IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Evaluation of Classifiers Performance in Cervical Cancer Detection

Rajpriya R.[1*], Saravanan M.S.[2]

[1]Dept. of Computer Science, Bharathiar University, Coimbatore, India
[2]Dept. of Information Technology, Saveetha School of Engineering, Chennai, India

Corresponding Author: rajpriyapranav@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.10291035 | Available online at: www.ijcseonline.org

Accepted: 19/Apr/2019, Published: 30/Apr/2019

Abstract-Artificial Intelligence (AI) plays an important role in many medical diagnosis systems. AI techniques uses for classifying the normal and abnormal cells are present in the cervix in the region of uterus. The classification of cancerous and non-cancerous cervical cells is detected by using AI techniques which gives accurate results. Compare to manual screening techniques like Pap smear test, the AI techniques gives better results and less time consuming. This paper presents several classifiers are used to classifies the normal and abnormal cells of Pap smear images.

Keywords: Cervical cancer, Support Vector Machine, Discriminant analysis, Decision tree, K-nearest neighbor.

I. INTRODUCTION

Cervical cancer is the second most common cancer in worldwide affected by all age groups. If the cervical cancer is detected and identified in early stage means, it can be cured. Pap smear test is an important technique used for diagnosing the cervical cells in manual method. The primary cause of cervical cancer is affected by Human Papilloma Virus (HPV) infection. Pap smear procedure is fully dependent by humans and suffers accurate results. Artificial intelligence techniques used to detecting cervical cancer which produces high accurate results in easy manner. The process of extracting features from Pap smear images using with automated diagnosis system. It can be achieved through size, shape, texture, and color features used to detect the cellular changes are present in the Pap smear slides. The main objective of this paper is to analysis of dataset used with several classification techniques to detect the cancerous and non-cancerous cervical cells are present in the dataset. Table.1 explains the implementation details of classification of cancerous and non-cancerous cells are present in Pap smear images.

Section I contains the introduction of cervical cancer detection, Section II contains the related work of detection of cervical cancer, Section III explains materials and methods of classifiers, Section IV contains results and discussion, and Section V contains the conclusion.

Table 1: Implementation details of proposed method

Software	MATLAB R2015a
Data sets	Pap Smear Slide images
Classification Techniques	Support vector machine, K-nearest neighbor, Decision tree, Discriminant analysis
Operating System	Windows 7
Data set File Format	Image (jpeg format)
Purpose	Classification of cancerous and non-cancerous Pap smear images

II. RELATED WORK

Rajeev Gupta, Abid Sarwar [1], proposes analysis of artificial intelligence for detection of cervical cancer from digitalized Papanicolaou Smear images. The proposed algorithm detects the random subset space, decorate, ensemble of nested dichotomies called END. Random forest algorithm and decision tree are used as classifiers for detecting random sub-space. Binary trees are used to transform the multiclass categorization problem. Naïve Bayes classifier assumes the independent of the presence or absence of other attributes. Backpropagation algorithm determines the activation of neurons in the network. Radial basis function network detects the non-linear hidden layers and linear output layers. Decision tree converts the data into a decision table by tracing the rows of the table. The author implements the algorithms for detecting the cells infections and classifying the cells. The author achieves the efficiency of algorithms was in the range of 69% to 76%. Priyanka K.Malli, Dr. Suvarna Nandyal [2], submits a machine learning technique for detection of cervical cancer. The comprehensive machine learning proposed to obtain the color and shape of nucleus and cytoplasm of the cervix. Fuzzy C-means clustering aims to clustering the one piece of data to belong to two or more clusters. Morphology operations are extracting the features of nucleus and cytoplasm (area, size and shape). K-nearest neighbor classification used to detect the shape and color features of image. To analyze the classes and classifying the features of images are expeted by Artificial Neural Network (ANN). The author achieves the result of accuracy 88.04% for classification techniques. M.Anousouya Devi, S.Ravi et.al [3] in their publication aims to presents a methods or classifying the cancerous cells. The classification of normal, abnormal and cancerous cells is identified by artificial neural network. Multilayer Perceptron (MLP) used two phases, image pre-processing and feed forward MPL neural networks. To extract the cervical features for contour detection used with MLP. To analyze the feature parameters of nucleus measured by principal component analysis. The author express the neural network algorithms for increases the efficiency of the accuracy in results. Chandraprabha R, Seema singh [4] presents a method for detection of cervical cancer. Artificial Neural Network (ANN) and Fuzzy Logic (FL) are used to classifying the cancerous cells. Back-Propagation algorithm identifies the features of nucleus (area, perimeter, eccentricity, ratio, and color-intensity). Hybrid Multi Layered Perceptron (HMLP) network used to segmenting the nucleus and cytoplasm for computing morphometric. Support Vector Machine (SVM) classifier is used to distinguish between normal and abnormal cells. The authors are intending to improve the limitations of the algorithm for detecting cancerous cells. Abid Sarwar, Vinod Sharma et.al [5], proposes the hybrid ensemble technique for improving the screening of cervical cancer and classifying the Pap smear images. The proposed method used to screen the cervical cancer and classification of Pap smear. The author achieves better results by using this method and the efficiency result is 96% for 2-class problem. N.Ganesan, K.Venkatesh et.al [6], work to diagnose the cancer disease by neural network Multi-Layer Perceptron (MLP) specifies the number of units are present in these layers and the number of hidden layers. The author intends to plan for automatic diagnosis used by neural network. Miao Wu, Chuanbo yan et.al [7], proposes the deep convolutional neural network model for recognizing and classifying the cytological images. The propsed method achieves the classification accuracy is 93.33% and the error rate is 3.85% was achieved. M.K.Soumya, K.Sneha and C.Arunvinodh [8], proposes texture analysis of cervical cancer and classification with Support Vector Machine (SVM) in Matlab. The proposed system accuracy is measured as 81% of dataset1, 82% of dataset2, and 83% of dataset3. Masakazu Sato, Koji Horie, Aki Hara [9], submits their work for classification of images from colposcopy using deep learning model. The proposed method of deep learning with keras neural network and Tensor flow libraries gives the accuracy was ~50%. Yessi Jusman, Siti Noraini Sulaiman et.al [10], proposes the Multi-Layered Perceptron (MLP) network using with Levenberg-Marquardt Backpropagation algorithm for classification of cervical cell types. The result of the proposed method gives 97.3% of highest accuracy.

III. MATERIALS AND METHODS

The classification system of Pap smear images involves: preprocessing, feature enhancement, feature extraction, and classification of cancerous and non-cancerous cells are present in the Pap smear slides. Fig.1 shows the classification systems of proposed methods.

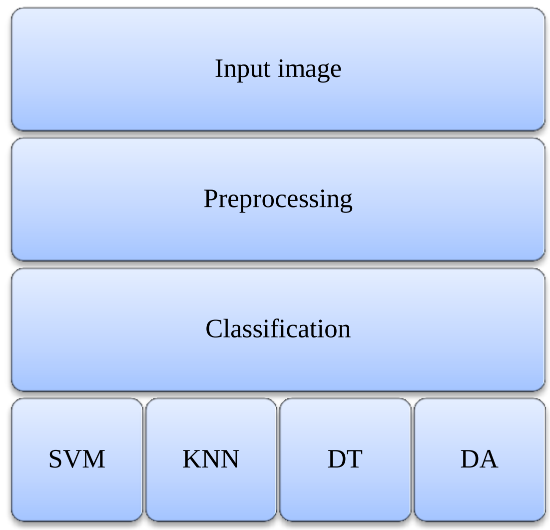

Fig: 1 Classification Systems

A. Preprocessing of Pap smear Images

The Pap smear images are acquired by skilled cytotechnicians. The publically available database is given by Herlev University Hospital, Denmark. In this database, Table.2 shows several Pap smear images with seven stages of cancer and cell images were captured with a resolution of $0.201 \frac{\mu m}{pixel}$. The collected images are enhanced by filtering techniques for removing noise reduction and segmenting the cell nucleus and cytoplasm. The enhanced images are used to extract the size, shape, texture, and color-intensity information for classifying the cancerous and non-cancerous cells.

Table.2 Summary of Datasets

Datasets (Pap smear slides)	Number of instances	Feature types
Normal superficial	54	Integer
Normal intermediate	37	
Normal columnar	36	

Light dysplastic	37	real categorical
Moderate dysplastic	21	
Severe dysplastic	23	
Carcinoma_in_situ	20	
Total number of extracted features= 22(7 classes) Total number of input slides= 228 images		

A. Classification Techniques

There are several classifiers are used to classify the normal and abnormal cancerous cells are present in the Pap smear images. In this paper, four classifiers are implemented for classifying the cancerous and non-cancerous cells from Pap smear slide images. Input (Pap smear) images are used to extract the size, shape, texture, and color- intensity information of cells and nucleus. The extracted information's are converted into table for detecting and classifying the cancer.

1) Support Vector Machine (SVM)

It is a supervised machine learning algorithm which can be used for both classification and regression challenges. In SVM, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinates [11]. The objective of SVM is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Support vectors are data points are closer to the hyperplane and influence the position and orientation of the hyperplane [12]. Using these support vectors, we maximize the margin of the classifier. Fig.2 shows the results of the confusion matrix for SVM.

Linear kernel: Class function $G(x_1, x_2)$, linear space S and a function ψ mapping x.

$$G(x_1, x_2) = <\psi(x_1), \psi(x_2)> \quad (1)$$

Polynomial kernel:

$$G(x_1, x_2) = (1 + x_1' x_2)P \quad (2)$$

Where P = positive integer.

Radial basis function (Gaussian):

$$G(x_1, x_2) = \exp(-\|x_1 - x_2\|^2) \quad (3)$$

Sigmoid:

$$G(x_1, x_2) = \tanh(P_1 x_1' x_2 + P_2) \quad (4)$$

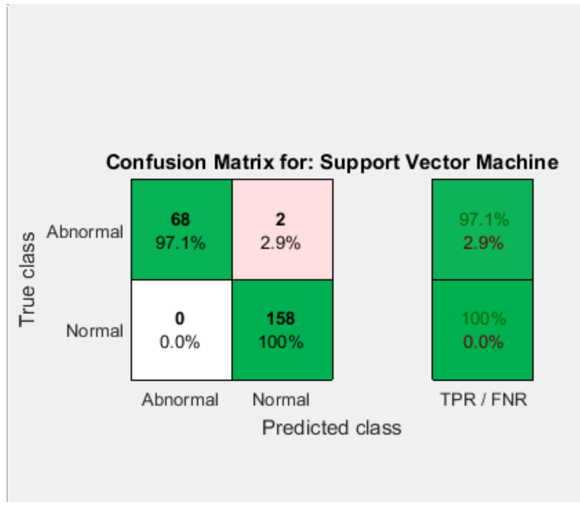

Fig. 2 Confusion Matrix for SVM

2) K-nearest Neighbors algorithm (KNN)

It is non-parametric method used for classification and regression. KNN is used to assign weight to the contributions of the neighbors, so that the nearest neighbors contribute more to the average than the more distant ones [13]. This algorithm is based on feature similarity, which is how closely out-of-sample features resemble our training set determines how we classify a given data point. This algorithm stores all available cases and classifies new cases based on a similarity measure. In classification, the output can be calculated as the class with the highest frequency from the K-most similar instances. Each instance in essence votes for their class and the class with the most votes is taken as the prediction. The purpose of this algorithm is to classify a new object base on attributed and training samples [14].

If you also specify prior or weights, then the software takes the observation weights into account. Specifically, the weighted mean of predictor *j* is

$$\bar{x_j} = \sum_{B_j} w_k \, x_{jk} \quad (5)$$

The weighted standard deviation is

$$S_j = \sum_{B_j} w_k \left(x_{jk} - \bar{x}_j\right) \quad (6)$$

Where B_j is the set of indices k for which x_{jk} and w_k are not missing. Fig.3 shows the result of the confusion matrix for KNN.

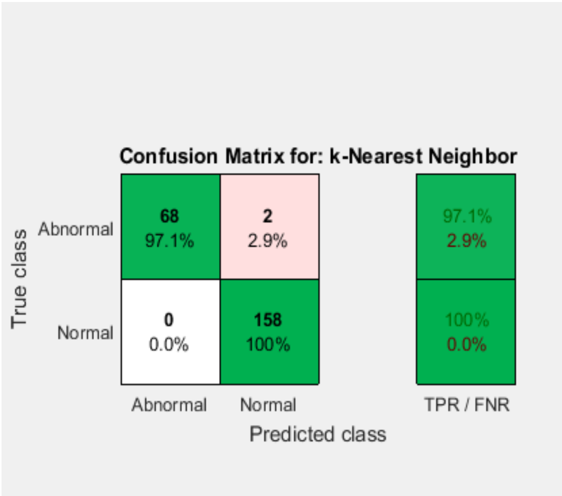

Fig.3 Confusion Matrix for KNN

3) *Decision Tree (DT)*

It is the most powerful tool for classification and prediction. It is like a tree structure, where each internal node denotes test as an attributes, each branch represents an outcome of the test and each leaf node holds a class label [15]. The tree can be splitting the source set into subsets based on an attribute value test. In recursive manner, each subset is derived by partitioning the subset. It can handle high dimensional data and gives good accuracy. It is a predictive modeling tool that splits a data set based on different conditions. The goal of decision tree is to create a model and it predicts the value of a target variable by learning decision rules inferred from the data features [16].

Estimates the probability that an observation is in node t using

$$P(T) = \sum_{j \in T} w_j \qquad (7)$$

Where w_j is the weight observation j, and T is the set of all observation indices in node T. If you do not specify prior or weights, then $w_j = 1/n$, where n is the sample size.

Classification tree splits nodes based on either impurity or node error [17]. Impurity means one of several things, depending on your choice of the split criterion name-value pair argument.

Gini's Diversity Index (gdi) — The Gini index of a node is

$$1 - \sum_i p^2 (i) \qquad (8)$$

Where the sum is over the classes *i* at the node, and $p(i)$ is the observed fraction of classes with class *i* that reach the node. A node with just one class (a *pure* node) has Gini index 0; otherwise the Gini index is positive. So the Gini index is a measure of node impurity.

Deviance ('deviance') — with $p\ (i)$ defined the same as for the Gini index, the deviance of a node is

$$\sum_i P(i) \log_2 P(i) \qquad (9)$$

A pure node has deviance 0; otherwise, the deviance is positive.

Twoing rule ('twoing') — Twoing is not a purity measure of a node, but is a different measure for deciding how to split a node. Let $L(i)$ denote the fraction of members of class *i* in the left child node after a split, and $R(i)$ denote the fraction of members of class *i* in the right child node after a split. Choose the split criterion to maximize

$$P(L)P(R)\left(\sum_i |L(i) - R(i)|\right)^2 \qquad (10)$$

where $P(L)$ and $P(R)$ are the fractions of observations that split to the left and right respectively. If the expression is large, the split made each child node purer. Similarly, if the expression is small, the split made each child node similar to each other, and therefore similar to the parent node. The split did not increase node purity.

Node error — The node error is the fraction of misclassified classes at a node. If *j* is the class with the largest number of training samples at a node, the node error is

$$1 - P(j) \qquad (11)$$

A surrogate decision split is an alternative to the optimal decision split at a given node in a decision tree. The optimal split is found by growing the tree; the surrogate split uses a similar or correlated predictor variable and split criterion [18]. When the value of the optimal split predictor for an observation is missing, the observation is sent to the left or right child node using the best surrogate predictor. When the value of the best surrogate split predictor for the observation is also missing, the observation is sent to the left or right child node using the second-best surrogate predictor, and so on. Candidate splits are sorted in descending order by their predictive measure of association. The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association. Fig.4 shows the results for DT.

Suppose x_j and x_k are predictor variables *j* and *k*, respectively, and $j \neq k$. At node *t*, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{min(P_L, P_R) - \left(1 - P_{L_jL_k} - P_{R_jR_k}\right)}{min(P_L, P_R)} \qquad (12)$$

P_L is the proportion of observations in node *t*, such that $x_j < u$. The subscript *L* stands for the left child of node *t*.

P_R is the proportion of observations in node *t*, such that $x_j \geq u$. The subscript *R* stands for the right child of node *t*.

P_{LjLk} is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.
P_{RjRk} is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.
Observations with missing values for x_j or x_k do not contribute to the proportion calculations.
λ_{jk} is a value in $(-\infty,1]$. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

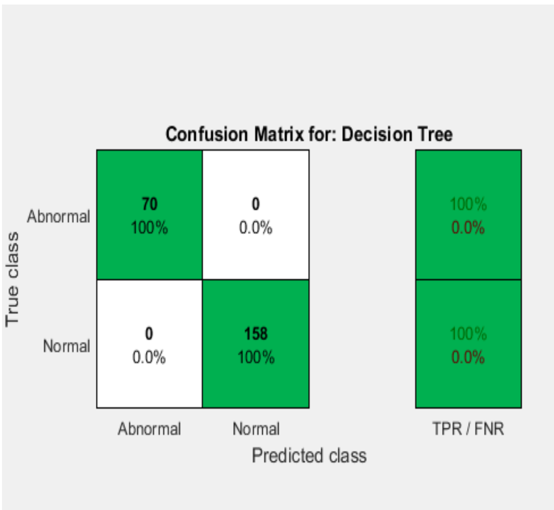

Fig.4 Confusion Matrix for DT

4) Discriminant Analysis (DA)

It is a statistical tool with an objective to assess the adequacy of a classification. It gives the group memberships or to assign objects to one group among a number of groups [19]. This algorithm is used to determine which predictor variables are related to the dependent variable and to predict the value of the dependent variable given certain values of the predictor variables. It is used to find a set of prediction based on independent variables that are used to classify individual into groups. The objectives of discriminant analysis are to finding a predictive equation to understand the relationships that may exist among the variables [20]. Fig.5 shows the results of the confusion matrix for DA.
The model for discriminant analysis is:
Each class (Y) generates data (X) using a multivariate normal distribution. That is, the model assumes X has a Gaussian mixture distribution (gmdistribution).
For linear discriminant analysis, the model has the same covariance matrix for each class, only the means vary.
For quadratic discriminant analysis, both means and co-variances of each class vary.
Predict classifies so as to minimize the expected classification cost:

$$\hat{y} = arg \min_{y=1,2,..k} \sum_{k=1}^{k} \hat{P}\left(\frac{k}{x}\right) C\left(\frac{y}{k}\right) \quad (13)$$

Where

- $\hat{y}$ is the predicted classification.
- k is the number of classes.
- $\hat{P}\left(\frac{k}{x}\right)$ is the posterior probability of class k for observation x.
- $C\left(\frac{y}{k}\right)$ is the cost of classifying an observation as y when its true class is k.

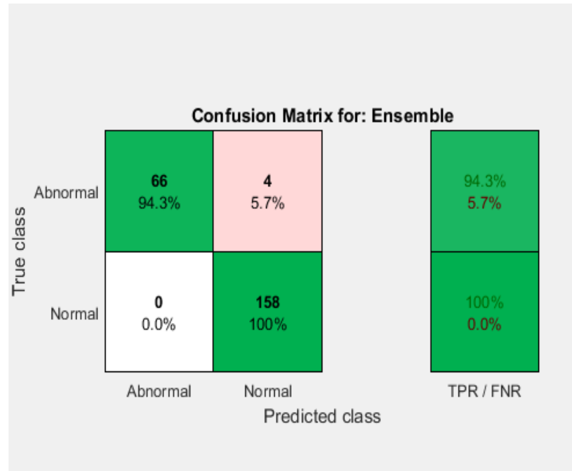

Fig.5 Confusion Matrix for DA

IV. RESULTS AND DISCUSSION

The aim of this paper is to find the best artificial intelligence algorithms for cervical cancer detection. The four different classifiers are implemented in MATLAB. The classifiers are applied with 10-fold cross validation in order to check the results of every class. There are 228 slides used to extract the features and classify the normal and abnormal cells are present in the Pap smear images. Fig.6 shows the image of normal and abnormal Pap smears. The Pap smear slides are used to extract the information of cervical cancerous and non-cancerous cells. Table.3 shows the extracted information of inputs. The classifiers are implemented for classifying the cancerous and non-cancerous extracted data from the Pap smear slides. Table.4 shows the abnormal and normal accuracy results and error rates of 2-class problems using various techniques such as SVM, KNN, DT and DA.

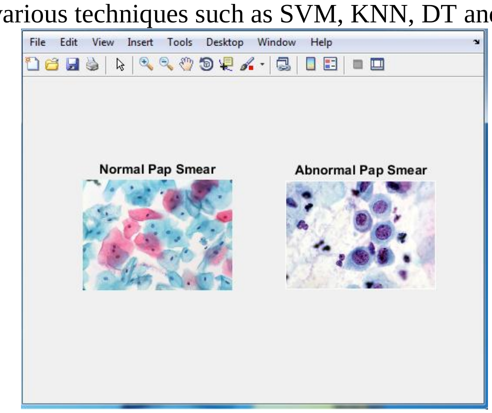

Fig.6 Normal and Abnormal Pap smear

Table 3: Features information (extracted)

Feature Category	Features	Number of Features
Fuzzy logic (Edge-based)	Contrast, Correlation, Energy, Homogeneity, Entropy, and Kurtosis	7
Threshold (Region-based)	Area, MajorAxisLength, MinorAxisLength, Eccentricity, Orientation, ConvexArea, EquivDiameter, Solidity, Extent, and Perimeter	10
Fuzzy C-means Clustering (Color intensity-based)	Mean, Median, Standard Deviation, Covariance, and Range	5
Total number of features extracted:22 features		

Table 4: Accuracy and Error rate of 2 Class-problems

Techniques	Normal		Abnormal	
Validation (10 fold)	Over all Accuracy	Over all Error rate	Over all Accuracy	Over all Error rate
SVM	100%	0.0%	99.1%	0.9%
KNN	100%	0.0%	99.1%	0.9%
DT	100%	0.0%	100%	0.0%
DA	100%	0.0%	98.2%	1.8%

V. CONCLUSION

Artificial Intelligence (AI) plays an important role in medical image applications due its accuracy and better results. AI uses many algorithms which classifies the normal and abnormal cervical cells at the earliest stage. In this paper, different types of method based on AI for detection of cervical cancer are discussed.

REFERENCES

[1] Rajeev Gupta, Abid Sarwar, Vinod Sharma, "Screening of cervical cancer by Artificial Intelligence based Analysis of Digitized Papanicolaou-Smear Images", International Journal of Contemporary Medical Research 2017; 4(5): 1108-1113.

[2] Priyanka K Malli, Dr.Suvarna Nandyal, "Machine learning technique for detection of cervical cancer using KNN and Artificial neural network", International journal of Emerging Trends and Technology in Computer Science(IJETICS), volume 6, Issue 4, July-Aug 2017.

[3] M.Anousouya Devi, S.Ravi, et.al, "Classification of cervical cancer using Artificial Neural Network", IMCIP-2016, Procedia computer science, 89 (2016), 465-472.

[4] Chandraprabha R, Seema singh, "Artificial Intelligent System for diagnosis of cervical cancer: A Brief Review and Future outline", IJLRET, NC3PS-2016, pp.38-41.

[5] Abid sarwar, Vinod sharma, Rajeev gupta, "Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis", Medicine Universe, July 2015, vol-4, p.54-62.

[6] Dr.N.Ganesan, Dr.K.Venkatesh, et.al, "Application of Neural Networks in Diagnosing Cancer Disease using Demographic Data", International Journal of Computer Application, 2010, volume 1-No.26.

[7] Miao wu, Chuanboyan, Huiquiang Liu, et.al, "Automatic Classification of Cervical Cancer rom Cytological images by using Convolutional Neural Network", Portland press, Bioscience Reports (2018), 38, BSR20181769.

[8] M.K.Soumya, K.Snwha and C.Arunvinodh, "Cervical Cancer Detection and Classification using Texture Analysis", Biomedical and Pharmacology Journal, vol.9 (2), 663-671 (2016).

[9] Masakazu Sato, Koji Horie, Aki Hara, "Application of Deep Learning to the Classification of images from Colposcopy",Oncology letters, 15:3518-3523, 2018.

[10] Yessi Jusman, Siti Noraini Sulaiman, Nor Ashidi, " Capability of New Features from FTIR Spectral of Cervical Cells for Cervical Precancerous
Diagnostic System using MLP Networks", IEEE, 978-1-4244-4547-9/09, TENCON 2009.

[11] A. S. Phatak and B, P. A., Classification of MR Images of Cervical Cancer Using SVM and ANN Engineering, issue 2277, (2015).

[12] B.Ashok, Dr.Aruna "comparison of Feature selection methods for Diagnosis of Cervical cancer using SVM classifier" Journal of Engineering Research Applications IISN:2248-9622,vol.6,Issue 1,(Part-1) January2016 pp.94-99

[13] Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification,"Computers in Biology and Medicine,vol.39,no.1,pp.69–78,2009

[14] P.K.Malli,S.Nandyal,"Machine learning Technique for detection of Cervical Cancer using k-NN and Artificial Neural Network", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2017

[15] Y. LeCun, Y. Bengioand, and G. E. Hinton. "Deep learning", Nature, 521:436–444, 2015

[16] M.A.Devi,S.Ravi,J.Vaishnavi and S.Punitha,"Classification of Cervical Cancer using Artificial Neural Networks", Procedia Computer Science 89, pp.465 – 472, 2016.

[17] B.Sharma and K.K.Mangat, "Various Techniques for Classification and Segmentation of Cervical Cell Images - A Review", International Journal of Computer Applications (0975 – 8887), vol 147, 2016.

[18] L.Thampi and V.Paul,"Automatic Segmentation and Classification in Cervical Cancer Images: Evaluation and Challenges", International Journal of Pure and Applied Mathematics, vol 119, 2018

[19] Kitchener HC, Blanks R, Cubie H, Desai M, Dunn G, Legood R, Gray a, Sadique Z, Moss S. MAVARIC "A comparison of automation-assisted and manual cervical screening: A randomised controlled trial", Health technology assessment (Winchester,

England) [Internet]. 2011 Jan;15:iii–iv, ix–xi, 1–170. pubmed/21266159.

[20] Trevisan J, Angelov PP, Carmichael PL, Scott AD, Martin FL. "Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) bio-spectroscopy datasets: current practices to future perspectives", The Analyst. 2012 Jul; 137:3202-3215.

Author's Profile

R.RAJPRIYA is pursuing Ph.D. in the stream of computer science from Bharathiar University, Coimbatore. She is working as an Assistant Professor in the Department of Computer Application, Bhaktavatsalam Memorial College, Chennai. Her research area is based on detection of cervical cancer using artificial intelligence techniques.

M.S.Saravanan has obtained Ph.D. from the Bharathiar University in the year 2013 in India. He is currently working as Professor in the Department of Information Technology in Saveetha University, Chennai, India. He is also a member of IEEE, CSI (Computer Society of India), ISTE (Indian Society of Technical Education) and the registered member of IAENG (International Association of Engineers). He has published eighty-six international publications and presented twenty-four research papers in international and national conferences, having 19 plus years of teaching experience in various institutions in India and Abroad.