International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I10P230

Volume 70 Issue 10, 307-318, October 2022

Original Article

Classifying Pap Smear Images with an Advanced Composite Random Forest Model

Sharmistha Bhattacharjee[1], Dipankar Ray[2], Diganta Saha[3], D. Sobya[4]

[1]*National Institute of Electronics & Information Technology (NIELIT, KOLKATA), Jadavpur University Campus, Kolkata, West Bengal, India.*
[2]*Automation Centre, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand, India.*
[3]*Department of Computer Science & Engineering, Jadavpur University, Kolkata, West Bengal, India.*
[4]*Department of Electronics and Communication Engineering, RajaRajeswari College of Engineering, Bengaluru, Karnataka, India*

[1]*Corresponding Author : sharmi9870@yahoo.com*

Received: 29 August 2022 Revised: 13 October 2022 Accepted: 18 October 2022 Published: 20 October 2022

Abstract *- Manual screening and diagnosis of conventional Pap-smear slides for cervical cancer diagnosis is slow and suffers from human error. Here we have proposed a hybrid-deep-learning model achieved using k-means cluster and Random Forest models, which aims to identify the most prevailing characteristics of cervical tissues and classify them into different cytopathological classes. Just because the texture, shape (morphometric), and color of the nucleus and cytoplasm together or individually play a vital role in PAP smear image classification, fifteen prominent features are extracted based on it to classify images collected from the Herlev Pap Smear dataset. Gray Level Covariance Matrix and Gabor Filter helped extract the texture-based features, whereas morphometric and color-based characteristics were abstracted using Canny's edge detection and histogram analysis. In addition, a new and advanced cutting-edge compound random forest model is constructed to categorize the PAP smear photos. It was noted that the suggested hybrid approach offers up to 99% effectiveness. Additionally, this study also demonstrated a thorough comparison of the suggested model. It was observed that the suggested model also performs admirably when measured against Support Vector Machine (SVM) and Deep-Multilayer Perceptron methods.*

Keywords *- Cervical Cancer, Herlev Pap Smear dataset, Gray Level Covariance Matrix, Random Forest, Deep-Multilayer Perceptron.*

1. Introduction

One of the most prevalent diseases that impacts all women who receive optimal care if it is diagnosed early is the circumstance of cervical cancer. The death rate in India is relatively high. According to GLOBOCAN 2020, India reports 123,907 occurrences per year and close to 77,348 cancer-related fatalities, accounting for approximately one out of three global cancer incidences [1]. A Pap smear test is commonly used in gynecology to test pre-malignant and malignant tissues in the cervix uteri. Sample can be taken from the cervix, identified using Papanicolaou methods [2], and applied to a microscope slide. The microscopic analysis then identifies the cell construction and embryological abnormalities of cell nuclei. These examinations identify pre-cancerous conditions of cervical tissue if any, and early preventive treatment gets done.

In the manual screening procedure, many images are scrutinized using conventional techniques. The complete method is time-consuming, expensive, and encompasses observer biases. Moreover, false-positive and false-negative diagnostic errors often doubt the manual screening process. A machine learning-based computerized screening system of Pap-smear images can assist cytopathologists in reducing the screening time and observer biases. Standard wet-fixed Parameters are used in the Smear test diagnostic test, including variations in light and color intensity of the cellular components of the coloring operations. In addition, air-drying and rehydration method is used for a smear with excessive blood, mucus, bacteria, and inflammation. All these features together make the identification of apprehensive cells. The primary cells for examining the abnormality are many cells (50 thousand to 3 lakhs on average) on a slide to be examined by a proficient at illustrating the smear as ordinary or maligned. As the categorizing of PAP smear is constructed on well-established cell characteristics, it is desirable to model the automated classifier encompassing all those characteristic features. A direct image-based classifier may overlook some of these critical features, which become considered while classifying the images.

Currently, computational approaches like image processing and classification are greatly interesting in clinical data analysis. The biomedical images and their feature datasets are routinely analyzed using machine learning techniques. Machine learning practices like decision trees, artificial neural networks, convolution neural networks, and support vector machines are used to classify datasets into different pathological classes. Neural network models often suffer from generalization problems, and due to overfitting, they become unstable for field data classification. Considering the above criteria, images were classified into their respective classes based on texture, spatial directional texture, shape, and color. The decision tree-based classifier creates a classifier tree, which helps justify the classification process and explore the feature space to get in-depth knowledge about the classification process. These models become helpful to the cytopathologist to justify the classifications of the Pap smear slides.

This study demonstrates the following goals:

1. To create an automated Pap smear image classification model that utilizes advanced machine learning models so that every Pap-smear picture may be categorized into the correct cytological category.
2. In addition to identifying the optical and first- and greater statistical aspects of the cervical cell pictures, the work aims to designate the images with the corresponding cytological classifications genuinely.
3. The primary objective of this study is to develop a reliable, mechanized Pap smear cervical image categorization framework that is significantly accurate compared to any other traditional test.

2. Related Work

Several scientists and researchers have published scientific papers and journals on cervical cancer classification worldwide. We will recapitulate a glimpse of some of their works below.

[Tang and Foong, 2014] [3] investigated feature learning by expanding random forests to evaluate and rebuild original data implementing the learned future.

[Riana, Widyantoro, and Mengko, 2015] [4] worked with cervical inflammatory cells. They implemented a texture-based method for feature extraction and applied a Decision Tree model for evaluation, which gives a good accuracy result.

[Vens et al., 2011] [5] proposed an unembellished yet functional method to persuade a task-dependent feature illustration utilizing ensembles of random decision trees. The novel feature mapping is effectual in space and time and offers a non-parametric metric transformation; and cannot be articulated through kernel matrix but is flexible for regression and classification problems.

[Nithya et al., 2019] [6] explored numerous feature selection methods to fix the significant attributes in projecting cervical cancer through several model training iterations and ultimately established an optimized feature selection model. They have shown that the proposed model performed the best.

[Nanni et al., 2020] [7] offered a structure grounded on increasing the function by an ensemble-based network model for biological-image categorization datasets of color images using Hervlev Pap smear images. It considered numerous categories of ensembles with diverse structures besides diverse learning parameters.

[Das et al., 2017] [8] presented an intellectual ensemble structure to identify cervical dysplasia depending on contour, surface, and tint structures. The established system was evaluated using two clinical database centers with single and smeared images and the Herlev dataset. They have also contributed a novel segmentation technique for mining shape features.

[William W et al., 2019] [9] developed a tool for cervical cancer that condenses the time needed by rejecting the apparent normal ones to utilize more time on the doubtful images. They applied the Trainable Weka Segmentation model with a sequential elimination approach. Wrapper filters have been used for feature extraction, while the fuzzy C-means procedure has been used for categorization.

[Kyi Pyar Win et al., 2020] [10] developed another method for cervical cancer. They employed digital image analysis of Pap smear images. They used an innovative shape-based iterative model with watershed and random forest techniques. Finally, a bagging ensemble classifier has been used for classification.

[M. A Devi et al., 2016] [11] reviewed and investigated different categories of ANN architecture along with precision results and enactment. A rapid representation and recognition of cervical cancer presented to classify normal and diseased cells.

[Song Y et al., 2019] [12] presented a technique for cutting the contour of images with several overlapping cytoplasms into multiple contour fragments and then reconstructing for every cytoplasm and refining segmentation outcomes. This process has been applied to two unique cervical smear datasets and proved very effective.

3. Materials and Methods

3.1. Dataset Details

The model has been designed based on a meticulously collected and classified Herlev dataset [13]. The dataset comprises single-nucleus cervical cells of seven different

histological classes and provides a good platform for a practical machine-learning exercise.

3.2. Texture-based features Extraction Model

3.2.1. GLCM (Gray Level Co-Occurrence Matrix)

In this scheme [14], the size of picture pixels corresponds to the amount of gray-level components. The second data collection probability values for variations between grey levels "m" and "n" at a specific range (d) and an exact orientation (θ) are contained in the matrix component "$b(m,n).$" Utilizing GLCM, texture-based characteristics were retrieved. Energy, contrast, homogeneity, and entropy, four qualities, have been gathered in four orthogonal directions with an interval d. Using a unit displacement d, thorough features were extracted. The following equations show the mean variation of these criteria, which offers a continuous change of these features with the labelled classes.

$$Energy = \sum_{m,n}^{Ng}\{b(m,n)\}^{2} \quad (1)$$

$$Entropy = -\sum_{m,m}^{Ng} b(m,n)\, log\, b(m,n) \quad (2)$$

$$f_{contrast} = \sum_{m,n}^{Ng}(m-n)^2 * b(m,n) \quad (3)$$

$$f_{homogeneity} = \sum_{m,n}^{Ng} \frac{b(m,n)}{1+|m-n|} \quad (4)$$

Where $b(m,n)$ represents the GLCM, Ng = No of distinct gray levels in the image.

3.2.2. Gabor Bank-Based Feature Extraction

Texture feature extraction denotes the method of figuring characteristics of an image numerically to develop some quantifiable figures used to categorize the image. Gabor filters [15] are an example of this feature extraction method and proved to be a worthy model as it can capture the total frequency range in every direction. The Gabor filter [16] is a Gaussian distribution that can operate in both the frequency and spatial domains and is regulated by a compound bandwidth and direction. The ability to provide apparent discrepancies of different textures made Gabor Filter popular. Here a 2-D Gabor filter is applied to collect features from three-dimensional medical images. The Equation is represented as follows:

$$G(x,y,\omega,\theta,\sigma_x,\sigma_y) \equiv \frac{1}{2\,\pi\sigma_x\sigma_y} e^{-\frac{(x)^2}{2\sigma_x^2}-\frac{(y)^2}{2\sigma_y^2}} e^{j\omega(xcos\theta+ysin\theta)} \quad (5)$$

Where σ is the spatial spread, ω is the frequency, θ is the orientation.

3.3. Geometric shape features

Invariant shape descriptors [17] of cell objects refer to precise physiognomies concerning the geometry of Pap smear slide images. They form potent feature vectors indicating cytological stages of cervical cells. In this experiment, critical features like area, regularity of the boundary, the eccentricity of the nucleus, and cytoplasm are considered for the cervical cells into different cytological stages. The area and boundary smoothness are two crucial criteria for the visual classification process. The eccentricity between the principal and minor axes was examined to determine the circularity of the cytoplasm and the nucleus. The ratio of these two objects has also been considered a classification feature to compare the sizes of cell nuclei and cytoplasm. These shape-based features are invariant and regarded as reliable classifiers in a controlled environment.

3.4. Color Features

Stained Pap smear images are relatively easy to classify. Staining makes the cytoplasm of superficial cells eosinophilic. These make squamous and intermediate squamous cells basophilic. Lastly, Cervical Squamous Cancer cells generally become bluish-green. In the case of the presence of keratin in the cervical epithelium, some cells also become orange. As keratin in the cervix region is abnormal, this indicates an alarming pathological condition. Cytoplasm and colors of the cell nucleus have been collected separately and converted to an index image. The highest number of color index values within an index image becomes the object's predominant color value. The normalized RGB color value becomes used for identifying the dominant color value from the nucleus's color map and cytoplasm of a cell image. This study uses a linear interpolation scheme to map the principal RGB value to a valid RGB value for determining an RGB color code from the color map table. To better represent the color scale, we also considered the dynamic range of the color distribution.

3.5. The Classifiers

3.5.1. Random Forest (RF)

Random forest [18] was constructed from several classification trees during its training period. Every tree yields a class, after which the forest selects the class with the maximum output votesAll trees in the forest are supported by the elements of a randomly generated vector that was separately collected and had an equitable distribution across all trees. Leo Breiman [19] implemented the bagging method to diminish variance and circumvent overfitting. Ho pooled Breiman's idea of bagging plus random feature selection and created a set of decision trees for resolving overfitting issues. Andy Liaw and Matthew Wiener [20] emphasized the benefits of implementing random forest as the instrument for classification. Random forest requires neither any prototype nor model as allusion, making the learning method uncomplicated and producing tremendous predictors.

3.5.2. Support Vector Machine (SVM)

In SVM [21], a hyperplane is produced, which separates the feature space into two disjointed areas to create a

maximum boundary between two classes. This hyperplane sketches the class margin allowing new features to get classified. Hyperplanes or clusters of hyperplanes get constructed by SVM higher dimension spaces. Finally, the hyperplane with the maximum margin is considered. Some advantages of SVM are that it can handle an extensive database, is simple to use and understand, is fast, and models are small in size.

Suppose there are 'm' training samples (x_i, y_i), where $x_i \in R_N$ and y_i is the corresponding label ($yi \in \{-1, 1\}$). The best hyperplane that divides the pieces of data adequately and optimises the separation of any category from the hyperplane is found by the SVM model. Calculating the preeminent hyperplane is modelled as an optimization problem with constraints and resolved by expanding quadratic programming methods. The discriminant hyperplane is well-defined through the level set as shown in Eq. (6):

$$f(x) \equiv \sum_{i=1}^{\mathrm{m}} y_i \alpha_i . k(x, x_i) + b \quad (6)$$

Where k = kernel function and the sign of $f(x)$ regulates the association of x.

3.5.3. Multilayer Perceptron (MLP)

The MLP [22-24] is a popular learning algorithm. It is a feed-forward artificial neural network comprising a directed graph with multiple layers of nodes. The MLP network is depicted in Fig.5 below. It can also use multiple hidden layers. An MLP classification network comprising one hidden layer with N hidden nodes is well-defined by two weight matrices, $W \in R_N X_d$ and $V \in R_N X_1$, an activation function $f(\mathbf{x})$ which operates element-wise, the step-function F(**x**) with the equation as in Eq. (7).

$$\hat{y} = \mathrm{F}\big(\mathrm{V}^T f(Wx)\big) \quad (7)$$

Since there are so many criteria that need to be modified, MLP learning takes a long time than it does for other predictors. The main parameters to tune are the number of hidden nodes, learning rate, weight decay, and momentum.

3.6. Resampling Technique: Cross Validation

A cross-validation is a resampling approach used to construct machine learning algorithms when data is insufficient. The technique lies on a parameter termed k, which denotes the number of splits of equal size applied to given data. The overall method is as follows:

- Randomly Shuffled the input dataset.
- Then, the given data is split into k groups
- The following steps are repeated for a distinctive individual group.
- The proficiency of the model has been summarized by expanding the different model evaluation scores k is usually selected as 5 or 10, but there is no strict rule. For experimental purposes using 5-fold cross-validation assessment was applied.

4. Proposed Method

Figure 1 depicts the process flow of the suggested approach. According to Fig. 1, in the first step, the images are read from the Herlev Pap smear dataset and sent for pre-processing. In this stage, before feeding the input feature vectors, the sample dataset has been filtered to distinguish the outliers depending on the population's standard deviation.

4.1. Implementation Details

The steps involved in the experiment are described below:

4.1.1. Pre-Processing

Automatization of the classification of Pap smear images requires specific preprocessing to get them ready for high-level analysis. For feature extraction, it is necessary to segment the images into the nucleus and cytoplasm. The segmentation, in turn, requires image enhancement for noise and artefact reduction. The nucleus area of any cervical cell usually has greater duskier pixel dispersal than the cytoplasm region. Input images are first binarized and trailed by morphological closing operation using a structuring element of size 3×3. Next, a morphological filling operation gets performed. Following that, the obtained core picture is subtracted from the initial image. The detracted image is then transformed to LAB color space and classified into three clusters for cytoplasm, RBC, and background, implementing K-means clustering with Euclidean distance metric. The mean of the red band of all three clusters obtained from the original image is estimated, and the cluster has the least mean taken as a cytoplasm cluster [25]. In this experiment, the original RGB images and segmented nucleus and cytoplasm images were to understand the feature space better and considered texture (directional), shape, and color-based features as extraction mechanisms based on understanding.

4.1.2. Feature Extract Process

The Herlev image database comprises diverse labelled images with distinct visual features like shape, color, and texture. The proposed model extracted and stored a feature extraction matrix corresponding to an individual image and trained a model with this data. Features are extracted when any new image comes, and the query image features matrix is fed into the model to get its corresponding label.

The gray level co-occurrence matrix (GLCM) technique is the most frequently used process for texture features. Since Gabor filters are dominant for use in border areas of an image, clubbing Gabor features and GLCM features are experimented with. Here further analyzed, the feature set with Linear Discriminant Analysis (LDA) for dimensionality reduction. Finally, the reduced feature set is used to model a

classifier to classify the images into their respective seven groups. The shape of cell objects plays an essential role in Pap smear image classification. For study purposes, invariant shape-based features like area, circularity, perimeter, and eccentricity of the nucleus and cytoplasm are considered. Since color scheme schemes are important in the classification phase, one of the most dominating and expressive varieties of a nucleus and cytoplasm color schemes were evaluated as characteristics for image categorization. Sixteen features from texture, shape, and color to represent 917 images were extracted from images. To identify the role of different features, the proposed study moves to the average of all the features over the seven categorical classes of the test cases. In all cases, their rolling averages show an inevitable trend from class label 1 to class label 7. They indicate that capturing these feature predicates in a proper order may precisely classify all the samples into their respective classes.

4.1.3. Distribution of dataset using K-Means

K-Means [26] successfully divides entities into clusters that share similarities but vary from entities in other clusters. It requires selecting K random positions as cluster centers known as centroids. In this problem, the selected value of K is10 obtained through the trial-and-error method, although there is a rule of thumb to choose the suitable value of clusters shown in Eq. (8):

$$K = \frac{\sqrt{N}}{2} \tag{8}$$

4.1.4. Classification using Proposed Hybrid Random Forest

The proposed hybrid model consists of a central image database connected to k-means and Random Forest (RF) models in succession. The unsupervised k-means classifier initiates the classification process and creates the Random Forest [27] model training, validation, and testing datasets. The algorithms of the proposed model illustrate in the algorithm section.

4.2. Algorithm

The algorithm of the proposed model is given below:

1. Read sample Pap smear slide images
2. Enhance and filter the image dataset
3. Perform GLCM feature selection to estimate texture-based features
4. Perform Gabor feature selection to estimate spatial features
5. Estimate the invariant morphological operation to extract shape-based features
6. Estimate the color features to extract color-based features
7. Combine all the features to form a global feature set of all the slide images.
8. Partition the sample space of each class of Pap smear image dataset using k-means clusters
9. Prepare (i) training, (ii) validation, and (iii) model testing datasets with training (70%), validation (15%), and testing (15%) datasets from each k-means cluster of each image class.
10. Design a deep-learning Random Forest model with proper network parameters.
11. Evaluate the image class of the test samples and compare the predicted value with the actual class level of the model verification dataset and performance evaluation done
12. Designed an SVM model with three different kernels: Linear, Quadratic and Gaussian, and a Multilayer Perceptron model with one and three Hidden Layers and compared with our Hybrid Random Forest model.

5. Experimental Details and Discussions

5.1. Experimental Results

The investigation has been carried out using the Herlev Pap Smear image dataset using different classifiers [28,29] with distinct feature sets such as Color, Texture, and geometric shape features [31-35] in Table1 and the classification accuracy compared. The experiment has been carried out on the Herlev Pap Smear dataset, having both Healthy and diseased images implementing different classifiers and matched grounded on the number of features.

Table 1. Features Used

Serial No	Feature Category	Feature Algorithms	Features
1	Texture	GLCM	Energy, Entropy, Homogeneity, Contrast
		Gabor	Gabor Magnitude
2	Color		Color of the cell nucleus, Color of the cell cytoplasm, Dynamic range of colors
3	Shape Feature		Nucleus area, Cytoplasm Area, Regularity of boundary-nucleus, Regularity of boundary-cytoplasm, Eccentricity of the nucleus, Eccentricity of cytoplasm, Nucleus-cytoplasm size proportion

The study of all the classifiers has been compared based on the number of features extracted. The following feature sets used for comparison are in Table 2 below.

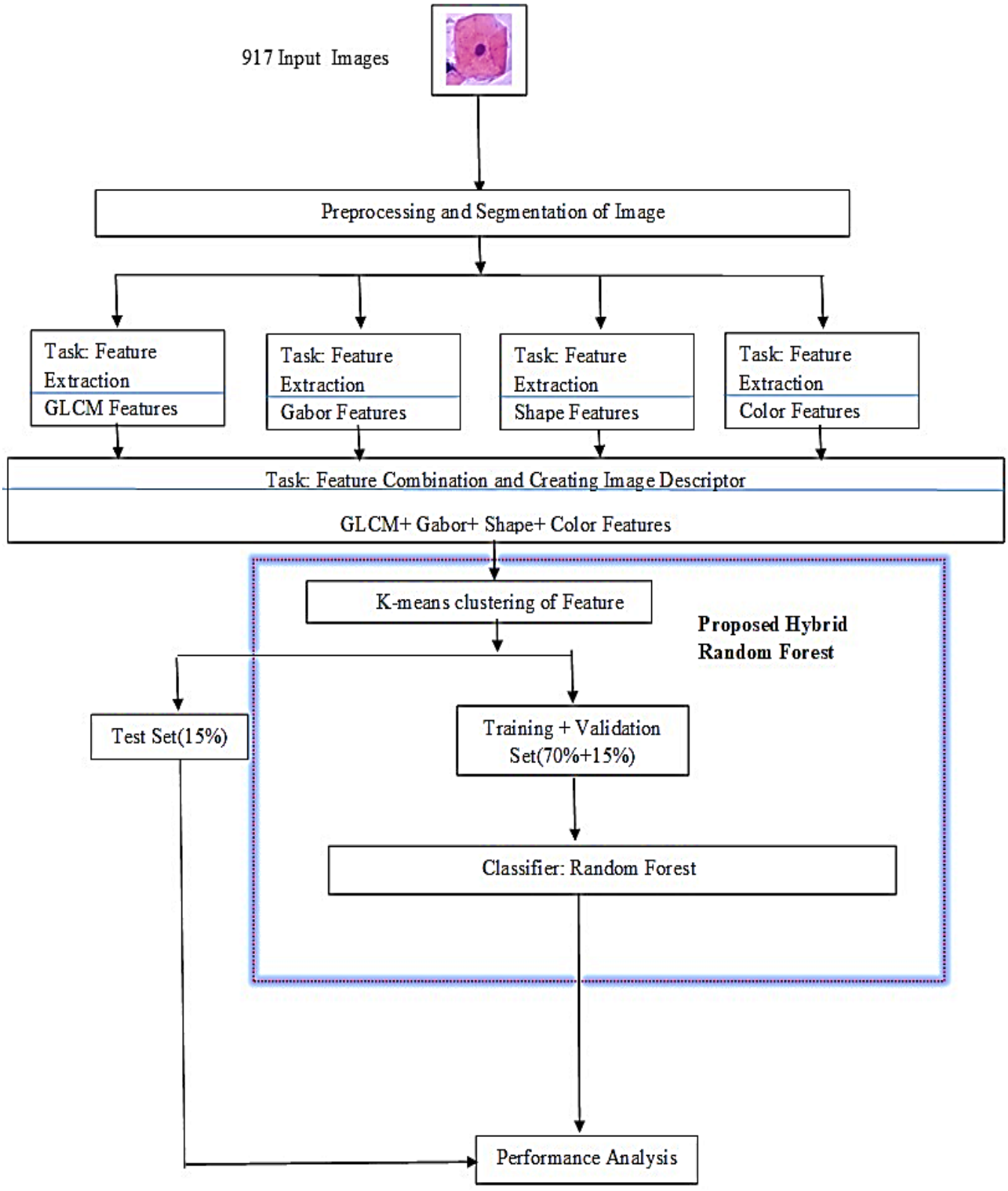

Fig 1. Proposed Hybrid Classification Model Work Flow Diagram

Table 2. Feature Sets used

Feature Set	Feature Set Type	Feature Standard Algorithm	Features
FS1	Color + Shape Feature		Color of the cell nucleus, Color of the cell cytoplasm, Dynamic range of colors, Nucleus area, Cytoplasm Area, Regularity of boundary-nucleus, Regularity of boundary-cytoplasm, Eccentricity of a nucleus, Eccentricity of cytoplasm, Nucleus-cytoplasm size proportion
FS2	Texture	GLCM +Gabor	Energy, Entropy, Homogeneity, Contrast, Gabor Magnitude
FS3	FS1 + FS2		Combining all the above 15 features

The proposed Hybrid Random Forest model's performance evaluation is compared with the Deep-learning Multilayer Perceptron (MLP) model and Support Vector Machine (SVM). Here an average of 10-fold cross-validation data for all cases were considered.

A hybrid of K-means clustered feature set and Random Forest classification algorithm with the following parameters have been used. The number of trees in the forest has been taken as 50 with the Gini impurity criterion with a maximum tree depth.

While creating the SVM classifier, defining an explicit kernel function as an essential learning parameter is indispensable. For this experiment, three types of SVM were considered: Linear, Quadratic, and Gauss SVM.

The deep-learning MLP model has been configured with a ReLU activation function having 1 and 3 hidden layers. Fully connected, each hidden layer has been designed with ten nodes and (10,10,10) nodes, respectively. A softmax function maps the multinomial distribution in the final layer to model the categorical response in the outcome. MLP models are tested with different numbers of epochs, learning rates, and early stopping with validation to achieve the maximum level of generalization. For experimental purposes, the deep-learning neural networks were trained with training datasets and confirmed the best fit model with the validation dataset.

The first step of this experiment was to consider seven shapes and three colors features and feed them into the six classifiers as all features may not yield better performance to the classifiers, so ten features to evaluate the classifier comportment with the help of 8 performance evaluators. Table 3 and Figure 2 and 3 below shows that Random Forest performs the best.

Next, we performed our experiment and showed classification evaluation based on five texture features, indicating that Random Forest serves well than other classifiers. Table 6 and Fig 4 and 5 illustrate it.

Table 3. Performance evaluator with shape and color features

Measuring Parameter	Proposed Hybrid Random Forest	SVM linear	SVM Quadratic	SVM Gauss	MLP- 1 layer	MLP- 3 layer
Accuracy	0.9854	0.9635	0.9562	0.9051	0.9270	0.9635
Error	0.0146	0.03649	0.0438	0.0949	0.0730	0.0365
Sensitivity	0.9795	0.96852	0.9643	0.9005	0.9070	0.973
Specificity	0.9976	0.99373	0.9922	0.9836	0.9878	0.9936
Precision	0.9766	0.96093	0.9682	0.9055	0.9093	0.9710
False Positive Rate	0.0024	0.00627	0.0078	0.0163	0.0122	0.0064
F1_score	0.9770	0.96355	0.9658	0.9023	0.9041	0.9711
MCC	0.9753	0.95798	0.9584	0.8865	0.8945	0.9653

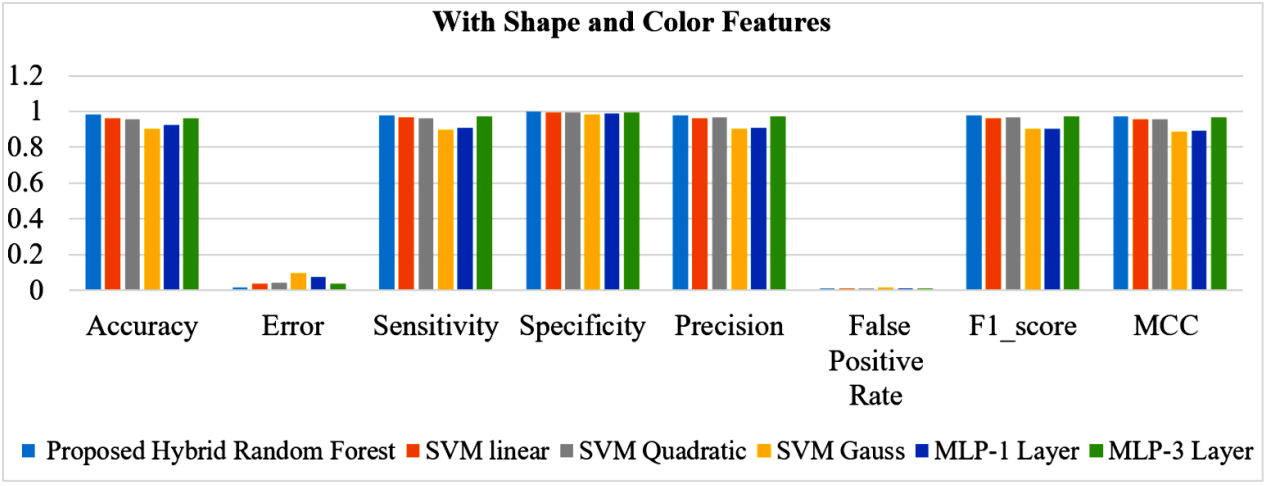

Fig. 2 Performance evaluator with shape and color features

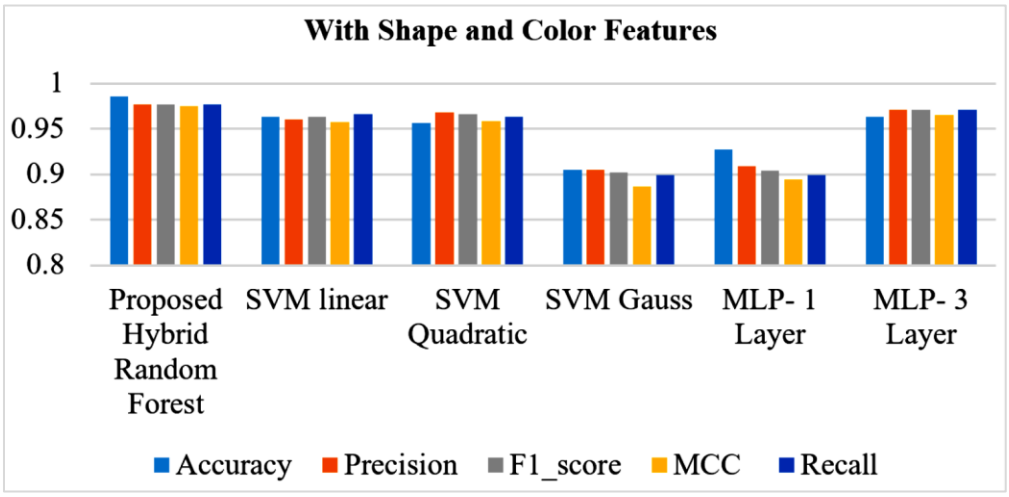

Fig. 3 Classifier assessment with shape and color features

Table 4. Performance evaluator with texture feature

Measuring Parameter	Proposed Hybrid Random Forest	SVM linear	SVM Quadratic	SVM Gauss	MLP- 1 layer	MLP- 3 layer
Accuracy	0.9708	0.9635	0.9416	0.9124	0.9343	0.9416
Error	0.0292	0.0365	0.0584	0.0876	0.0657	0.0584
Sensitivity	0.9743	0.9591	0.9446	0.9215	0.9206	0.9372
Specificity	0.9953	0.993	0.9902	0.9846	0.9885	0.9901
Precision	0.9558	0.9772	0.9364	0.9164	0.9354	0.9208
False Positive Rate	0.0047	0.007	0.0098	0.0154	0.0115	0.0099
F1_score	0.9632	0.9671	0.9372	0.9182	0.9263	0.9275
MCC	0.9594	0.9613	0.9294	0.9033	0.9159	0.9184

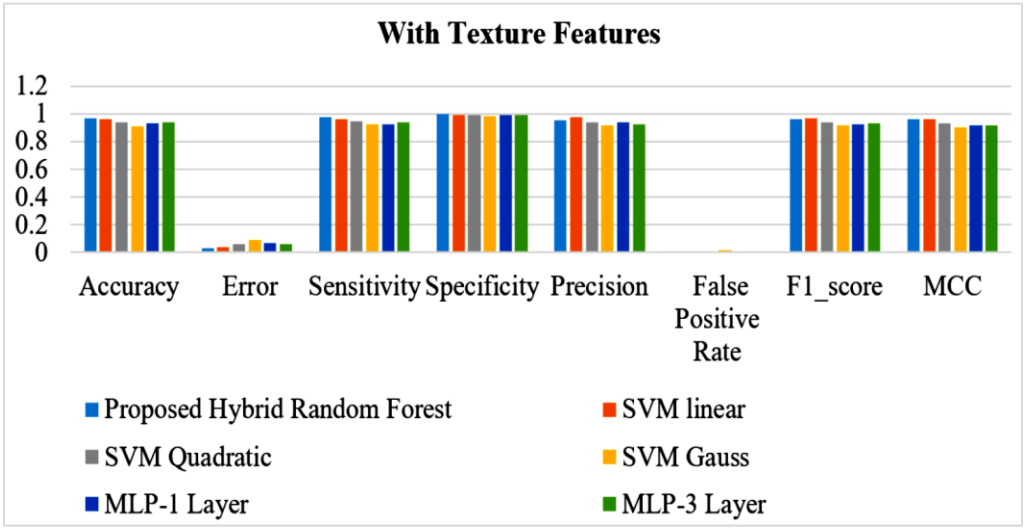

Fig. 4 Performance evaluator with texture features.

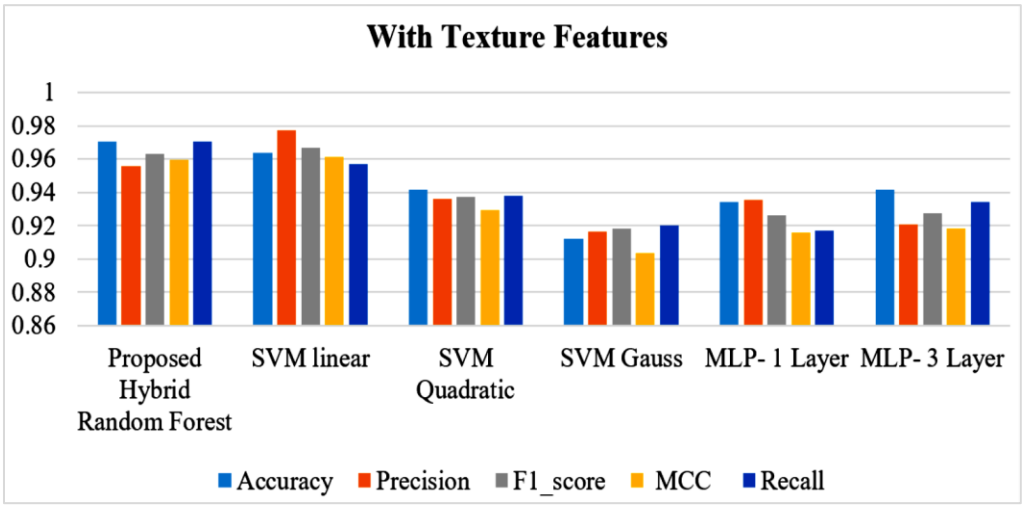

Fig. 5 Classifier assessment with texture features.

Finally, we experimented with five texture features, three color features, and seven shape-related features and fed them into all six classifiers. This time also, we found that Random Forest performed the best. Table 5 and Fig 6 and 7 below demonstrate it.

Table 5. Performance evaluator with all shape, color, and texture features

	Proposed Hybrid Random Forest	SVM linear	SVM Quadratic	SVM Gauss	MLP- 1 layer	MLP- 3 layer
Accuracy	0.9927	0.9635	0.9781	0.9124	0.9562	0.9416
Error	0.0073	0.0365	0.0219	0.0876	0.0438	0.0584
Sensitivity	0.989	0.9725	0.9744	0.9264	0.945	0.9418
Specificity	0.9987	0.994	0.9961	0.9837	0.9924	0.9901
Precision	0.9947	0.9673	0.9831	0.9423	0.9583	0.9427
False Positive Rate	0.0013	0.006	0.0039	0.0163	0.0076	0.0099
F1_score	0.9916	0.968	0.9777	0.9308	0.9499	0.9385
MCC	0.9905	0.9629	0.9745	0.9172	0.9435	0.9308

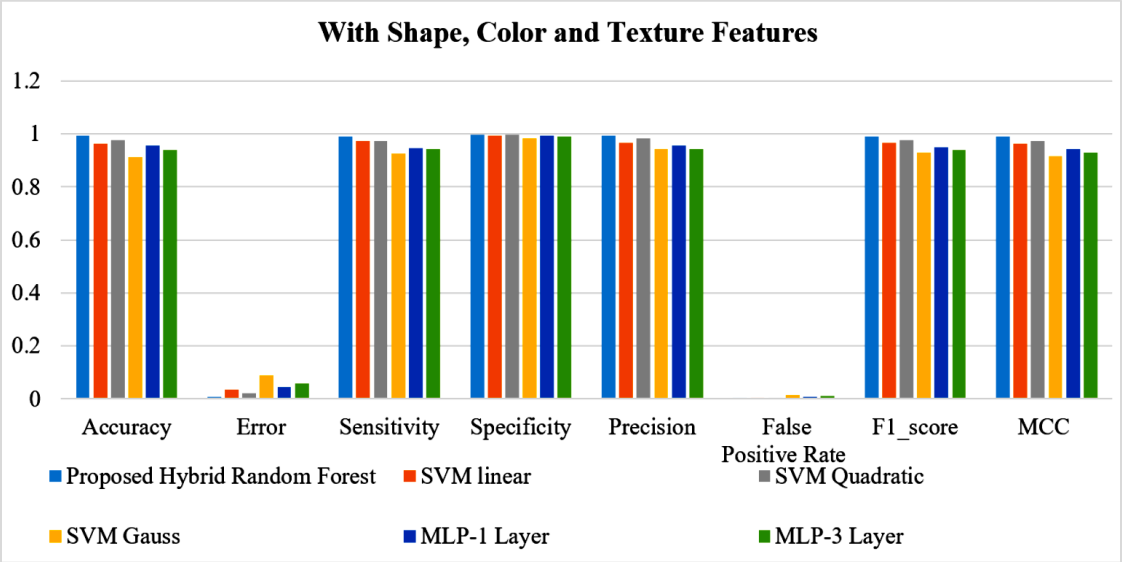

Fig. 6 Performance evaluator with all 15 shapes, color, and texture features

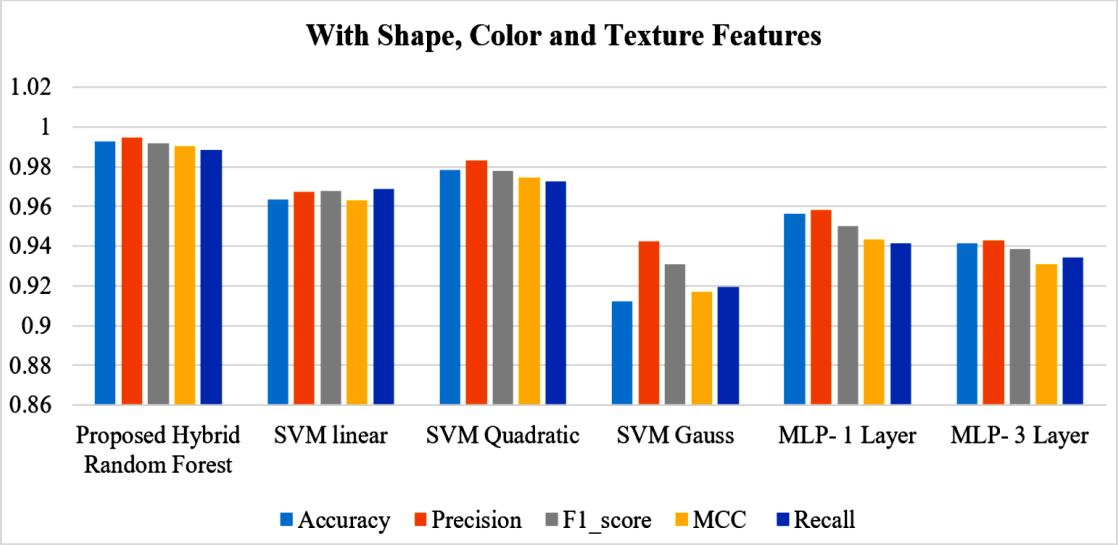

Fig. 7 Classifier assessment with all 15 shapes, colors, and texture features

From Table 6 and its graphical representation in Fig 8, it is apparent that Random Forest is naturally ahead of the other five models, and combining all the features of shape, color, and texture gives the maximum accuracy.

Table 6. Accuracy evaluator of the different classifiers based on all 15 features

	Proposed Hybrid Random Forest	SVM linear	SVM Quadratic	SVM Gauss	MLP- 1 layer	MLP- 3 layer
With shape and color features (FS1)	0.9854	0.9635	0.9562	0.9051	0.9270	0.9635
With texture features (FS2)	0.9708	0.9635	0.9416	0.9124	0.9343	0.9416
With shape, color, and texture features (FS3)	0.9927	0.9635	0.9781	0.9124	0.9562	0.9416

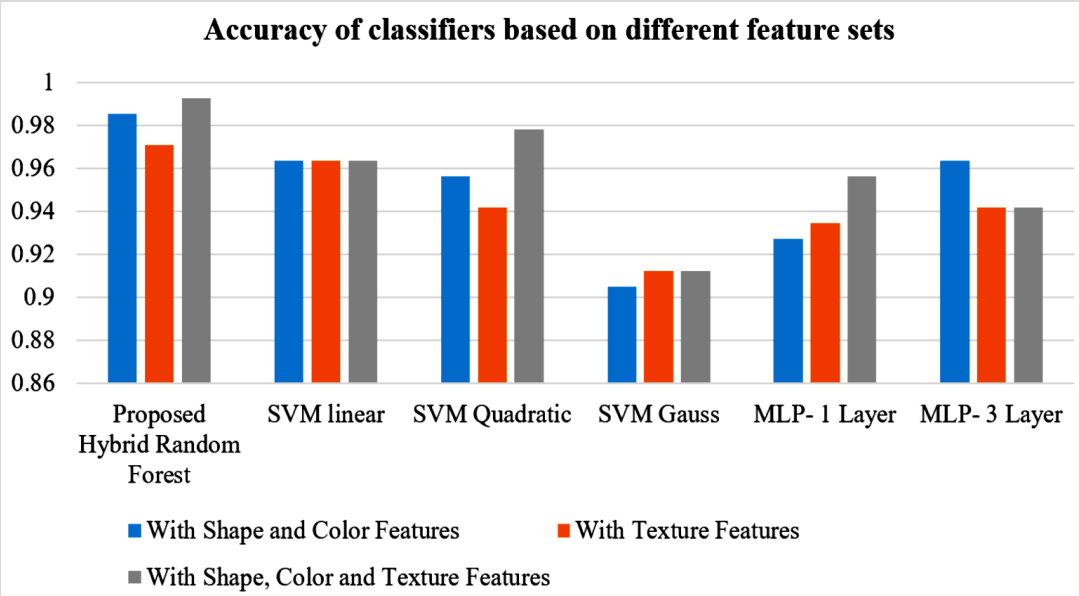

Fig. 8 Accuracy of classifiers based on all 15 features

5.2. Discussions

The suggested hybrid approach of the random forest technique was found from the aforementioned experimental study to have the maximum accuracy. The suggested model offers greater than 97 percent accuracy for all feature-collection models,

It demonstrates that the anticipated study on solitary cervical cell pictures performed extremely well. The suggested approach will be used for photos with numerous cells in the future. This novel strategy works well in the classification process, and this technique will certainly perform well enough on pictures through a bulky amount of cells.

6. Conclusion

The suggested work provided an involuntary classification technique of Pap smear slide images depending on selectively identified features. Texture, Shape, and Color are the three types of features identified to represent all the images into their respective cytological classes. The hybrid Random Forest model has been designed with a K-means classifier to make clusters of each class label. The purpose was to prepare experimental datasets with the best possible representation of all feature values within the training, validation, and testing datasets. This nonlinear supervised hybrid model provides a generalized classifier and delivers the logical analysis of the classification steps. Thus, it helps the cytopathologist verify the classification results and even justify their judgments concerning the logical analysis of the decision tree. This logical analysis of the Random Forest decision trees also provides a deep insight into the highly subjective manual classification process to identify the most promising feature set and their order to classify the Pap smear images. The approach also reduces the possibility of errors arising due to incompleteness and inconsistency in the visual slide image classification process. The method is robust and fast as it is repeatable under various model design constraints. The model has also shown a clinically acceptable solution for a test dataset.

References

[1] H. Sung et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” CA: A Cancer Journal for Clinicians.

[2] E.K.W. Schulte,“Standardization of Biological Dyes and Stains: Pitfalls and Possibilities,” *Histochemistry*, vol. 95, no. 4, pp. 319–328, 1991.

[3] A. Tang,Foong, and J.T, “A Qualitative Evaluation of Random Forest Feature Learning,” *Recent Advances on Soft Computing and Data Mining, Springer, Cham*, pp. 359-368, 2014.

[4] D.Riana, D. H. Widyantoro, and T. L. Mengko, “Extraction and Classification Texture of Inflammatory Cells and Nuclei in Normal Pap Smear Images,” *4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, pp. 65-69, 2015.

[5] C. Vens and F. Costa, “Random Forest Based Feature Induction,” *IEEE 11th International Conference on Data Mining*, pp. 744-753, 2011.

[6] B.Nithya, V.Ilango, “Evaluation of Machine Learning Based Optimized Feature Selection Approaches and Classification Methods for Cervical Cancer Prediction,” *SN Applied Sciences*, vol. 1, no. 641, 2019.

[7] L. Nanni, S. Ghidoni, S. Brahnam, "Ensemble of Convolutional Neural Networks for Bioimage Classification," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 19-35, 2020.

[8] Bora et al., "Automated Classification of Pap Smear Images to Detect Cervical Dysplasia," *Computer Methods and Programs in Biomedicine*, vol 138, pp. 31-47, 2017.

[9] W. William et al., "A Pap-Smear Analysis Tool (PAT) for Detection of Cervical Cancer from Pap-Smear Images," *BioMedical Engineering OnLine*, vol. 18, 2019.

[10] Kyi Pyar Win et al., "Computer-Assisted Screening for Cervical Cancer using Digital Image Processing of Pap smear Images," *Applied Sciences*, 2020.

[11] M.A Devi et al., "Classification of Cervical Cancer Using Artificial Neural Networks," *Procedia Computer Science*, vol. 89, pp. 465-472, 2016.

[12] Y Songet al., "Segmentation of Overlapping Cytoplasm in Cervical Smear Images via Adaptive Shape Priors Extracted from Contour Fragments," *IEEE Transactions on Medical Imaging*, vol. 38, no. 12, pp. 2849-2862, 2019.

[13] MDE-Lab : The Management and Decision Engineering Laboratory, 2022. [Online]. Available: http://mde-lab.aegean.gr/

[14] R.E. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, 1973.

[15] Y Hamamoto et al., "A Gabor Filter-Based Method for Recognizing Handwritten Numberals," *Pattern Recognition*, vol. 31, no. 4, pp. 395-400, 1998.

[16] Jain A. K, F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters," *Pattern Recognition*, vol. 24, no. 12, pp. 1167-1186, 1991.

[17] Cruz Jerome et al., "Object Recognition and Detection by Shape and Color Pattern Recognition Utilizing Artificial Neural Networks," *IEEE*, pp. 140-144, 2013.

[18] See Yuen Chark and Norliza Mohd Noor, "Integrating Complete Gabor Filter to the Random Forest Classification Algorithm for Face Recognition," *Journal of Engineering Science and Technology*, vol. 14, no. 2, pp. 859-874. 2019.

[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[20] A Liaw,and M. Wiener, "Classification and Regression by Random Forest," *News - Reddit*, vol. 2, no. 3, pp. 18-22, 2002.

[21] Srivastava, Durgesh & Bhambhu, Lekha, "Data Classification using Support Vector Machine," *Journal of Theoretical and Applied Information Technology*, vol. 12, pp. 1-7, 2010.

[22] Mahmudul et al., "An Algorithm for Training Multilayer Perceptron (MLP) for Image Reconstruction using Neural Network without Overfitting," *International Journal of Scientific & Technology Research*, vol. 4, no. 2, pp. 271-275, 2015.

[23] Laurene Fausett, "Fundamentals of Neural Networks: Architectures, Algorithms, and Applications," *Pearson Education*, 2008.

[24] Dr. Surendiran R, Dr. Thangamani M, Monisha S, Rajesh P, "Exploring the Cervical Cancer Prediction by Machine Learning and Deep Learning with Artificial Intelligence Approaches," *International Journal of Engineering Trends and Technology*, vol. 70, no. 7, pp. 94-107, 2022. Crossref, https://doi.org/10.14445/22315381/IJETT-V70I7P211.

[25] M. M. Puranik, S.V.Halse, "A Review Paper: Study of Various Types of Noises in Digital Images," *International Journal of Engineering Trends and Technology*, vol. 57, no. 1, pp. 40-43, 2018. Crossref, https://doi.org/10.14445/22315381/IJETT-V57P208

[26] O. Sarrafzadeh, andA. Dehnavi, "Nucleus and Cytoplasm Segmentation in Microscopic Images using K-Means Clustering and Region Growing," *Advance Biomedical Research*, vol. 4, no. 174, 2015.

[27] Plissiti, Marina E., and Christophoros Nikou, "A Review of Automated Techniques for Cervical Cell Image Analysis and Classification," *Biomedical Imaging and Computational Modelling in Biomechanics, Springer Netherlands*, pp. 1-18, 2013.

[28] Breiman et al., "Classification and Regression Trees," *Routledge*, 2017.

[29] E. Martin, "*Pap-Smear Classification*," Master's thesis, Technical University of Denmark:-DTU, 2003.

[30] V .P. Amadi, N.D Nwiabu, V. I. E. Anireh, "Case-Based Reasoning System for the Diagnosis and Treatment of Breast, Cervical and Prostate Cancer," *SSRG International Journal of Computer Science and Engineering*, vol. 8, no. 8, pp. 13-20, 2021. *Crossref*, https://doi.org/10.14445/23488387/IJCSE-V8I8P103

[31] S. Banerjee and D. Dutta Majumdar, "A 2D Shape Metric and its Implementation in Biomedical Imaging," *Pattern Recognition Letters*, vol. 17, no. 2, pp. 141–147, 1996.

[32] S. Parui, E. Sarma, and D. Majumder, "Studies on Some Multimodal Medical Image Registration Approaches for Diagnostic and Therapeutic Planning: with Some Case Studies," *Pattern Recognition Letters*, vol. 4, pp. 201-204, 1986.

[33] D. Dutta Majumder, and D. Ray, "Approaches of Multimodal Medical Images Registration and Fusion: Efficacy on Diagnostic and Therapeutic Planning," *IETE Journal of Research*, vol. 57, no. 6, pp. 498-514, 2011.

[34] W. Lisheng, Q. Gan, and T. Ji, "Cervical Cancer Histology Image Identification Method Based on Texture and Lesion Area Features," *Computer-Assisted Surgery Abingdon*, vol. 22, pp. 186-199. 2017.

[35] S. Bhattacharjee, Y.J Singh, and D. Ray, "Comparative Performance Analysis of Machine Learning Classifiers on Ovarian Cancer Dataset," *Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE*, pp. 213-218, 2017.