissues, and the resultant Raman peaks correlating to nucleic acid and proteins are evident [24]. Raman spectroscopy has significantly high specificity for the recognition of tumor tissue. It is hence condensed the necessity for adjunct therapy [25].

For cancer analysis, Raman spectroscopy has been broadly experimented. Krishna *et al.* identified that Raman spectroscopy is sensible to the spectral sensitivity against morphological changes and molecular composition related to malignant tissue. Oral leukoplakia, Oral squamous cell carcinoma, and oral sub-mucous fibrosis are differentiated from healthy tissues by statistical multivariate methods [26]. Carvalho *et al.* found the application of a high-wavenumber pattern in Raman spectroscopy in the oral tumor and detected the oral tissue by multivariate analysis called PCA [27]. Krishna *et al.* experimented with oral tissue in an ex-vivo environment and used Raman spectroscopy with a multivariate approach called PCA to compare the changes in protein in healthy tissues and oral squamous cell carcinoma [37]. Venkatakrishna *et al.* employed an optic fiber medium as a probe tool for Raman spectroscopy and validated the Raman spectrum using the technique called PCA, employing residual squared sum and Mahalanobis distance as a standard to differentiate oral tumor and healthy tissue [28]. Malini *et al.* experimented with the application of Raman spectroscopy in the recognition of oral cancer, using the multivariate approach PCA to characterize normal, precancerous, malignant, and inflammatory tissues [29]. Sonis *et al.* employed Raman spectroscopy to examine the spectral patterns obtained from dissimilar oral regions [30]. Another multivariate analysis approach called linear discriminant analysis (LDA) is used to show the potential of Raman spectroscopy for specificity for specific mucosal types and oral mucosal diseases of the oral region [37]. Deep learning approaches need not need any manual involvement to obtain features compared to conventional machine learning approaches. Hence, deep learning is employed in healthcare monitoring systems, as electrocardiogram identification [32-34], segmentation of lung on x-ray [35], and characterization of bacteria [36]. This technical approach can excerpt more characteristic features of cancer. Consequently, the cancer tissue recognition accuracy is enhanced.

Many cancers related studies employed several machine learning models, conducted on different subclasses of their data. In particular, many research compared the effectiveness of various ML models, typically conventional ML models including LDA and PCA with deep learning models as CNNs. **Table 1** includes the model with comparatively the best performance. We provide this metric unless otherwise mentioned, despite the fact that its applicability to classification tasks is questioned, due to its ubiquitousness in the literature reviewed and its intuitive interpretation.

Table 1 Literature review overview.

Authors	**Sample**	**Model**	**Accuracy**
Li C *et al.* [48]	cell lines-Breast Cancer (BC)	PCA SVM	99.0 %
Shang L *et al.* [49]	Tissue-BC	CNN	92 %
Brusatori M *et al.* [50]	Tissue-BC	CNN	90 %
Wu S *et al.* [51]	Tissue-Oral Cancer	PCA-QDA	82 %
Zhu L. *et al.* [52]	Tissue-Tongue Cancer	CNN-SVM	99.5 %
Lin K. *et al.* [53]	Tissue-Nasopharyngeal Cancer	GA-LDA-PLS	98 %

Materials and methods

Preparation of sample and protocol

18 females diagnosed with BC histologically and 8 control females were chosen for the study. Particularly, the available BC samples belong to stages 2 - 4. Since candidates rarely report to the clinic at stage 1, stage 1 samples are not available. Candidates from similar socioeconomic and ethnic backgrounds were grouped. To get plasma samples, blood samples were taken in EDTA vials and centrifuged at the rate of 3,400 rpm for 5 min.

Acquisition of raman spectral data

20 µl of a plasma sample from whole blood was placed on a substrate (aluminum) and spectral characteristics of Raman were obtained. 20 µl of the same plasma sample are placed on the aluminum substrate 3 times, with a new aluminum substrate each time, yielding a total of 25 spectra from each sample of blood plasma. Raman spectra were collected from all 18 females with BC and 8 controlled blood samples using a Raman micro spectrometer. The spectrometer has a laser source of 785 nm diode with 60 MW power. The Raman spectra were obtained between 600 and 1,800 cm^{-1} in 30 s. **Table 2** lists the Raman spectral feature assignments [38-42].

Table 2 Raman spectral feature assignments.

Spectral wavenumbers	**Assignment names (Biomarkers)**
625 cm^{-1}	Nucleotide conformation
640 cm^{-1}	Stretching of C-S & twisting of C-C in $C_9H_{11}NO_3$
675 cm^{-1}	CO-NH I (b-sheet)
689 cm^{-1}	Nucleotide conformation
714 cm^{-1}	C-N adenine CN_2 $(CH3)_3$ (lipids)
752 cm^{-1}	C_5 H_5 N_5 O
761 cm^{-1}	C_{11} H_{12} N_2 O_2, d (ring)
770 cm^{-1}	PO_4^{3-}
778 cm^{-1}	C_{47} H_{83} O_{13} P
788 cm^{-1}	phosphodiester bands in DNA

Spectral wavenumbers	Assignment names (Biomarkers)
798 cm^{-1}	CH deformation
828 cm^{-1}	$C_9H_{11}NO_3$ /protein
857 cm^{-1}	$C_{57}H_{91}N_{19}O_{16}$
885 cm^{-1}	Disaccharide (cellobiose), (C-O-C) skeletal mode
901 cm^{-1}	Monosaccharides (b-glucose)
917 cm^{-1}	C_5 H_9 NO_2, C_5 H_9 NO_3
932 cm^{-1}	C-C, a-helix
1,100 cm^{-1}	C-C gauche-bonded chain
1,138 cm^{-1}	n (C-C)- fatty acids, lipids
1,145 cm^{-1}	n (C-C)- fatty acids, lipids
1,173 cm^{-1}	$C_4H_5N_3O$, $C_5H_5N_5O$
1,185 cm^{-1}	Anti-symmetric PO_4^{3-} vibrations
1,268 cm^{-1}	d (=C-H) (CH_2OH–CHOH–CH_2OH)
1,285 cm^{-1}	CH_2OH–CHOH–CH_2OH
1,307 cm^{-1}	Twisting of CH_3/CH_2, wagging &/or $C_{65}H_{102}N_{18}O_{21}$ bending mode & $CH_3(CH_2)$ nCOOH
1,319 cm^{-1}	$C_5H_5N_5O$
1,410 cm^{-1}	ns COO_2 (IgG)
1,440 cm^{-1}	deformation of CH_2 in normal breast tissue
1,609 cm^{-1}	$C_4H_5N_3O$
1,660 cm^{-1}	$C_{15}H_{31}N_3O_{13}P_2$

Preprocessing of data

All Raman spectral data processing was done with Python [43]. Smoothing, vector normalization, substrate removal, and baseline correction were all part of the data pre-processing. The Savitzky-Golay smoothing method was used to vector normalize and smooth all spectra, including substrate backgrounds (order 5,13-point window). All of the spectra were rubber band corrected, and the spectra from the substrate were deducted from every spectrum to remove the baseline.

Smoothing of the spectral data is a typical pre-processing technique. Smoothing is a mathematical process done on raw spectral data to eliminate (random) noise. This is particularly necessary when attempting to isolate significant spectral characteristic features that may be partially occluded by noise. For Raman spectroscopy, derivatives of the real spectral data are evaluated. The numerical distinction of data intensifies noise highly, and smoothing is required to obtain meaningful and significant spectral derivatives. The Savitzky-Golay (SG) method is a standard in data smoothing, especially in the case of time series data. Rubber band corrections are performed with the ends attached to the ends of the spectrum or at least 1 component of the spectrum to be corrected, and wrapped around the curve profile of the spectrum from below. The rubber band is then positioned against the spectrum's curve characteristic pattern. These are used in all deep learning approaches and multivariate analyses for time series data. When this is subtracted from the spectrum, the required spectrum with a rectified baseline is attained, and the preprocessing results (Where the spectral sample from each class of blood samples such as cancer and controlled sample; are preprocessed and the resultant preprocessing results are illustrated in **Figure 1**), are shown in **Figure 1**.

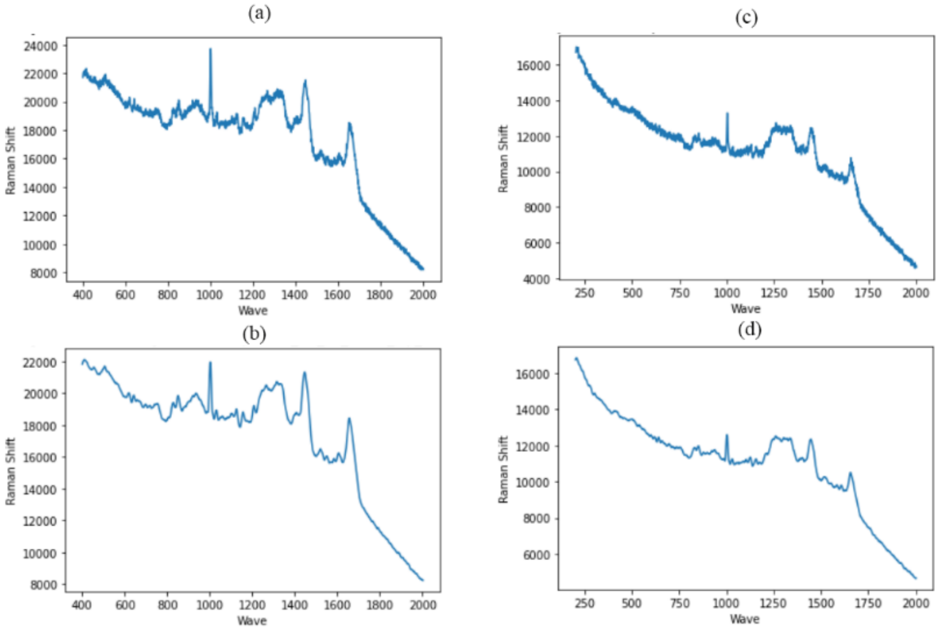


Figure 1 (a) the raw spectrum obtained from a BC sample and (b) the preprocessed spectrum of the raw BC sample, (c) the raw spectrum obtained form-controlled sample (d) the preprocessed spectrum of the raw controlled sample.

Augmentation of data

The deep learning models need sufficient data to process. Adequate training of data allows the neural network to properly train the intrinsic features of the spectral data, improve the steadiness of models, and minimize overfitting. However, the number of observations (spectral sample) made for the study was limited. To address this problem, a successful and simple spectral data augmentation technique was created to synthesize new spectral data with no incurring additional labeling requirements. Data augmentation is a process that entails adding repeated spectra to the data by replicating along with noise and/or other changes to enhance the no. spectra. The data augmentation technique is employed in DL to not only enhance the spectral data size but also to add noise into the spectral data so that the tendency of the model to likely get overfit against the training set is less significant. As a result, the data is regularized, and the learning process smoothed out. This technique doesn't need to confront biological proceedings, nonetheless is convenient as it can be used as a technique for data regularization, hence it reduces the tendency of a model to over-fit the data. The following are the 3 methods of data augmentation: (a) Moving the original Raman spectra left or right by 2 cm^{-1}, increasing the overall spectral resolution from 600 to 3,000; (b) Increasing the overall spectral resolution from 3,000 to 3,600 by applying random Gaussian noise to the source spectra; (c) Using a randomized scaling coefficient with a sum of 1 to linearly superimpose the original Raman spectra of the same kind, the total spectral resolution was increased from 3,600 to 5,000 [54]. Each sample in the training set underwent this process 10 times, and the results were added to the dataset.

Methodology

The aim of the proposed classifier is to apply DL techniques to effectively determine whether a patient has BC or not from the Raman spectrogram, and then use the same spectral data to identify the stages of the BC based on multivariate data analysis. Deep learning approaches are employed in the automatic recognition of features from raw spectral data with an end-to-end training procedure and supervised learning (SL) archetype. The SL method is used to classify blood samples and determine whether they are benign or malignant. For such predictions, the proposed method uses an ensemble OGRU and LSTM framework. The potential of OGRU-LSTM-based architecture is put forward in this research work, which has a different pattern of LSTM and OGRU layers with 4 dense layers (D-layer) incorporated with PCA-FDA multivariate analysis. Bidirectional neural network layers are used to implement the LSTM and OGRU layers. Except for the D-layer, the dropout layers follow the OGRU and LSTM Layers. The results obtained from the leading deep learning classifiers such as RNN-LSTM, GRU-RNN, and RNN are compared with the proposed ORGU-LSTM model. The validation of the models is performed by validating the throughputs obtained from the classifiers and multivariate data analysis with the ground truth value where it is histopathologically labeled for each sample. The labels of samples are as follows cancerous, non-cancerous, stage 2 BC, stage 3 BC, and stage 4 BC. The results of each classifier are cross-validated with the label marked against each sample to ensure classification performance.

The overview of typical GRU

The vanishing gradient concern with RNN is resolved by the use of GRU. GRU employs the "reset gate and update gate" tactic, which includes 2 vectors to determine what data must be allowed to pass to the output. The most unique feature of these vectors is that they are able to train to hold the previous data without eliminating or washing it, even though it is typically not relevant to prediction. GRU performs better than RNN and LSTM because of its easy structure. GRU in a modest way regulates the stream of data through gates similar to LSTM.

The GRU-NN is a type of RNN that can sustain longer-term data dependence and is commonly employed in clinical examination protocol. The output gate, input gate, and forgetting gate are the 3 gate units that make up the LSTM neural network model. The time-series data is processed by modeling the gate. Although the gradient fades to some amount, training is more likely as a result of the parameters actively engaged.

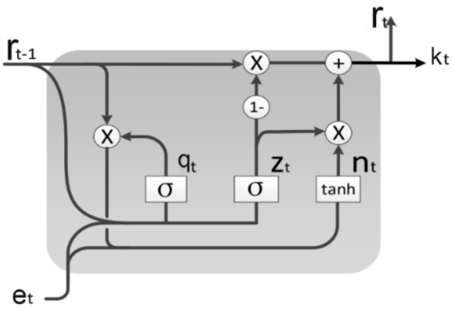


Figure 2 The general architecture of GRU [33].

The GRU-NN model is a variation of LSTM-NN. It improves the topology of LSTM by combining the LSTM's 3 gating units into 2, namely the reset gate and the update gate. As a result, the GRU model's parameters are modest, the training overhead is low, and a longer distance information dependency may be maintained. **Figure 2** illustrates the neural structure of a conventional GRU.

The implicit, output, and input layers constitute the GRU model. GRU neurons are used to create the hidden layers. The data stays at time *t* after data preprocessing is the GRU neural network's input data. In most cases, time-series data is used as input. The conventional GRU unit, reset gate, and update gate calculation formulas are as follows at time *t* if the input sequence is $e_1, e_2,..., e_t$.;

$$q_t = \sigma\,(W_r * [r_{t-1}, e_t]) \tag{1}$$

$$l_t = \sigma\,(W_l * [r_{t-1}, e_t]) \tag{2}$$

$$m_t = tanh\,(W * [q_{t-1} * r_{t-1}, e_t]) \tag{3}$$

$$r_t = (1 - l_t) * r_{t-1} + l_t * m_t]) \tag{4}$$

$$k_t = \sigma\,(W_o * r_t) \tag{5}$$

The output of the reset gate at time *t* is denoted by r_t; The output gate at time *t* is represented by l_t; In the update gate, W_l is the weight between the r_{t-1} and input; W_q is the weight in the reset gate between the r_{t-1} and input, where at time *t-1* the standard GRU unit output is r_{t-1}. The input at time t is denoted by e_t; at time *t*, m_t is a new vector formed using the tanh layer, i.e. in the hidden state, and added to the current state;

W denotes the weight between the inputs and the update gate's output l_t; The output of the standard GRU unit at time t is represented by r_t, it is utilized to update the current state of the neuron; at the present *t* time, multiply (1-l_t) the preceding hidden state r_{t-1} to create the hidden state h_t, to build the hidden state r_t at the present *t* time, remove extraneous data and add the product of l_t and m_t.; W_o signify the weight of r_t, and at time *t*, k_t is the output of the GRU-NN, that is, the predicted result; for the sigmoid activation function, 2 often utilized neuron activation functions are sigmoid and tanh. The GRU model preserves sensitive data across long distances by reducing the amount of gating units, eliminating redundant data continuously, and storing information dependencies in the hidden state, as shown in the formula above [33].

The proposed OGRU- LSTM model

This section details the hybrid model for breast cancer identification and classification and it is seen that the same model can improve the accuracy of classification. Considering all the potential advantages of hybrid models and focusing on enhancing the performance of classification approaches, this paper compares and evaluates the proposed hybrid model with the other 3 leading deep learning models. These methodologies are applied to classify the Raman spectrum as either benign or malignant.

GRU, on the other hand, has drawbacks such as low learning efficiency and slow convergence. To eliminate these drawbacks a neural network model with an OGRU is proposed. The OGRU architecture utilizes the reset gate to advance the prediction performance and learning efficiency of GRU by optimizing its learning process. As a result, an OGRU is proposed, in which the GRU neural unit's update gate is enhanced, the actual input of update gate e_t is modified to e_t multiplied with r_t, and the output of the reset gate is utilized to adjust the update gate with feedback. By filtering the present input data e_t through the reset gate, similar data's negative effects are reduced to a greater extent, and convergence will be accelerated, hence the goal of efficient learning will be achieved. The deep OGRU neural network is modeled from GRU architecture, whose neural structure is depicted in **Figure 3**.

$e_1, e_2..., e_t$ are considered as the input sequence, then the gate is updated at *t*, reset the gate, and apply OGRU output calculation formula;

$$q_t = \sigma\,(W_q * [r_{t-1}, e_t]) \tag{6}$$

$$l_t = \sigma\,(W_l * [r_{t-1}, e_t * r_t]) \tag{7}$$

$$m_t = tanh\,(W * [q_t, r_{t-1} * e_t]) \tag{8}$$

$$r_t = (1 - l_t) * r_{t-1} + l_t * m_t \tag{9}$$

$$k_t = \sigma\,(W_o * r_t) \tag{10}$$

Among them, In the formula, r_t and l_t have the same denotation as normal GRU neurons. The OGRU neuron varies from the GRU neuron in that, in order to hide the state weight, the r_t is multiplied with the preceding time at the update gate l_t, allowing the reset gate to re-screen the present input data e_t, and to optimize the neuron structure, the reset gate's output is used to modify the update gate.

The neuron topology of the OGRU is more logical than GRU, as shown in **Figure 3** and formula (7), and at each time the hidden state can be made simpler, and to some extent, the gradient attenuation can be controlled. As a result, the OGRU model may retain a higher distance information dependency, as well as a higher prediction accuracy and learning efficiency [33].

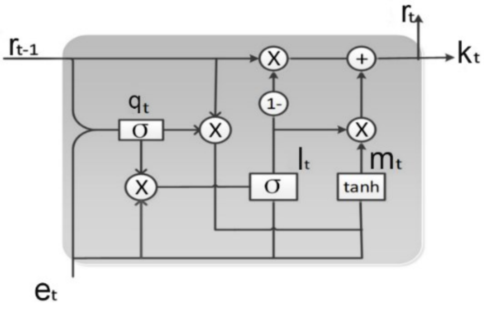


Figure 3 The OGRU neuronal structure [33].

The input layer, output layer, and hidden layer make up the deep OGRU-NN. The OGRU neurons constitute the hidden layers. The recursive transfer of data among neurons is increased, and the data preservation capability is enhanced, by optimizing the learning process of the GRU model. It is significant that the ratio of the mean of error and the sum of squares of error are employed to evaluate the accuracy of prediction. They are commonly used indicators for determining the variation among actual and expected values, and they can correctly reflect the model's accuracy of prediction [33]. Formulas (11) for AE and (12) for SSE illustrate the error ratio means and the sum of error squares;

$$AE = \frac{1}{M}\sum_{i=t}^{M} | \frac{k_i' - k_i}{k_i} | \tag{11}$$

$$SSE = \sum_{i=1}^{M}(k_i' - k_i)^2 \tag{12}$$

M denotes total no. samples in the validation, k_i' is the output predicted, and k_i is the real output. The accuracy of prediction is higher if AE is smaller than SSE.

To obtain maximum efficacy, hyperparameters must be tuned while designing this model. This section describes the model's specifications as well as its hyper-parameters. By considering the ability of LSTM with its less memory consumption, and fast and accurate performance in using datasets with longer sequences as pertaining to this work's spectral data, the OGRU model is ensembled with LSTM. The model proposed has 4 OGRU and 2 LSTM layers, each with 512, 256, 128, 64, 32 and 16 units. Each of these layers are having a 20 % dropout rate. Following that, 4 D-layer with 8, 4, 2 and 1 nodes are stacked in this model, correspondingly. With the exception of the dropout layers, in all levels, the sigmoid activation function is used. These layers are finally compiled using a 64-batch size Adam optimizer with 100 epochs. The hyper-parameters can be adjusted to help the model achieve the best prediction results. To obtain prediction, the NN takes a total parameter of 3, 32, 036 and trains them. **Table 3** has a comprehensive description of the model.

Table 3 Proposed stacked ensemble OGRU-LSTM Model.

Name of layers	Dropout rate/ No. of nodes	Shape of output	No. of parameters received
Bi-directional-OGRU	512	Nil-30-124	99,850
Dropout-layer	0.2	Nil-30-512	0
Bi-directional-OGRU	256	Nil-30-256	164,362
Dropout-layer	0.2	Nil-30-128	0
Bi-directional-OGRU	128	Nil-30-64	30,922
Dropout-layer	0.2	Nil-30-64	0
Bi-directional-OGRU	64	Nil-32	20,396
Dropout-layer	0.2	Nil-32	0
LSTM layer	32	Nil-32	10,474
Dropout-layer	0.2	Nil-32	0
LSTM layer	16	Nil-32	5,688
Dropout-layer	0.2	Nil-32	
D-layer	8	Nil-8	274
D-layer	4	Nil-4	46
D-layer	2	Nil-2	20
D-layer	1	Nil-1	4

Stacked RNN-LSTM model

Each of the 4 LSTM layers is followed by dropout layers in a stacked LSTM model. Four D-layers are included in the model again. The activation function 'sigmoid' is used to implement the final D-layer and the LSTM layers. This model trains with received 131,705 parameters for obtaining results prediction once it is implemented by selecting relevant hyper-parameters. **Table 4** describes this model.

Table 4 Stacked RNN-LSTM model.

Name of layers	Dropout rate/ No. of nodes	Shape of output	No. of parameters received
LSTM layer	128	Nil-30-128	66,560
Dropout-layer	0.2	Nil-30-128	0
LSTM layer	64	Nil-30-64	49,408
Dropout-layer	0.2	Nil-30-64	0
LSTM layer	32	Nil-30-32	12,416
Dropout-layer	0.2	Nil-30-32	0
LSTM layer	16	Nil-16	3,136
Dropout-layer	0.2	Nil-16	0
D-layer	8	Nil-8	136
D-layer	4	Nil-4	36
D-layer	2	Nil-2	10
D-layer	1	Nil,1	3

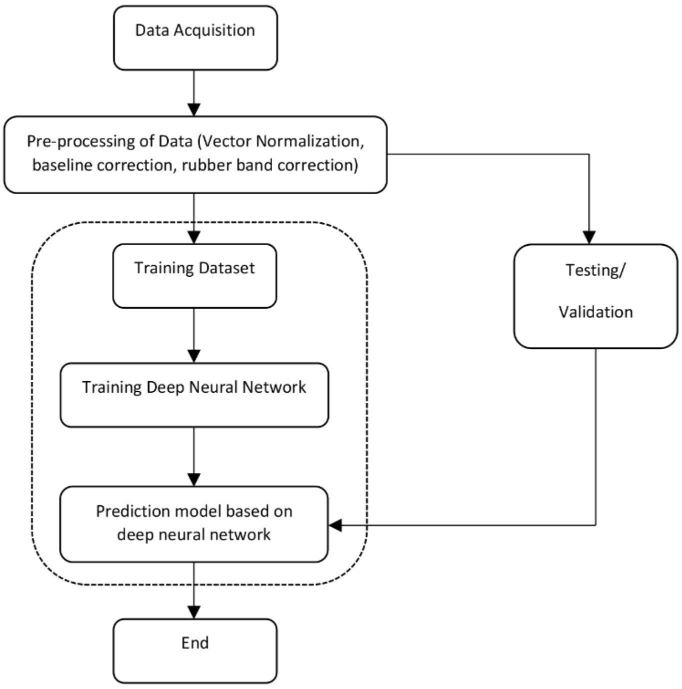

Data Acquisition
Pre-processing of Data (Vector Normalization, baseline correction, rubber band correction)
Training Dataset
Testing/ Validation
Training Deep Neural Network
Prediction model based on deep neural network
End

Figure 4 The flow diagram of the prediction model.

Stacked RNN-GRU model

This model, like the other 2 baseline classifiers, is made up of GRU sequences and dropout layers. This model similarly has 4 D-layers layered on top of each other. Like the previous 2 models, the final output layer and the GRU layers both take the activation function sigmoid. During the training phase, the model can make predictions by considering 98,825 parameters. **Table 5** gives a brief detail of the model.

Table 5 Stacked RNN-GRU model’s description.

Name of layers	Dropout rate/ No. of nodes	Shape of output	No. of parameters received
GRU-layer	128	Nil-30-128	49,920
Dropout-layer	0.2	Nil-30-128	0
GRU-layer	64	Nil-30-64	37,056
Dropout-Layer	0.2	Nil-30-64	0
GRU-layer	32	Nil-30-32	9,312
Dropout-Layer	0.2	Nil-30-32	0
GRU-layer	16	Nil-16	2,352
Dropout-Layer	0.2	Nil-16	136
D-layer	8	Nil-8	36
D-layer	4	Nil-4	10
D-layer	2	Nil-2	3
D-layer	1	Nil-1	136

Simple recurrent neural network

The general structure of RNN architecture is presented in this section. All 3 models have the same amount of layers and units with the same batch size. The significant part of the RNN Layer is different for each model. In each layer on the nodes, the current epochs, dropout layer rate, optimizer, and batch size of the developed models have the same structure. In terms of prediction, this will give a similar stage for comparing models. **Table 6** lists the stated parameters that are common to all of the executed models.

Table 6 Structure of the RNN model.

Layer no.	1	2	3	4	5	6	7	8
No. of nodes	128	64	32	16	8	4	2	1
Layer type	RNN				Dense			
Dropout layer	Yes (Dropout rate - 20 %)				No			
Size of batch	64							
No. epochs	100							
Loss	Log loss binary classification							
Optimizer	Adam							

This RNN model is made up of 4 RNN Layers stacked on top of each other. After each RNN Layer, dropout layers are utilized at a rate of 20 %. Nevertheless, these are followed by the addition of 4 dense-layers. ‘Sigmoid’ is the activation function for the last dense-layer and the first 4 RNN Layers. This model is fed a total of 33,065 parameters during the training phase. **Table 7** provides a concise description of this model.

Table 7 Stacked RNN model.

Name of layers	Dropout rate/ No. of nodes	Shape of output	No. of parameters received
RNN layer	128	Nil-30-128	16,640
Dropout-layer	0.2	Nil-30-128	0
RNN layer	64	Nil-30-64	12,352
Dropout-layer	0.2	Nil-30-64	0
RNN layer	32	Nil-30-32	3,104
Dropout-layer	0.2	Nil-30-32	0
RNN layer	16	Nil-16	784
Dropout-layer	0.2	Nil-16	0
D-layer	8	Nil-8	136
D-layer	4	Nil-4	36
D-layer	2	Nil-2	10
D-layer	1	Nil-1	3

An unsupervised method called PCA is used for exploring data sets and determining the underlying reasons for data variability. Correlated variables are transformed into uncorrelated variables by PCA called PC-scores (or scores), which can aid in reducing data dimensionality while retaining variability. The most variability in spectral data is explained by the first PC-score, and each subsequent score explains the next uppermost source of remaining variability. However, the fact that PCA is well adapted to highlighting the existence of clusters in a data set, spectral property variations not associated with certain pathological conditions but other less appropriate biological parameters can affect the transfer of data over scatter plots. As a result, when increasing the weight of underlying features of spectra in the classification, it seems that introducing a supervised constraint to the PCA for discrimination is essential. PCA-FDA (Factorial Discriminate Analysis) allows to compute and choose the appropriate PCs for your current application [44]. PCA-FDA, for instance, is employed to distinguish Raman spectra based on the BC stage, demonstrating that Raman spectroscopy is capable of diagnostic purposes. Cross-validation is necessary to evade overoptimistic classification rates and model overfitting due to the supervised nature of the model. As a result, the spectral data sets were divided into 2 3/4 of the spectra for training and 1/4 of the remaining spectra for validation using a random sample selection method. In addition, cross-validation of 100-fold was used to test the model's robustness with various combinations of training and validation data sets. In the confusion matrix, the aggregate result of 10 PC-scores is represented, which shows classification by means of sensitivity and specificity of categorization. This facilitates a more accurate assessment of the rate of differentiation acquired from a data set limited by the number of candidates.

Results and discussion

RNN models are being trained by feeding the dataset during the process of training. For testing and training, the resultant spectral dataset is employed. From the total of 26 blood samples around 2,340 Raman spectrums were obtained. When augmentation is performed on the obtained spectrums and thereby forms around 70,200 spectrums with characteristic features. With random sample selection, the data sets were split into a training set, which included 3-quarters (3/4) of the spectra, and a validation set, which included the final quarter (1/4) of the spectra. Over each epoch, the process of training is assessed against accuracy and loss. Baseline classifiers are trained using epoch sizes of varying lengths. The initial epoch sizes were 20, 50 and 100. These epoch sizes were used to train all the models.

In terms of testing accuracy, the models are analyzed and compared after they had been trained. The highest predictive efficiency is listed in **Table 8** with an epoch size of 100. Henceforth, the optimal training criterion is an epoch size of 100. Loss and Accuracy attained through each of the 100 epochs against each of the 4 models are specified in **Figures 5 - 8**. All of the models are implemented using the Keras [45] deep learning framework with the TensorFlow [46] backend.

Table 8 Accuracy of prediction for all different models of RNN for various Size of epoch.

Size of epoch	Stacked RNN (%)	Stacked RNN-LSTM (%)	Stacked GRU-RNN (%)	Proposed OGRU-LSTM (%)
20	64.83	69.43	76.53	81.27
50	67.69	76.83	86.73	91.27
150	77.38	86.34	94.26	97.89

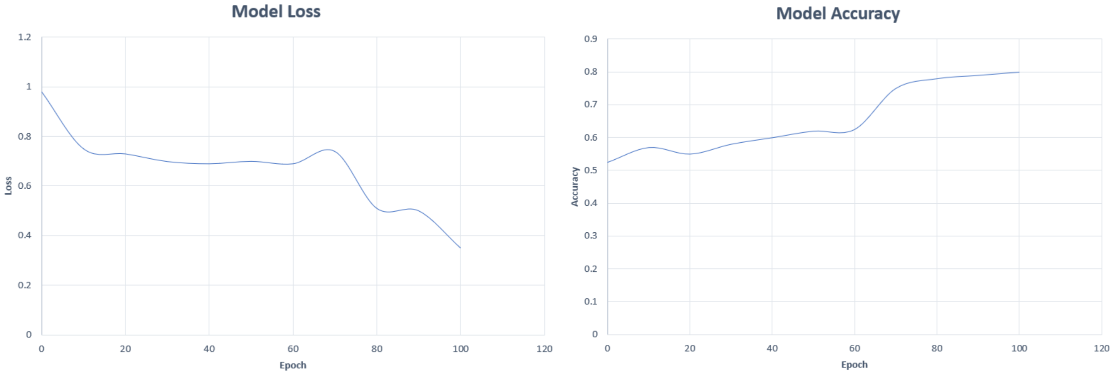
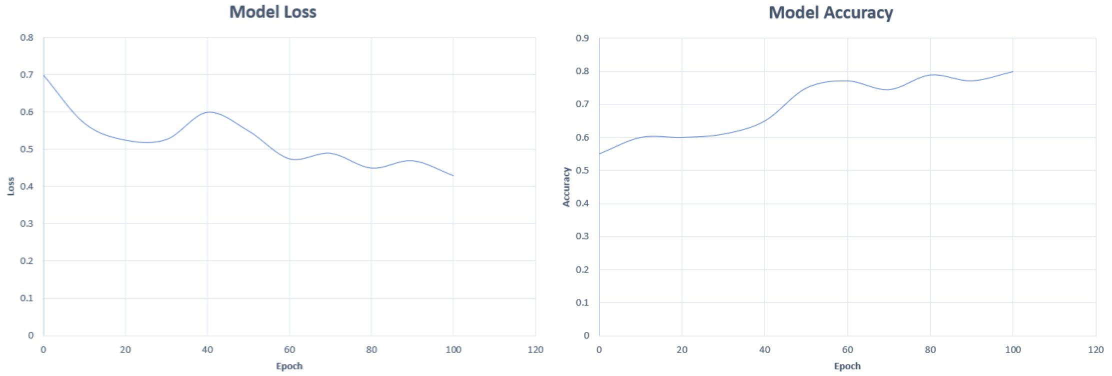


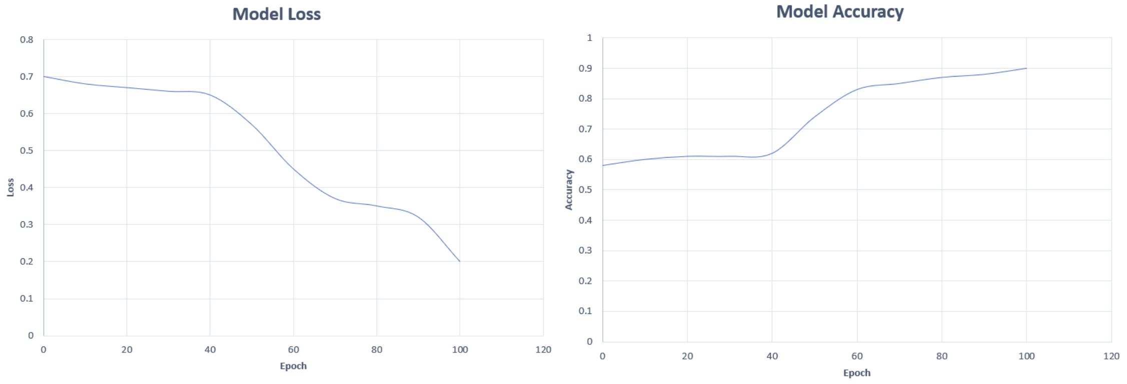
Figure 5 (a) Loss curve of stacked RNN and (b) accuracy curve of stacked RNN.

The F1-Score, accuracy, MSE, and Cohen-Kappa Score of the put forward stacked OGRU-LSTM model are all assessed. Losses that occur during testing are also recorded. It is then compared against the models of simple-RNN, and stacked GRU and stacked LSTM models as baseline classifiers. After the training procedure is completed, the accuracy of the test is calculated after the 100th epoch. Using evaluation metrics, all the executed models are illustrated after the completion of 100 training epochs in terms of performance. Predictions for the test dataset are obtained after this training session utilizing training data. Such prediction results are compared to actual observable values, resulting in a deep model evaluation based on the metrics utilized. **Table 9** shows the results of the comparison analysis. The comparison analysis shows that the proposed model outperforms existing classifiers in terms of promising results.



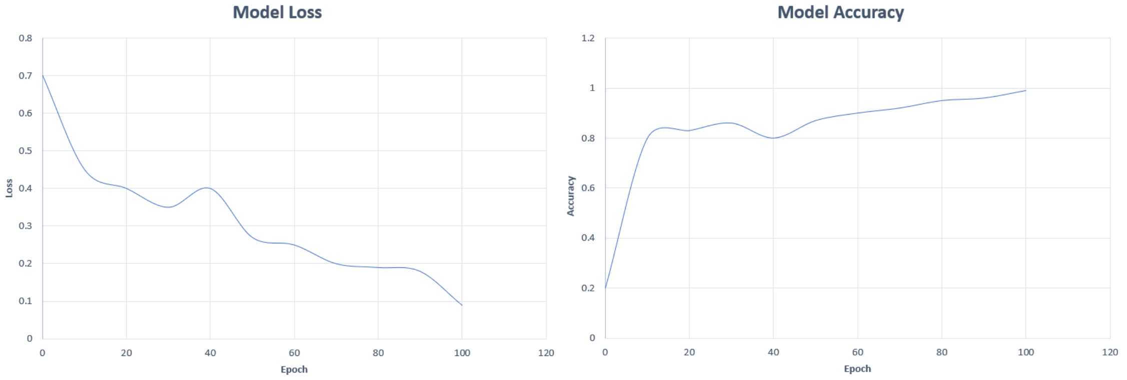
(a) (b)

Figure 6 (a) Loss plot of stacked LSTM model and (b) accuracy plot of stacked LSTM model.



(a) (b)

Figure 7 (a) Loss plot of stacked OGRU model and (b) Accuracy plot of stacked OGRU model.



(a) (b)

Figure 8 (a) Loss plot of stacked OGRU-LSTM model and (b) accuracy plot of stacked OGRU-LSTM model.

As demonstrated in **Table 9**, the stacked LSTM deep model outperforms the stacked simple RNN deep model in terms of classification accuracy. This is because, due to the structure of long-term dependencies, the LSTM outperforms the simple RNN. Again, it is known that over LSTM, GRU has better performance. As a result, the stacked GRU deep model shows an improvement in classification results. Even better classification results are obtained by combining the capabilities of LSTM and GRU into a single medium. As a result, this architecture is recognized as the most effective prediction model for BC identification with high efficiency.

Table 9 Performance of breast cancer classification on different models.

Parameters of evaluation	**Stacked RNN**	**Stacked RNN- LSTM**	**Stacked RNN- GRU**	**Stacked OGRU-LSTM**
Test loss	0.499775600514173	0.427584776595164	0.28454673589041267	0.09337276361547075
Accuracy (%)	77.38	86.34	94.26	97.89
F1-score	0.743	0.768	0.928	0.957
MSC	0.293	0.219	0.079	0.037
Cohen-kappa score	0.497	0.612	0.768	0.928

Figure 9 illustrates the mean spectra of healthy and BC plasma samples, together with the standard deviation (mean spectra of BC stages 4, 3 and 2), with possible distinguishing features specified by vertical lines. Raman spectral features in the mean Raman spectra of patient samples, at 689, 770, 788, 828, 848, 885, 1,138, 1,173, 1,185, 1,268, 1,285, 1,307 and 1,319 cm^{-1} have intensities significantly higher. In the mean Raman spectra of healthy candidates, however, the features of Raman spectra at 700, 761, 798 and 1,410 cm^{-1} had higher intensities than in the patient samples.

APCA scatter plot from the blood plasma of BC candidates at different stages vs healthy controls is shown in **Figure 10**. PC1 explicates 79 % of the distinction and provides reasonable discrimination among healthy and BC candidates' Raman spectral data. However, the Stage 2 and 3 samples are well distinguished from healthy ones, they are substantially overlapping, despite according to PC1 the Stage 3 samples being significantly distinct. The crossing spectral data of stage 3 and 2 clusters specifies that, though clinically classified by the intensity of cancer progression, they are both biochemically and spectrally equivalent in terms of bloodstream manifestation. The clusters of Raman spectra, in particular, have been related to stage 4 samples, which are grouped according to specific candidates. According to PC1, they are not considerably different from the previous stages, but according to PC2, they are significantly different, signifying a unique biochemical indication in the bloodstream. supporting information. The difference in all of these graphs is not potent when compared to **Figure 10** of the manuscript (PC2 vs PC1), where PC1 distinguishes healthy from cancer samples and PC2 distinguishes stage 4 from stages 2 and 3.

Figure 11 illustrates the loadings of Raman spectral data and PCA scatter plots of healthy against each stage of BC samples. In a pairwise comparison, the development of disease at these stages is well-differentiated, indicating the potential of PCA. The pair-wise PCA analysis better highpoint variability in spectral data and thus differ from the combined PCA data analysis of the full data set, as shown in the analysis reported in this manuscript.

In the PCA scatter plots (a), the PC loadings in **Figure 11** show changes that occurred in biological molecules that are relevant for the grouping of healthy and malignant samples. The malignant samples' Raman spectral properties are mentioned in Positive loadings, whereas healthy samples' Raman spectral features are mentioned in negative loadings, as grouped in the PC-1's negative and positive axes. Raman spectral signatures at 689, 770, 788, 828, 848, 1,138, 1,173, 1,185, 1,268, 1,285, 1,307 and 1,319 cm^{-1} have been found in positive loadings for all stages. These Raman features were significantly raised in the mean Raman spectra of BC, as shown in **Figure 9**, and so these positive loadings are connected to the disease. Raman features intensities at 700, 761, 798, and 1,410 cm^{-1} in the negative loadings were significantly raised in healthy volunteers' mean Raman spectra than in cancer samples. Other notable Raman spectral characteristics identified include positive loadings at 1,100 cm^{-1} and negative loadings at 1,440 and 1,663 cm^{-1}. Furthermore, Raman spectral characteristics at 1,663, 1,440, 1,319, 1,268, 1,185, 689, 1,173 and 1,138 cm^{-1} are raised in BC stage 4 and can thus be considered a marker connected with cancer development.

The PCA scatter plot in **Figure 12(a)** compares 2 different BC stages, stage 3 and stage 2. According to PC2, the Raman spectra of the various stages are being distinguished, showing that Raman spectroscopy can classify blood plasma of 2 stages of the disease. Positive PC2 loadings are connected to stage-3 Raman features, which are having intensities higher than stage-2 and could be observed at 625, 689, 770, 788, 828, 1,285, 1,307 and 1,319 cm^{-1}. Although the 2 stages of cancer development do not differ significantly (**Figure 10**), they do differ significantly in a pairwise comparison (**Figure 12(a)**).

Although stage 2 BC is still in its early stages, there is evidence that it is growing or spreading. Stage 3 BC, on the other hand, is a progressive stage of the disease in which cancer has spread to the breast's nearby tissues. In this context, Raman spectroscopy has the capability to aid in the separation of these phases of BC, potentially resulting in an earlier diagnosis and more successful therapy. The approach is shown to achieve this by identifying the underlying biochemistry, as well as the Raman spectrum properties, as shown in (**Figure 12(b)**). Positive PC2 loadings are connected to Raman features of stage-3 that are more intense than those of stage-2, specifically those of nucleic acids. Since migrating DNA and constituents of the cancer cell is seen in the blood, these characteristics connected with DNA are very unambiguous signs of malignancy. Furthermore, the presence of these features in stage-3 candidates implies a progression of cancer growth from stage 2 to 3, as demonstrated by the existence of these spectral features in the Raman spectra of stage 3 BC.

Figure 13(a) illustrates a PCA scatter-plot of BC stage-4 and 2, indicating how the stage-4 (magenta dots) and stage-2 (blue dots) Raman spectral data is distinguished by grouping in the negative and positive axes of the PC-1 correspondingly. In the Raman spectral data of stage-4 candidates, the differences seen as positive loadings have intensities significantly higher than those of stage-3 candidates, at 689, 1,185, 1,285, and 1,319 cm^{-1}, as shown in **Figure 13(b)**.

Figure 14(a) shows a PCA scatter-plot of stages 4 and 3 BC, with clustering of PC-1 on the positive and negative axes, respectively, indicating a separation between these 2 stages. Positive loadings indicate Raman spectral characteristics that have intensities significantly higher in stage-4 over stage-3 in **Figure 14(b)**, which indicates the PC1 loadings. These Raman spectral patterns, at 689, 1,185, 1,285 and 1,319 cm^{-1}, are also observed as metabolic features connected with stage-4 (**Figure 13**), and these spectral features had intensities higher in stage-3 than in stage-2 (**Figure 12**). As a result, the biochemical alterations detected in the BC stage-by-stage comparison reflect the biochemical changes that occur during the progression and evolution of BC.

Stage 4 is the most advanced stage of BC, in which cancer has spread to the breast's surrounding tissues. At this stage of cancer, effective treatment is quite difficult. The distinction between stages 2/3 and 4 is critical since it leads to an early diagnosis, which can help with appropriate therapy. Stage 2 of BC is particularly hard to diagnose by histology without a large number of biopsies. Furthermore, this technique consumes a lot of time for processing, and histology techniques are not always accurate. In this context, Raman-based spectroscopic analysis is a more rapid analysis technique for the detection of the early-stage BC and classification based on BC stages, mostly on the basis of structural variation of contents in DNA during the evolution of the BC from stages 2 to 4. The Raman spectra those at 689, 1,185, 1,285 and 1,319 cm^{-1}, are considered (**Figure 13**) as the spectral features connected to stage-4 breast neoplasm and in **Figure 12** these spectral features significantly have intensities higher in stage 4 as compared with stage 2 and 3. As a result, the biochemical variations identified in the stage-by-stage comparison for BC demonstrate that biochemical changes occur during BC progression and development. Notably, all candidates in stage 4 are found to be very different from others, based on the clustering pattern of the Raman spectral data in the scatter plots of PCA (**Figures 13(a)** and **14(a)**), which might be due to the increased severity of the BC and metastasis. The variation in these spectral features is due to the peak intensity variations at 640, 675, 689, 700, 752, 770, 788, 828, 901 and 932 cm^{-1}.

The PCA-FDA employed cross-validation of 100-fold and used ¾ of the spectral data for training the neural network and 1/4 for validation, resulting in 175×100 iterations = 17,500, which was then sub-divided by 4 to get a validation set = 4,375. To evaluate the diagnostic capability of the approach, a further classification of the healthy and BC using PCA-FDA was performed. **Table 11** illustrates the output, which specifies the progression of the BC which is detected by using blood plasma samples having a specificity of 100 % and a sensitivity of 99 %. In addition, PCA-FDA has a specificity of 94 % and a sensitivity of 95 % for spectral data of BC of stage 2 vs 3 (**Table 10**), a specificity of 82 %, and a sensitivity of 96 % for stage 4 vs stage 3 (**Table 12**), and specificity of 100 % and a sensitivity of 91 % for BC of stage 4 vs 2 (**Table 13**).

Table 10 PCA-FDA for controlled vs BC (stage 3 vs stage 2).

	BC	BC 3	Total	Specificity (%)
BC 2	2,770	140	2,910	94
BC 3	270	4,377	4,647	Sensitivity (%)
Total no. of spectra			7,557	95

Table 11 PCA-FDA for BC (stage 2 vs stage 3).

	BC 2	Controlled	Total	Specificity (%)
BC	4,360	40	4,400	100
Controlled	10	2,800	2,810	Sensitivity (%)
Total no. of spectra			7,210	99

Table 12 PCA-FDA for BC (stage 3 vs stage 4).

	BC 3	**BC 4**	**Total**	**Specificity (%)**
BC 3	4,300	250	4,550	82
BC 4	450	1,900	2,350	Sensitivity (%)
	Total no. of spectra		6,900	96

Table 13 PCA-FDA for BC (stage 2 vs stage 4).

	BC 2	**BC 4**	**Total**	**Specificity (%)**
BC 2	2,700	315	3,015	100
BC 4	10	1,843	1,853	Sensitivity (%)
	Total no. of spectra		4,868	91

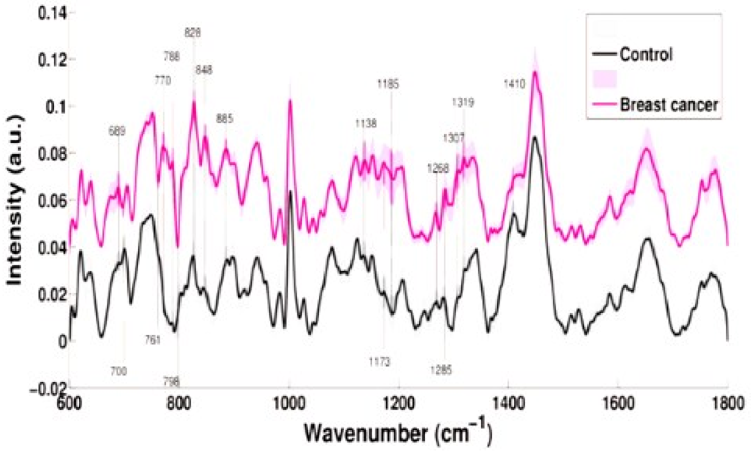


Figure 9 Healthy (black) and BC (magenta) blood plasma samples’ mean Raman spectra.

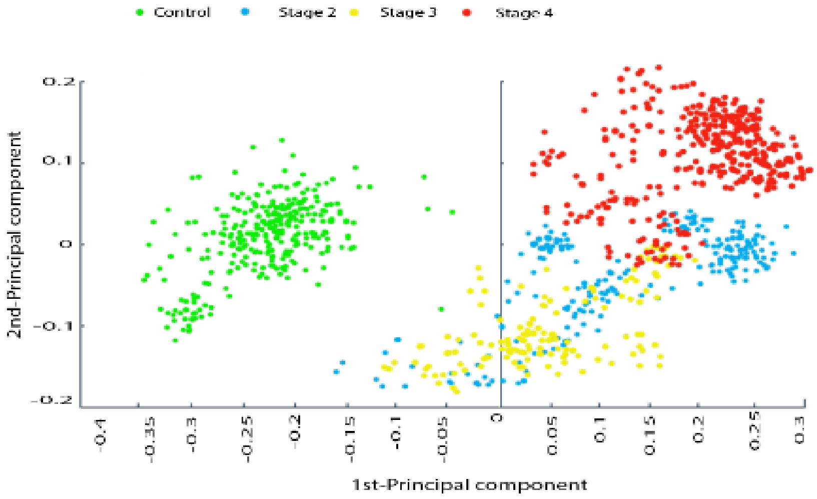


Figure 10 PCA scatter plot of healthy vs BC samples.

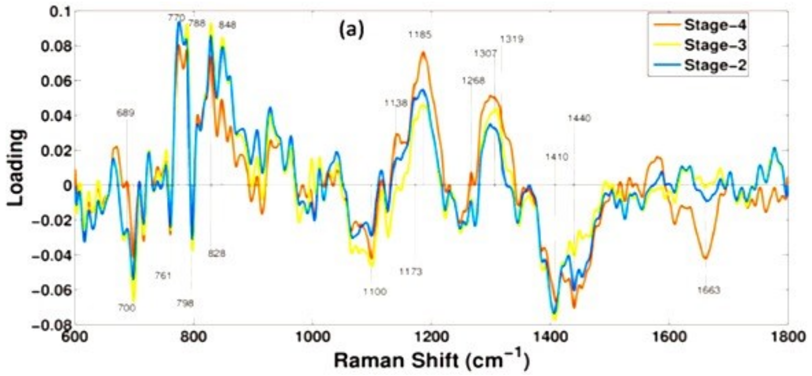


Figure 11 Pair-wise PCA analysis of different stages of BC vs healthy samples.

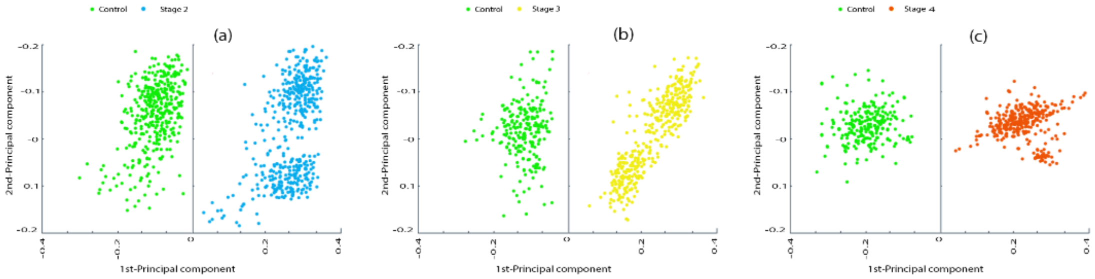


Figure 11 (a) PCA spectral plots of different stages of BC; healthy vs stage 2 (b), healthy vs stage 3 (c), and healthy vs stage 4 (d).

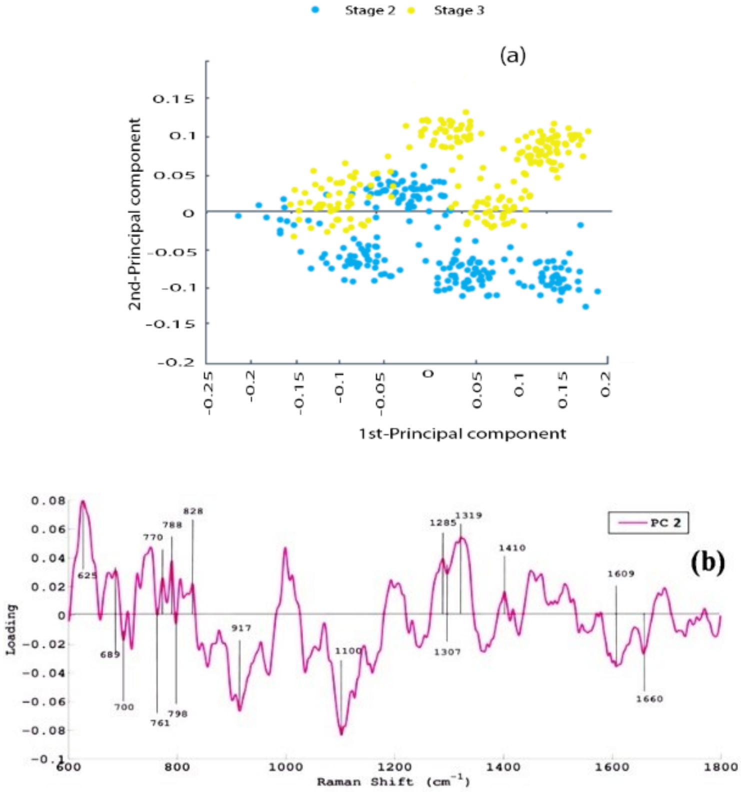


Figure 12 (a) BC analysis of stage 2 vs stage 3 spectral data using pair-wise PCA and (b) Pair-wise PCA loading of BC stage 2 vs stage 3 spectral data.

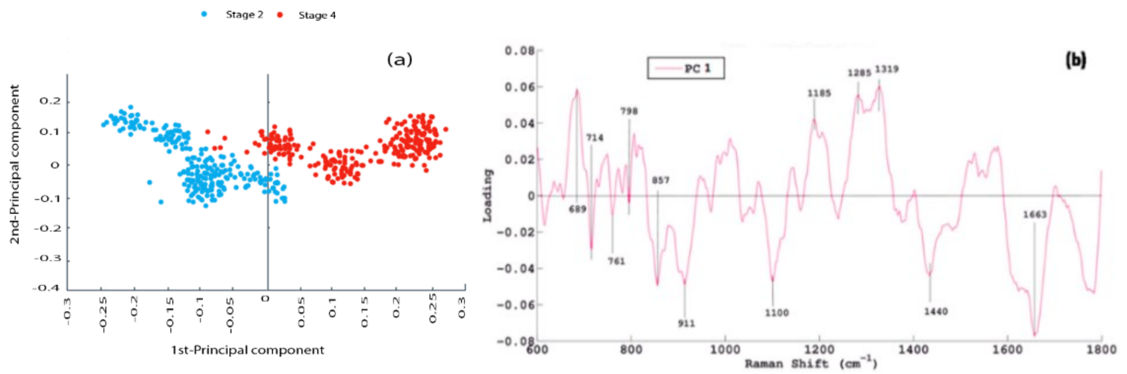


Figure 13 (a) BC stage 2 vs stage 4 Pair-wise PCA analysis and (b) PCA scatter plot - Raman spectral features having higher intensities in positive loadings for stage-4 than stage-3.

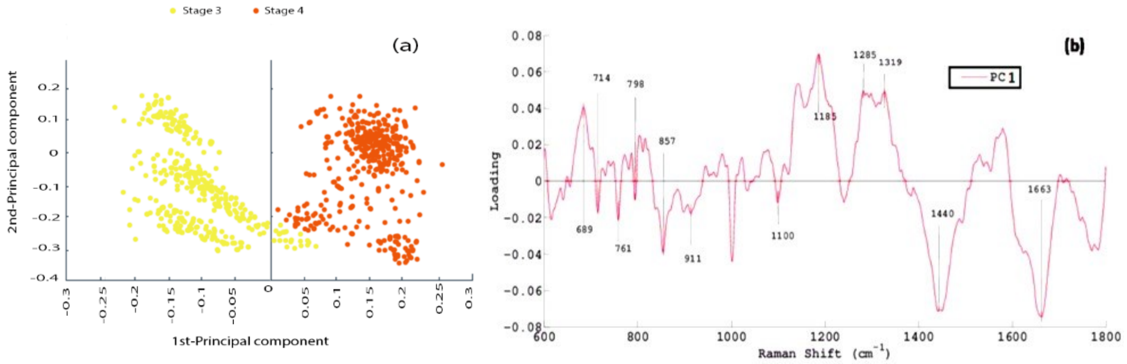


Figure 14 (a) Spectral patterns of BC stage 3 vs stage 4 pair-wise PCA analysis and (b) Raman spectral characteristics in PCA scatter plots have increased intensities in stage-4 compared to stage-3.

Conclusions

Breast neoplasm is a serious disease that must be treated with care. Early detection of this condition is extremely beneficial in saving millions of lives. The objective of this research work is to see if a Raman spectral dataset can be used to assess the probability of being affected by BC disease at an early stage. This research proposed and implements a stacked OGRU-LSTM hybrid model with Raman spectral data and multivariate data analysis. The approach of multivariate data analysis called PCA-FDA paired with OGRU-LSTM hybrid deep learning model has proved its ability for early diagnosis and staging of BC using blood samples obtained from clinically confirmed BC candidates. When developing the model with the necessary parameter tuning, interfering characteristics that have an impact on this condition were considered. Raman spectral characteristics related to DNA and proteins were monitored, which are only observed in BC candidates' plasma samples. Moreover, it is identified that several other spectral signatures in blood samples from candidates with various stages of BC differ significantly. PCA scatter plots revealed distinct characteristic differences between cancer states of BC. The distinction between stages 2 - 4 is critical because it leads to early detection, which can lead to more effective treatment. The stacked OGRU-LSTM model achieved the highest classification accuracy among other classifiers tested with the same spectral samples. Accuracy, Cohen-kappa score, F1-score, lowermost number of test losses, and MSE are indicating that the model outperforms other baseline classifiers with 97.89 % accuracy, 0.957 F1-score, 0.928 Cohen-kappa score, and 0.037 MSE, and the PCA-FDA gives potential differentiation ability to the system in the prediction of BC stages with a satisfactory result.

References

[1] VS Renjith and PSH Jose. A noninvasive approach using multi-tier deep learning classifier for the detection and classification of breast neoplasm based on the staging of tumor growth. *In*: Proceedings of the 2020 International Conference on Decision Aid Sciences and Application, Sakheer, Bahrain. 2020, p. 12-6.

[2] J Sathwara, S Bobdey and B Ganesh. Breast cancer survival studies in India: A review. *Int. J. Res. Med. Sci.* 2016; **4**, 3102-8.

[3] CK Anders, R Johnson, J Litton, M Phillips and A Bleyer. Breast cancer before age 40 years. *Semin. Oncol.* 2009; **36**, 237-49.

[4] AB Mariotto, KR Yabroff, Y Shao, EJ Feuer and ML Brown. Projections of the cost of cancer care in the United States: 2010-2020. J. Natl. Canc. Inst. 2011; **103**, 117-28.

[5] MM Rivera-Franco and E Leon-Rodriguez. Delays in breast cancer detection and treatment in developing countries. *Breast Canc. Basic Clin. Res.* 2018; **12**, 1178223417752677.

[6] PM Campeau, WD Foulkes and MD Tischkowitz. Hereditary breast cancer: New genetic developments, new therapeutic avenues. *Hum. Genet.* 2008; **124**, 31-42.

[7] AM Martin and BL Weber. Genetic and hormonal risk factors in breast cancer. *J. Egypt. Natl. Canc. Inst.* 2000; **92**, 1126-35.

[8] A Hans-Olov, D Hunter and D Trichopoulos. *Textbook of cancer epidemiology*. Oxford University Press, Oxford, 2008.

[9] B María-Ester and YV Navarro. Detección del cáncer de mama: Estado de la mamografía en México. *Cancerología* 2006; **1**, 147-62.

[10] JRF Caldeira, ÉC Prando, FC Quevedo, FAM Neto, CA Rainho and SR Rogatto. CDH1 promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer. *BMC Canc.* 2006; **6**, 48.

[11] G Agarwal, P Pradeep, V Aggarwal, CH Yip and PS Cheung. Spectrum of breast cancer in Asian women. *World J. Surg.* 2007; **31**, 1031-40.

[12] QB Li, XJ Sun, YZ Xu, LM Yang, YF Zhang, SF Weng, JS Shi and JG Wu. Diagnosis of gastric inflammation and malignancy in endoscopic biopsies based on Fourier transform infrared spectroscopy. *Clin. Chem.* 2005; **51**, 346-50.

[13] R Alfano, G Tang, A Pradhan, W Lam, D Choy and E Opher. Fluorescence spectra from cancerous and normal human breast and lung tissues. *IEEE J. Quant. Electron.* 1987; **23**, 1806-11.

[14] C Liu, R Alfano, W Sha, H Zhu, D Akins, J Cleary, R Prudente and E Cellmer. Human breast tissues studied by IR Fourier-transform Raman spectroscopy. *In*: Proceedings of the Conference on Lasers and Electro-Optics 1991, Maryland. 1991.

[15] Y Pu, W Wang, Y Yang and RR Alfano. Native fluorescence spectra of human cancerous and normal breast tissues analyzed with non-negative constraint methods. *Appl. Optic.* 2013; **52**, 1293-301.

[16] S Teh, W Zheng, K Ho, M Teh, K Yeoh and Z Huang. Near-infrared Raman spectroscopy for early diagnosis and typing of adenocarcinoma in the stomach. *Br. J. Surg.* 2010; **97**, 550-7.

[17] D Shen, G Wu and HI Suk. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 2017; **19**, 221-48.

[18] A Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. Nonlinear Phenom.* 2020; **404**, 132306.

[19] MA Mohammed, B Al-Khateeb, AN Rashid, DA Ibrahim, MKA Ghani and SA Mostafa. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. *Comput. Electr. Eng.* 2018; **70**, 871-82.

[20] RDH Devi and P Deepika. Performance comparison of various clustering techniques for diagnosis of breast cancer. *In*: Proceedings of the 2015 IEEE International Conference on Computational Intelligence and Computing Research, Madurai, India. 2015.

[21] MS Bergholt, W Zheng and Z Huang. Characterizing variability in *in vivo* Raman spectroscopic properties of different anatomical sites of normal tissue in the oral cavity. *J. Raman. Spectros.* 2012; **43**, 255-62.

[22] PJ Sudharshan, C Petitjean, F Spanhol, LE Oliveira, L Heutte and P Honeine. Multiple instances learning for histopathological breast cancer image classification. *Expert Syst. Appl.* 2019; **117**, 103-11.

[23] SP Singh, A Deshmukh and P Chaturvedi. *In vivo* Raman spectroscopic identification of premalignant lesions in oral buccal mucosa. *J. Biomed. Optic.* 2012; **17**, 105002.

[24] LP Choo-Smith, HGM Edwards, HP Endtz, JM Kros, F Heule, H Barr, JS Robinson, HA Bruining and GJ Puppels. Medical applications of Raman spectroscopy: From proof of principle to clinical implementation. *Biopolymers* 2002; **67**, 1-9.

[25] C Kallaway, LM Almond, H Barr, J Wood, J Hutchings, C Kendall and N Stone. Advances in the clinical application of Raman spectroscopy for cancer diagnostics. *Photodiagnosis Photodynamic Ther.* 2013; **10**, 207-19.

[26] H Krishna, SK Majumder and P Chaturvedi. *In vivo* Raman spectroscopy for detection of oral neoplasia: A pilot clinical study. *J. Biophotonics* 2014; **7**, 690-702.

[27] LFCS Carvalho, F Bonnier, C Tellez, LD Santos, K O'Callaghan, J O'Sullivan, LES Soares, S Flint, AA Martin, FM Lyng and HJ Byrne. Raman spectroscopic analysis of oral cells in the high wavenumber region. *Exp. Mol. Pathol.* 2017; **103**, 255-62.

[28] K Venkatakrishna, J Kurien and KM Pai. Optical pathology of oral tissue: A Raman spectroscopy diagnostic method. *Curr. Sci.* 2001; **80**, 665-8.

[29] R Malini, K Venkatakrishna, J Kurien, KM Pai, L Rao, VB Kartha and CM Krishna. Discrimination of normal, inflammatory, premalignant, and malignant oral tissue: A Raman spectroscopy study. *Biopolymers* 2006; **81**, 179-93.

[30] ST Sonis, RC Fazio, L Fang and BB Ward. Raman spectroscopic analysis of oral tissues at different anatomic sites. *J. Biomed. Opt.* 2006; **11**, 024006.

[31] ZC Lipton, J Berkowitz and C Elkan. A critical review of recurrent neural networks for sequence learning. *Mach. Learn.* 2015; **1**, 1-38.

[32] F Beritelli, G Capizzi, GL Sciuto, C Napoli and M Woźniak. A novel training method to preserve generalization of RBPNN classifiers applied to ECG signals diagnosis. *Neural Network.* 2018; **108**, 331-8.

[33] X Wang, J Xu, W Shi and J Liu. OGRU: An optimized gated recurrent unit neural network. *J. Phys. Conf.* 2019; **1325**, 012089.

[34] X Song and Y Li. Data gathering in wireless sensor networks via regular low density parity check matrix. *IEEE CAA J. Automatica Sin.* 2018; **5**, 83-91.

[35] D Polap and M Wozniak. Lung segmentation on x-ray images with neural validation. *In*: Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence, Hawaii. 2017, p. 1-7.

[36] D Polap and M Wozniak. Bacteria shape classification by the use of region covariance and convolutional neural network. *In*: Proceedings of the 2019 International Joint Conference on Neural Networks, Budapest, Hungary. 2019, p. 1-7.

[37] CM Krishna, GD Sockalingum, J Kurien, L Rao, L Venteo, M Pluot, M Manfait and VB Kartha. Micro-Raman spectroscopy for optical pathology of oral squamous cell carcinoma. *Appl. Spectros.* 2004; **58**, 1128-35.

[38] JD Gelder, KD Gussem, P Vandenabeele and L Moens. Reference database of Raman spectra of biological molecules. *J. Raman Spectros.* 2007; **38**, 1133-47.

[39] AD Meade, FM Lyng, P Knief and HJ Byrne. Growth substrate induced functional changes elucidated by FTIR and Raman spectroscopy in *in-vitro* cultured human keratinocytes. *Anal. Bioanalytical Chem.* 2007; **387**, 1717-28.

[40] I Notingher. Raman spectroscopy cell-based biosensors. *Sensors* 2007; **7**, 1343-58.

[41] P Jess, V Garcés-Chávez, D Smith, M Mazilu, L Paterson, A Riches, C Herrington, W Sibbett and K Dholakia. Dual beam fibre trap for Raman microspectroscopy of single cells. *Optic. Express* 2006; **14**, 5779-91.

[42] Z Movasaghi, S Rehman and IU Rehman. Raman spectroscopy of biological tissues. *Appl. Spectros. Rev.* 2007; **42**, 493-541.

[43] H Nawaz, F Bonnier, P Knief, O Howe, FM Lyng, AD Meade and HJ Byrne. Evaluation of the potential of Raman microspectroscopy for prediction of chemotherapeutic response to cisplatin in lung adenocarcinoma. *Analyst* 2010; **135**, 3070-6.

[44] D Bertrand, P Courcoux, JC Autran, R Meritan and P Robert. Stepwise canonical discriminant analysis of continuous digitalized signals: Application to chromatograms of wheat proteins. *J. Chemometr.* 1990; **4**, 413-27.

[45] F Chollet. Keras. Available at: https://keras.io, accessed March 2023.

[46] M Abadi, P Barham, J Chen, Z Chen, A Davis, J Dean, M Devin, S Ghemawat, G Irving, M Isard, M Kudlur, J Levenberg, R Monga, S Moore, DG Murray, B Steiner, P Tucker, V Vasudevan, P Warden, M Wicke, Y Yu and X Zheng. TensorFlow: A system for large-scale machine learning. *In*: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Georgia. 2016, p. 265-83.

[47] M Bilal, M Bilal, S Tabassum, M Saleem, H Mahmood, U Sarwar, H Bangush, F Munir, MA Zia and M Ahmed. Optical screening of female breast cancer from whole blood using Raman spectroscopy. *Appl. Spectros.* 2017; **71**, 1004-13.

[48] C Li, L Zhang, D Peng, X Yi, S He, F Liu, X Zheng, W Huang, L Zhao and X Huang. Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochim. Acta Mol. Biomol. Spectros.* 2021; **264**, 120300.

[49] L Shang, D Ma, J Tang, Y Bao, J Fu and J Yin. Classifying breast cancer tissue by Raman spectroscopy with one-dimensional convolutional neural network. *Spectrochim. Acta Mol. Biomol. Spectros.* 2021; **256**, 119732.

[50] SKi Koya, M Brusatori, S Yurgelevic, C Huang, CW Werner, RE Kast, J Shanley, M Sherman, KV Honn, KR Maddipati and GW Auner. Accurate identification of breast cancer margins in microenvironments of ex-vivo basal and luminal breast cancer tissues using Raman spectroscopy. *Prostag. Other Lipid Mediat.* 2020; **151**, 106475.

[51] J Ming-Jer Jeng, M Sharma, L Sharma, C Ting-Yu, H Shiang-Fu, C Liann-Be, W Shih-Lin and L Chow. Raman spectroscopy analysis for optical diagnosis of oral cancer detection. *J. Clin. Med.* 2019; **8**, 1313.

[52] J Xia, L Zhu, M Yu, T Zhang, Z Zhu, X Lou, G Sun and M Dong. Analysis and classification of oral tongue squamous cell carcinoma based on Raman spectroscopy and convolutional neural networks. *J. Mod. Optic.* 2020; **67**, 481-9.

[53] P Žuvela, K Lin, C Shu, W Zheng, CM Lim and Z Huang. Fiber-optic Raman spectroscopy with nature-inspired genetic algorithms enhances real-time *in vivo* detection and diagnosis of nasopharyngeal carcinoma. *Anal. Chem.* 2019; **91**, 8101-8.

[54] Y Zhao, K Rong and AL Tan. Qualitative classification of estrogen powder by Raman spectroscopy based on one-dimensional convolutional neural network. *Spectros. Spectral Anal.* 2019; **39**, 3755-60.