Received: 22 May 2020 | Revised: 18 August 2020 | Accepted: 21 August 2020
DOI: 10.1002/cnr2.1293

ORIGINAL ARTICLE

Cancer Reports WILEY

Study of morphological and textural features for classification of oral squamous cell carcinoma by traditional machine learning techniques

Tabassum Yesmin Rahman[1] | Lipi B. Mahanta[2] | Hiten Choudhury[1] | Anup K. Das[3] | Jagannath D. Sarma[4]

[1]Department of Computer Science & IT, Cotton University, Guwahati, India
[2]Mathematical and Computational Sciences Division, Institute of Advanced Study in Science and Technology, Guwahati, India
[3]Pathology, Arya Wellness Centre, Guwahati, India
[4]Pathology, Dr B. Borooah Cancer Institute, Guwahati, India

Correspondence
Lipi B. Mahanta, Mathematical and Computational Sciences Division, Institute of Advanced Study in Science and Technology, Guwahati, Assam 781036, India.
Email: lbmahanta@iasst.gov.in
Tabassum Yesmin Rahman, Department of Computer Science & IT, Cotton University, Panbazar, Guwahati, Assam 781001, India.
Email: yesmin.a15@gmail.com

Abstract

Background: Oral squamous cell carcinoma (OSCC) is the most prevalent form of oral cancer. Very few researches have been carried out for the automatic diagnosis of OSCC using artificial intelligence techniques. Though biopsy is the ultimate test for cancer diagnosis, analyzing a biopsy report is a very much challenging task. To develop computer-assisted software that will diagnose cancerous cells automatically is very important and also a major need of the hour.

Aim: To identify OSCC based on morphological and textural features of hand-cropped cell nuclei by traditional machine learning methods.

Methods: In this study, a structure for semi-automated detection and classification of oral cancer from microscopic biopsy images of OSCC, using clinically significant and biologically interpretable morphological and textural features, are examined and proposed. Forty biopsy slides were used for the study from which a total of 452 hand-cropped cell nuclei has been considered for morphological and textural feature extraction and further analysis. After making a comparative analysis of commonly used methods in the segmentation technique, a combined technique is proposed. Our proposed methodology achieves the best segmentation of the nuclei. Henceforth the features extracted were fed into five classifiers, support vector machine, logistic regression, linear discriminant, *k*-nearest neighbors and decision tree classifier. Classifiers were also analyzed by training time. Another contribution of the study is a large indigenous cell level dataset of OSCC biopsy images.

Results: We achieved 99.78% accuracy applying decision tree classifier in classifying OSCC using morphological and textural features.

Conclusion: It is found that both morphological and textural features play a very important role in OSCC diagnosis. It is hoped that this type of framework will help the clinicians/pathologists in OSCC diagnosis.

KEYWORDS
cell level, cell nuclei, ML, morphological features, oral cancer, OSCC, textural features

1 | INTRODUCTION

Any cancerous tissue growth situated in the oral cavity is known as oral cancer and is a type of head and neck cancer. Oral cancer is very commonly occurring cancer worldwide, especially in India. Besides various types of oral cancer, squamous cell carcinoma (SCC) is the commonest one.[1] There are three grades of oral squamous cell carcinoma (OSCC): well-differentiated SCC, moderately differentiated SCC, and poorly differentiated SCC. This type of cancer (OSCC) starts at the epithelial layer of the oral cavity. This layer consists of a layer of cells in a definite manner. When a cell is affected by cancer, not only the size and the shape of the cell but also the nucleus gets altered. Generally, the size of the nucleus becomes larger when affected by the disease.

A pathologist usually diagnoses a patient with cancer by observing his or her biopsy slides. A biopsy is a gold standard test for cancer diagnosis in which a small number of cells or tissues are extracted from the cancerous region of the patient. Many types of biopsies such as excisional, incisional, punch, fine needle, can be conducted to determine the presence and extent of cancer development. An incisional biopsy involves removing a small portion of the tissue from the suspicious region, whereas excisional biopsy involves removing the entire suspicious region or the whole organ where cancer takes place. From these samples, clinician prepares hematoxyline and eosin (H&E) stained slides. Pathologists then observe these slides under a microscope. But, this traditional method has its drawbacks as it is very tedious work, requiring a lot of experience and takes more time. It is also possible to get an erroneous report sometimes, even by an experienced pathologist. Visual assessments can be subject to inappropriate inter and intraobserver variations.[2] Hence, the development of computer-assisted software, which will diagnose cancerous cells automatically, is both pertinent and significant.

The main objective of this study is to develop a computer-assisted system, which will differentiate benign cell nuclei from malignant cell nuclei of OSCC based on morphological and textural features. Studies on nuclear size, shape, and texture analysis of biopsy images of OSCC are very limited. Hence, we have undertaken this study intending to compare these features of benign and the malignant cells of OSCC. Here, we have used a combined/hybrid segmentation technique to segment out the nucleus, which gives us a very good result in distinguishing benign cells from malignant cells. Researches on other domains of image processing have revealed adequately that the combination of correct methods most often leads to a stable approach in the extraction of meaningful information from the images.[3]

While studies on the use of artificial intelligence (AI) in the development of efficient algorithms have been profound in almost all areas of cancer, we present here a comprehensive summary of studies directed on oral cancer. It is observed that there have been few studies concentrating on OSCC of the buccal cavity, where texture analysis has been widely used.[1,4-7] Al-Kadi[8] obtained meningioma texture features by four different texture measures.

However, most studies have been done in automatic oral submucous fibrosis (OSF) detection at cell levels.[9-13] Fractal analysis on time-resolved autofluorescence measurement in tumors and healthy tissues of the oral cavity ex vivo have been shown too.[14] Studies also deal with developing classifiers for classification of oral cancer based on clinical values of chosen factors from patient records.[15-17] Analyzing clinically significant and biologically interpretable features from microscopic biopsy images, a framework for automated detection and classification of cancer is proposed and examined.[18] Again, using support vector machine (SVM) classifier, shape features were extracted for distinguishing inflammatory and fibroblast cells[19]; while a study on classifier improvement using Gaussian transformation was implemented by Krishnan et al[20] on oral submucous fibrosis.

Studies of oral cancer on domains apart from microscopic images include endoscopic images of the esophagus where active contour-based segmentation and fuzzy rule-based classification were proposed[21]; a quantitative histomorphometric classifier on oropharyngeal SCC[22]; deep learning techniques on hyperspectral images[23,24]; random forest and SVM on confocal laser endomicroscopy (CLE) images[25,26]; combined machine learning (ML) techniques using asymmetry of a temperature distribution of facial regions as principle cue on digital infrared thermal imaging.[27] Hameed et al[28] proposed a novel cell nuclei feature extraction method for immunohistochemical scoring of oral cancer tissue images. They got a good result with 96.09% accuracy through linear discriminant analysis classifier.

It is clear from the above that deep learning techniques have not been attempted on automated analysis of microscopic images of biopsy for OSCC. The main deterrent to the venture may be the absence of an adequate number of images to carry out the study. As our dataset is not large enough, we have also not applied deep learning or data mining techniques in our case. On the other hand, no online dataset is available. Moreover, traditional ML techniques are cost-effective as it takes less time to train a model. In the case of deep learning, a GPU is required to train properly. In our approach, use of CPU makes the developed algorithm more usable at low level, that is, by using simple laptops and desktops (which are more viable for use in remote areas or economically lower countries).

The main idea of this study is to develop a very accurate algorithm which could be used as a screening tool for OSCC. Hence the binary classification approach was taken, to make it a tool for sieving out the malignant cases automatically. This would reduce the high number of slides every pathologist has to encounter and analyse daily and concentrate only on the malignant cases, thus increasing their efficiency.

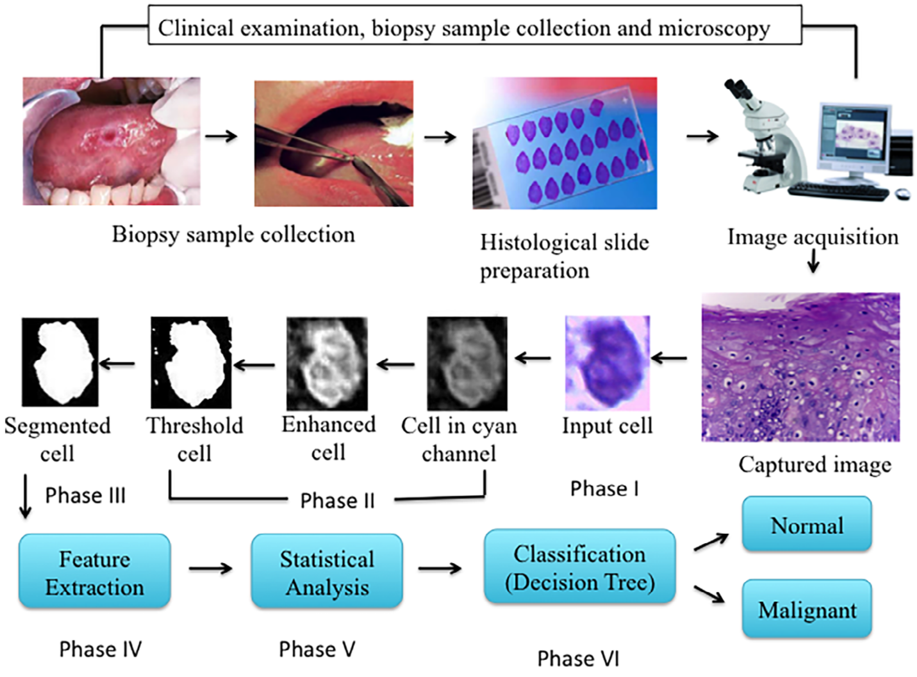

FIGURE 1 Graphical abstract with different phases involved in the classification of cells

2 | MATERIALS AND METHODS

The graphical methodology depicting the different phases involved in the proposed method is shown in Figure 1. It consists of six phases. Phase 1 is the imaging and database generation step. Phase 2 involves preprocessing. Segmentation of the nucleus is done in phase 3; Extraction of features is performed in phase 4; Phase 5 is the statistical analysis step and finally, the classification of cells is done in phase 6.

2.1 | Imaging and database generation

For this study, H&E stained biopsy slides were collected from two diagnostic centers: Ayursundra Healthcare Pvt. Ltd. (a prominent pathological and diagnostic center of the city) and Dr. B. Borooah Cancer Institute (the Regional Cancer Centre under Department of Atomic Energy, Government of India), following all ethical protocols. The slides were of 3 μm, one of the preferably used thicknesses of biopsy section used for preparing a slide in the histopathological analysis.[29] The microscopic digital imaging was done at Ayursundra Healthcare Pvt. Ltd. The reports of the patients were also collected for labeling the images. The biopsy slides were initially graded according to the Broder's grading system (namely normal, moderately differentiated SCC [MDSCC], well-differentiated SCC [WDSCC], and poorly differentiated SCC [PDSCC]) as followed by the laboratories of the region.[30] These were finally grouped to “normal” or “malignant” in the report. The details of the acquired images are shown in Table 1.

As there is no OSCC cell level database available online which may be considered as a benchmark dataset, we have prepared a database for the study with the indigenous data collected. Only those cases, which had confirmed histopathological correlation, were included in the study. At the time of capturing the images, the size of all the images were 2048 × 1536 pixels. Later, skilled and certified pathologists have selected the region of interest (ROI), that is, cell nucleus, which was used for ground truth preparation. After that, we have hand-cropped the cells from the original image and created the nuclei dataset.[31,32] The cropped images were of various sizes

TABLE 1 Details of captured images

Images	Details
Biopsy slides	40 (From 40 patients)
Slides with normal cells	13 (Normal histology slides [ie, those without any dysplasia or tumor] were considered as normal control)
Slides with malignant cells	27
Normal cells (cropped out)	118
Malignant cells (cropped out)	334
Image acquired equipment	Leica ICC50 HD microscope
Image format	JPEG
Magnification used	×400 (×40 objective lens × ×10 eyepiece)
Camera pixel size	3.2 μm
Effective pixel size	0.16 μm
Numerical aperture	0.65
Size of acquired image	2048 × 1536 pixels, 3.1 megapixels

depending upon the varying size of the cell. Since size based features are an important consideration in our case, resizing is not an option as it would distort the feature's information, that is, distort the cell's original size. After segmentation, the segmented image (nuclei) carried the actual information of size which were used to extract all the feature information (morphological as well as textural) quantitatively. Here, we are not considering connected nucleus. We manually cropped the cells as manual cropping generates a more reliable training dataset and hence any technique based on it would be the most appropriate. A total of 118 normal cells and 334 malignant cells were manually cropped out from the grabbed images.

2.2 | Data pre-processing

Some of the collected slides were dark and some others were lightly stained. Hence, there was a color difference in the captured images from these slides. Therefore, pre-processing techniques were applied to these images to eliminate staining difference so that it does not affect our result. As a pre-processing step, color channeling was done on the images. The different color components provide clear and distinctive diverse information about an image. Accordingly, we converted the images from the original red (R), green (G), and blue (B) components into different components using the following mathematical calculations:

$$\text{Cyan}\,(\text{C}) = (\text{G}+\text{B}), \text{Magenta}\,(\text{M}) = (\text{R}+\text{B}), \text{Yellow}\,(\text{Y}) = (\text{R}+\text{G}).$$

Hue (H) is obtained by using the following calculation:

$$\text{num} = \frac{(\text{R}-\text{G})+(\text{R}-\text{B})}{2},$$

$$\text{den} = \sqrt{(\text{R}-\text{G})^2+(\text{R}-\text{B})(\text{G}-)}$$

$$\theta = \cos^{-1}\left(\frac{\text{num}}{\text{den}}\right),$$

$$H = \begin{cases} \theta & \text{if } B \le G \\ 360-\theta & \text{if } B > G \end{cases}.$$

Next, Saturation (S) was obtained by using the calculation:

$$\text{num} = \min(\min(\text{R},\text{G}),\text{B});$$

$$\text{den} = \text{R} + \text{G} + \text{B};$$

$$\text{S} = 1 - \frac{3*\text{num}}{\text{den}}.$$

Intensity (I) was obtained as follows:

$$\text{I} = \frac{(\text{R}+\text{G}+\text{B})}{3}.$$

And finally, HIS was extracted as: HIS = cat (3,H,S,I), that is, combination of the three components.

In Figure 2, one image from our database is visualized in different color channels.

The channel which gives the aimed information can be considered for further processing ignoring the other channels. Hence, we chose the cyan channel for further processing, as it gives us the best outcome. Next, to eliminate the illumination differences of the images during acquisition, it was required to adjust the contrast of the images. Hence, the inbuilt function of MATLAB software, *imadjust*() was applied. This function maps the values in intensity image I to new values in I1 such that 1% of data is saturated at low and high intensities of I. The output after applying *imadjust* function on one of the image is shown in Figure 3.

2.3 | Nucleus segmentation

Following image pre-processing step, we used image segmentation techniques to separate the nucleus from the background. A segmentation technique is proposed in this paper to segment out the nucleus.

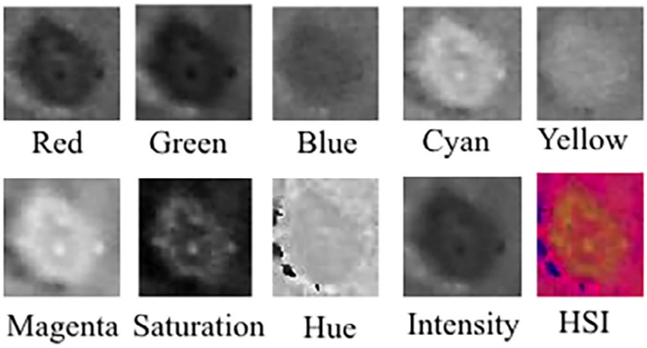

FIGURE 2 One of the images from our database in different color channels

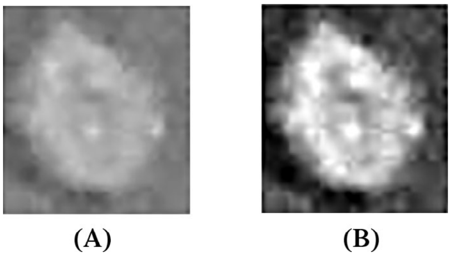

FIGURE 3 Image, A, before applying *imadjust*, B, after applying *imadjust* function

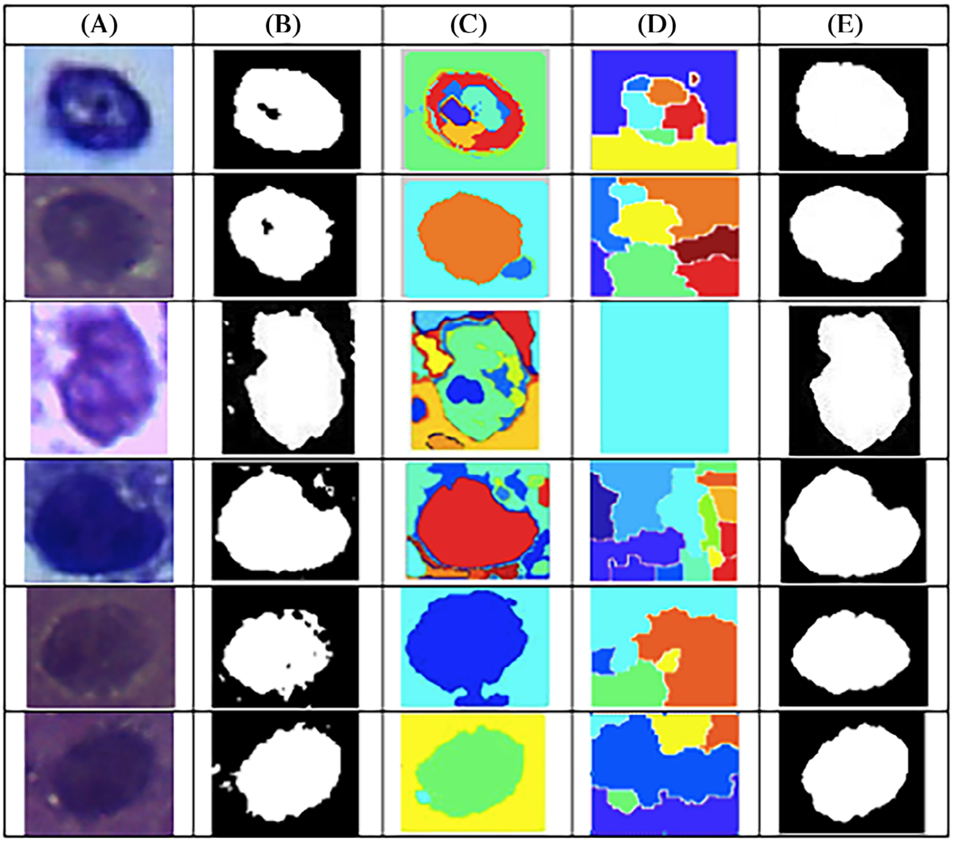

FIGURE 4 Outcomes of different segmentation method of some cells, A, original, B, Otsu, C, MSER, D, Watershed, E, proposed.

Several standard segmentation techniques like Otsu's segmentation, Watershed segmentation and maximally stable extremal regions (MSER) were applied, but none of them alone gave satisfactory results. Finally, we merged Otsu's method with the morphological operation (erode and dilate) to achieve the desired segmentation level of the nucleus.[33] The comparison of the segmented images with Otsu, MSER, and Watershed segmentation is depicted in Figure 4 and the various steps for achieving the same are shown in Figure 5. It is clear from these figures that our proposed methodology achieves the best segmentation results of the nuclei. The ROI, that is, the nucleus was not segmented out properly using MSER or Watershed segmentation techniques. In both the techniques, areas other than nucleus were segmented out along with the nucleus. In some cases, it fails to segment the nucleus. MSER technique may have failed because here regions are created by areas which are connected and typified by almost uniform intensity, surrounded by contrasting background. It has inadequate performances on blurred and/or textured images, particularly with scale variations.[34] It is clear that the images in our dataset belong to this category and hence was not successfully tackled by this method. Watershed based approach, too, may have not achieved success being an example of region-based threshold technique, where the image is portioned based on the similarity of gray levels. In natural images, this method almost always has an over-segmentation outcome.[35] In comparison, the Otsu method's principle is to select the threshold that minimizes the intraclass variance of the threshold black and white pixels. It is a nonparametric and unsupervised method of automatic threshold selection for picture segmentation. Hence, using the Otsu method, we achieve a better result than MSER or Watershed segmentation. But the Otsu method alone was not sufficient to get the best result. From Figure 4, we can see that in some images, after applying Otsu, some holes and in some other cases some other bodies were present along with the nucleus. To eliminate these, we combined the Otsu method with morphological erosion and dilation, which provides the best result. A structuring element of size 3 pixels was used for erosion and for dilation a structuring element of size 4 pixels was applied. It was applied only for one time in each image and was fixed for all images. The morphological functions enable to focus only on the required nucleus part, ignoring the holes and bodies.

2.4 | Extraction of features

At this stage, morphological and textural features were extracted from the segmented out nuclei.

2.4.1 | Morphological features

As morphological features, we have extracted area, perimeter, eccentricity and compactness, moments and Fourier descriptor. The morphological features were used to analyze the size and shape of the

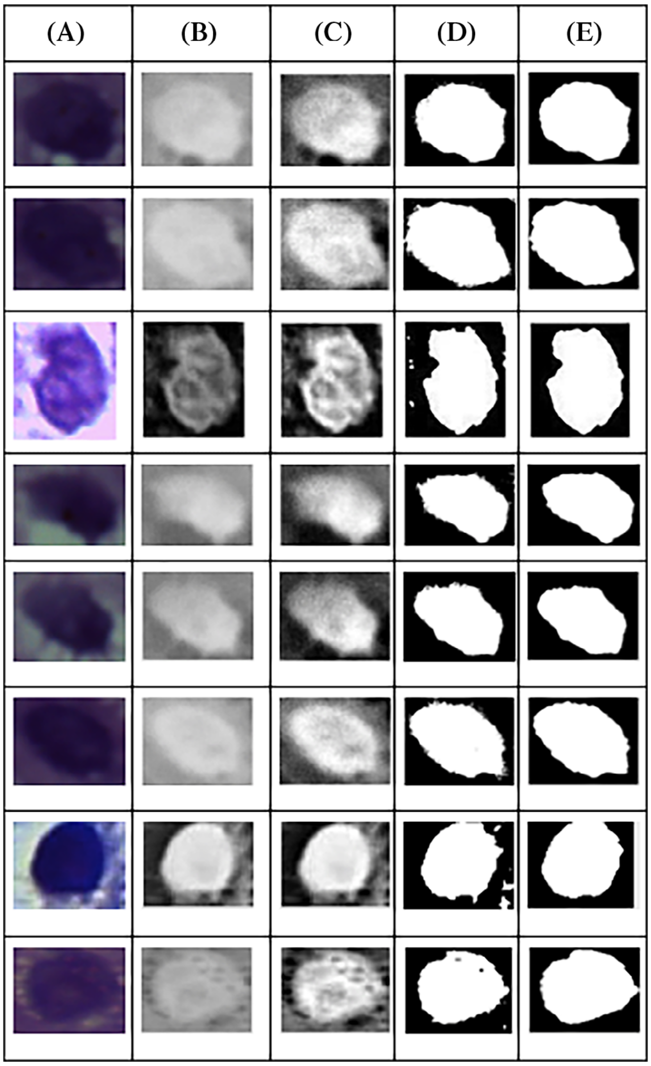

FIGURE 5 Different steps used for segmentation, A, original, B, cyan image, C, enhanced, D, threshold image, E, segmented out the nucleus

nucleus, that is, any enlargement or distortion in shape. These parameters are the most common features used for diagnosis by pathologists and hence most interpretable as well as preferable by them in connection with automation. All total we have extracted 522 morphological features.

2.4.2 | Textural features

Secondly, the textural features of the images were extracted. First-order statistics of histogram features like mean, variance, skewness, kurtosis, energy, and entropy,[36] as well as second-order statistics of Gray level co-occurrence matrix (GLCM), features namely contrast, variance, autocorrelation, cluster prominence, angular second moment, correlation, maximum probability, difference variance, entropy, cluster shade, dissimilarity, difference entropy, sum entropy, homogeneity, inverse difference moments, sum average, the sum of squares, sum variance, information measure of correlation 1, information measure of correlation 2, sum homogeneity, and the sum of energy were extracted.[37,38] We have examined GLCM in all four directions of $\theta = 0°$, $\theta = 45°$, $\theta = 90°$, and $\theta = 135°$. To make the GLCM symmetrical, the transpose of the matrices were added to the matrices formed along with the above four angles. Moreover, gray level run length matrix (GLRLM) features such as run percentage (RP), run length nonuniformity (RLN), low gray level run emphasis (LGRE), high gray level run emphasis (HGRE), short-run emphasis (SRE), long-run emphasis (LRE) and gray level nonuniformity (GLN) were extracted.[39] It was used to relate the connected length of pixels. Then Local binary pattern (LBP)[40,41] of individual cells, Tamura features (coarseness, contrast, and directionality) and histogram of gradient (HOG) were also extracted.[42-44] Applying the aforementioned techniques, a total of 441 textural features were extracted.

2.5 | Statistical analysis

It is evident from the above section that now we have a total of 963 features for further analysis of the cells. To select the statistically significant features, that is, to check whether the differences in the values of the extracted features of the cells for the different classes were statistically significant or not, a two-level feature selection technique was applied with two well-known methods: namely Karl Person's *t* test and principal component analysis (PCA). Since the level of accuracy in medical diagnosis must be high the selected level of confidence was 1%. Hence the null hypotheses were $H_0^{a,p,e,c}$ (there is no difference in the values of the parameters for different classes) vs alternative hypotheses $H_1^{a,p,e,c}$ (there is a difference in the values of the parameters for different classes). A two-tailed test was conducted for the purpose as different features were compared and initially which parameters would show higher value were unknown. On the other hand, PCA is a well-known statistical data compression method, which was used to identify a smaller number of uncorrelated variables.

2.6 | Classifiers

As classification urposes, we have employed five classifiers namely, decision tree, SVM, *k*-nearest neighbor (KNN), discriminant analysis, and logistic regression to find out the best suitable one.

2.6.1 | Decision tree classifier

We have first applied decision tree classifier on our data, as it is a simple and commonly used classifier. It applies a sequence of carefully crafted questions concerning the attributes of the test record. Decision tree breaks down a dataset into smaller and smaller subsets and at the same time, a related decision tree is incrementally grown. The ultimate result is a tree with decision nodes and leaf nodes. The top node of the tree is called the root node. Leaf node denotes a classification or decision.[45]

2.6.2 | KNN classifier

Next, we applied KNN classifier, which is a simple but powerful supervised ML technique. It uses data and categorizes new data based on similarity measures. Here, a feature vector is allocated the class that is most common among its KNNs.[46]

2.6.3 | SVM classifier

To improve the accuracy further, we applied SVM classifier.[12,47,48] The SVM classifier separates the data into normal and malignant classes by creating a line or a hyperplane. Support vectors are data points that are nearer to the hyperplane that influences the position and orientation of the hyperplane. We expand the margin of the classifier using these support vectors. SVM is based on the idea of margin expansion.[49,50]

2.6.4 | Linear discriminant analysis

LDA, which is a supervised ML technique, yields a set of prediction by estimating the probability that a new set of inputs belongs to each class. The class, which acquires the maximum probability, is the output class and a prediction is made.[51]

2.6.5 | Logistic regression

Logistic Regression is an ML classification algorithm, which is a predictive analysis. Logistic regression analyzes the output of one or more independent variables to predict the possibility of a categorical dependent variable. The dependent variable is a binary variable that contains data coded as 1 (in our case, malignant) or 0 (normal) in logistic regression.[52]

3 | IMPLEMENTATION

Implementation of the whole technique is done on MATLAB_R2016b version (The MathWorks, Inc., Natick, Massachusetts). Some coding was done by inbuilt Matlab function, customized routines, and available source code.

4 | ASSESSMENT OF THE RESULTS

The assessment of each classifier is done based on sensitivity, precision, specificity, and overall accuracy. The used formulas are as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%,$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%,$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\%,$$

$$\text{Overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\%,$$

where TP, true positive; TN, true negative; FP, false positive; FN, false negative.

Moreover, the receiver operating characteristic (ROC) is also plotted to assess the results of the classifier.

5 | RESULTS

The proposed segmentation method was employed on 452 cells (118 normal cells and 334 malignant cells) to segment out the nuclei. Total of 522 morphological features was extracted from these nuclei as reported earlier. Then, as mentioned previously, Karl Person's *t* test was applied to the extracted features to test the statistical significance. It reduces the feature set to 516 features. Again, PCA was applied to the *t* test selected features. Upon implementation of PCA, 340 features were found to be significant. PCA was used to check whether PCA selects only those features, which leads to better classification accuracy or in other words reduces the feature set. Table 2 summarizes these outcomes.

Now, we got two feature sets of 452 × 516 (from *t* test) and 452 × 340 (from PCA) dimensions for further analysis. These two feature sets were fed into five classifiers namely, SVM, KNN, decision tree classifier, logistic regression, and linear discriminant separately to find out the best suitable one. Five-fold cross-validation technique was used for testing the classifiers. When there is limited dataset there is a fear of overfitting of the model used. Cross-validation is a potent precautionary measure against overfitting. For this, the entire dataset was divided into five parts, four parts were used for classifier development (training set) and the other remaining part was used as a test set. This part was used to evaluate the classifier. The same technique was done for five times (folds). Four different parts were used for training the classifier whereas the remaining one was used for validation each time. The attained classification accuracy is the average of all the five test classification results. Figure 6 shows the

TABLE 2 Outcomes of the two-level statistical analysis of the features

Feature	Original	*t* test (*P* < .05)	PCA
Morphology	522	516	340
Texture	441	349	103
Total	963	865	443

Abbreviation: PCA, principal component analysis.

classification results for both the sets. Decision tree classifier performed better than the other four classifiers.

Once again, *t* test was applied on the extracted 441 textural features for identifying significant ones. This process reduced the feature set to 349 features. After that PCA was implemented on these selected features. This time the feature set was reduced to103 features (Table 2). Now, these two feature sets with 452 × 349 (from *t* test) and 452 × 103 (from *t* test) dimensions were fed into the said five classifiers.

All the steps used in case of morphological features were now repeated for these two feature sets also. Classification accuracies of different classifiers for *t* test selected features and PCA selected features is depicted in Figure 7. Finally, we combine both the PCA selected feature set of morphological features and textural features and this combined feature set (with dimension 452 × 443) was used for classification.

Accuracies of 99.78%, 93.6%, 62.9%, 93.6%, and 54.3% were obtained for the decision tree, linear discriminant, logistic regression, SVM, and KNN, respectively.

It is clear from the results that the decision tree classifier performed better than the other four classifiers with higher accuracy.

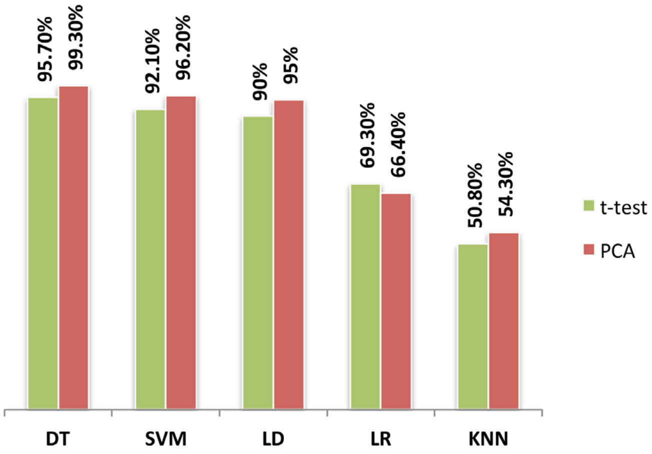

FIGURE 6 Classification accuracies for *t* test and PCA selected morphological features. PCA, principal component analysis

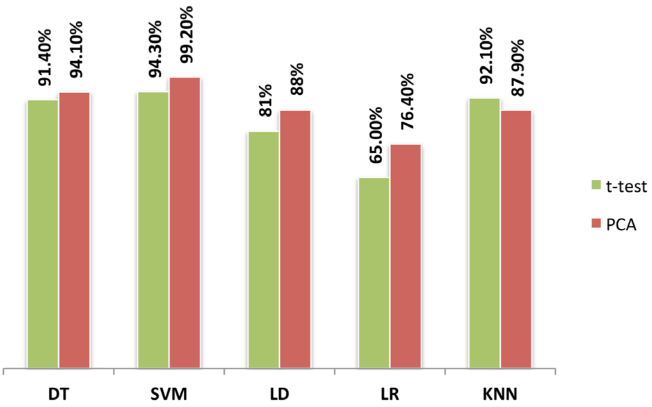

FIGURE 7 Classification accuracies for *t* test and PCA selected textural features. PCA, principal component analysis

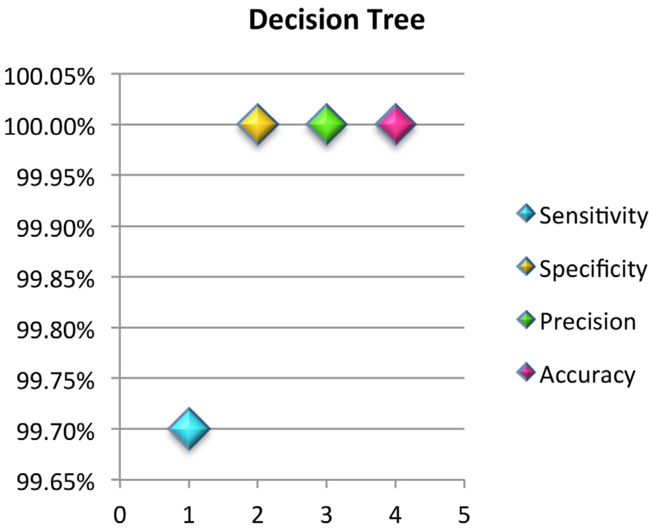

FIGURE 8 Sensitivity, specificity, precision, and overall classification accuracy for all 443 PCA selected features using decision tree classifier. PCA, principal component analysis

TABLE 3 Confusion matrix for Decision Tree classifier

n = 452	Predicted: No	Predicted: Yes
Actual: No	118	0
Actual: Yes	1	333

Note: n = total number of cases.
Abbreviation: SVM, support vector machine.

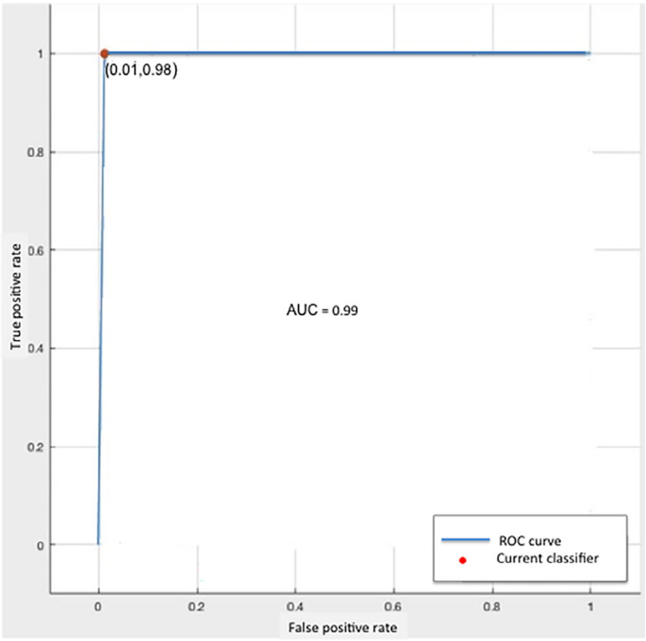

FIGURE 9 ROC for decision tree classifier with AUC score 0.99. AUC, area under the curve; ROC, receiver operating characteristic

TABLE 4 Analysis of different classifiers used for the combined feature set by training time and accuracy

	Number of images									
	10		100		200		300		400	
Classifiers	Accuracy (%)	Training time (s)	Accuracy (%)	Training time (s)	Accuracy (%)	Training time (s)	Accuracy (%)	Training time (s)	Accuracy (%)	Training time (s)
DT	52	2.098	66	2.587	89	2.990	95	3.6986	99.78	4.021
LD	91	2.998	91.8	3.087	93	3.549	93.4	3.945.5577	93.6	4.3907
LR	60	4.065	61	4.890	62.2	5.309	62.5	5.8573	62.9	6.722
SVM	70	2.530	81.5	2.587	89.7	2.580	92.2	3.0880	93.6	3.5973
KNN	78	2.540	73.4	2.786	69.9	3.352	59.7	4.015674	54.3	4.5873

Abbreviations: DT, decision tree; KNN, *k*-nearest neighbor; LD, linear discriminant; LR, logistic regression; SVM, support vector machine.

TABLE 5 Some of the oral cancer classification techniques based on morphological and textural features of histopathological images used in literature and our proposed method

Topics	Image type	Number of images	Type of cells	Features extracted	Classifier used	Results (%)
Discrimination of inflammatory and fibroblast cells of OSF from normal oral mucosa[19]	Normal, OSF	Normal = 451 (Cells) OSFWD = 1912 (Cells) OSFD = 1914 (Cells)	Connective	Size and shape	SVM	Normal 96.55, OSFWD 94.65, OSFD 92.33
Analysis of histomorphometric features of the basal cell nuclei for their size, shape, and intensity of staining[11]	Normal, OSF	Normal = 885 (Cells) OSFD = 470 (Cells)	Epithelial	Morphometric and mean intensity of nuclei	Bayesian classifier	99.04
Identification of OSF from morphological and textural properties of the basal cell nuclei[12]	Normal, OSF	Normal = 771 (Nuclei) OSF = 423 (Nuclei)	Epithelial	Morphological and textural features of the basal cell nuclei	SVM	99.66
Textural analysis of OSCC[1]	Normal, OSCC	Normal = 223 (Tissue images) OSCC = 253 (Tissue images)	Epithelial	Textural features	SVM	100
Automated identification of OSCC from whole image strip by morphological, textural, and color features[55]	Normal, OSCC	Normal = 237 (Nuclei) OSCC = 483 (Nuclei)	Epithelial	Morphological, textural, and color	DT SVM and logistic regression SVM, logistic regression, and linear discriminant	99.4 100 100
Proposed study	Normal, OSCC	Normal = 118 (Nuclei) OSCC = 334 (Nuclei)	Epithelial	Morphological Textural Combined	Decision tree SVM Decision tree	99.3 99.2 99.78

Abbreviations: DT, decision tree; OSCC, oral squamous cell carcinoma; OSFD, oral submucous fibrosis with dysplasia; OSFWD, oral submucous fibrosis without dysplasia; SVM, support vector machine.

Hence, decision tree classifier was selected as the suitable classifier for this work. The following graph in Figure 8 displays the classification results for the decision tree.

The confusion matrix for decision tree classifier formed from the four outcomes (true positive, true negative, false positive, and false negative) is shown in Table 3.

Further, the performance of the classifier is again evaluated by plotting a ROC curve.[53] Area under the curve (AUC) of the ROC was achieved 0.99 (Figure 9), which means the built classifier is an efficient one.

We have also analyzed the different classifier used for the final result (considering 443 features) concerning their training time. It was found that training time increases as the number of input data size increase for all classifiers. This analysis is summarized in Table 4. It was also found that in most of the cases accuracy increases as the number of images increases. This implies that the accuracy of the classifiers increases with the size of training data.

6 | DISCUSSION

In the evaluation of malignancy of oral cancer, great importance is placed on the changes in the morphological features, like nuclear size and shape. To overcome the unreliability in the biased analysis of these features, a more objective method would be the worth of computer-based image analysis techniques. In this study, we have observed that OSCC cells showed higher numeric values for the nuclear area, perimeter, compactness and eccentricity as compared to normal cells. The increase in the nuclear area may be biologically revealing of the malignant potentiality of the malignant cells, which may be interrelated with their increased and abnormal metabolic action.[54]

Numerous studies have been carried out for the measures of texture analysis to characterize lesions in the various areas of the body and to differentiate them from the normal tissues.[1,5-7] 100% classification accuracy was obtained in OSCC detection using textural features.[1] In another study, Rahman et al[55] carried out research on shape, texture and color features of OSCC on whole image strips and achieved good results. In that study, the whole image strip was used whereas here in this study cropped out cells are considered to understand which process gives better performances. Only a few pieces of research have been carried out for oral cancer diagnosis based on morphological features using histopathological images. But many of this type of studies have been carried out on OSF. Extracting shape features of the subepithelial connective tissue cells, Krishnan et al[19] classified inflammatory and fibroblast cells of OSF from normal oral mucosa. In another study, using Bayesian classifier, they observed an increase in the dimensions (area and perimeter), shape parameters and decreasing mean nuclei intensity of the nuclei in OSF with dysplasia than normal oral mucosa.[11] Again, morphological and textural properties of the basal cell of OSF were studied to distinguish from normal oral mucosa employing SVM classifier. They achieved a good result with 99.66% accuracy.[12] Thus, based on the above pieces of evidence, we felt that it is a necessity to build a tool for OSCC diagnosis using morphological and textural features. Although OSCC segmentation and classification is not directly comparable with OSF segmentation and classification, a summary result of classification techniques based especially on morphological and textural features of histopathological images till date, and our proposed method is shown in Table 5. These features play a very important role in oral cancer diagnosis. Here, we have used these features quantitatively for automated diagnosis of OSSC. Our system is capable of identifying the unknown class with good accuracy. Hence, it can be used as an efficient tool for making accurate decisions in oral cancer diagnosis.

From Table 5, it is evident that in our previous two studies,[1,55] we have achieved higher accuracy rates than this study. It is due to the number of images, that is, cells, that we have considered. In both the previous two studies, the number of cells (dataset) is more than this study. In the first case where we have analyzed only texture features, we had considered tissue level images, hence the number of training cells are quite larger than this study.[1] Thus in this study, we have also tried to establish a relationship between the number of trained cells and the subsequent result. So, it can be said that accuracy increases with the number of trained images, as further revealed from Table 4.

7 | CONCLUSIONS

From this study, it can be said that both morphological and textural features play a very important role in distinguishing malignant cells from normal cells in OSCC diagnosis. Here, best classification accuracies were obtained employing decision tree classifier. This algorithm can be developed to a software to be integrated with the microscope system, where the pathologist observes the slides and takes note of numerical features on hand-cropped portions, which mostly include cell nuclei. As an ML algorithm, this is more feasible than DL algorithms. We have achieved an accuracy of 99.78% in OSCC detection by applying our algorithm.

ACKNOWLEDGMENTS

T. Y. R. wishes to extend her thanks to Ayursundra Healthcare Pvt. Ltd., Guwahati, Assam and Dr B. Borooah Cancer Institute, Guwahati, Assam for proving histopathological slides for this research work.

AUTHOR CONTRIBUTIONS

T. Y. Rahman: Conceptualization; methodology; investigation; formal analysis; resources; writing-original draft; writing–review and editing. **Lipi B. Mahanta:** Conceptualization; methodology; project administration; resources; supervision; visualization; writing-review and editing. **Hiten Choudhury:** Conceptualization; formal analysis; supervision. **Anup Das:** Resources; supervision. **Jagannath D. Sarma:** Resources; supervision.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ETHICS STATEMENT

All ethical protocols were maintained during this phase. The institutional ethics duly approved this study (No. IEC [HS]/IASST/1082/2014-15/2).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. The images can be found online at https://doi.org/10.1016/j.dib.2020.105114, https://doi.org/10.17632/ftmp4cvtmb.1 and https://github.com/Tabassum2019/A-histopathological-image-repository-of-normal-epithelium-of-Oral-Cavity-and-OSCC/blob/master/README.md

ORCID

Lipi B. Mahanta https://orcid.org/0000-0002-7733-5461

REFERENCES

1. Rahman TY, Mahanta LB, Chakraborty C, Das AK, Sarma JD. Textural pattern classification for oral squamous cell carcinoma. *J Microsc*. 2017;269(1):85-93.
2. Kong J, Sertel O, Shimada H, Boyer KL, Saltz JH, Gurcan MN. Computer-aided evaluation of neuroblastoma on whole-slide histology images: classifying grade of neuroblastic differentiation. *Pattern Recognit*. 2009;42:1080-1092.
3. Pinto TW, de CMAG, Pedronette DCG, Martins PS. Image segmentation through combined methods: watershed transform, unsupervised distance learning and normalized cut. *IEEE Southwest Symposium on Image Analysis and Interpretation*. San Diego, CA: IEEE; 2014:153-156.
4. Kolarevic D, Tomasevic Z, Dzodic R, Kanjer K, Vukosavljevic DN, Radulovic M. Early prognosis of metastasis risk in inflammatory breast cancer by texture analysis of tumour microscopic images. *Biomed Microdevices*. 2015;17(5):92.
5. Raja JV, Khan M, Ramachandra VK, Al-Kadi O. Texture analysis of CT images in the characterization of oral cancers involving buccal mucosa. *Dentomaxillofac Rad*. 2012;41:475-480.
6. Al-Kadi OS, Watson D. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. *IEEE Trans Biomed Eng*. 2008;55:1822-1830.
7. Petkovska I, Shah SK, McNitt-Gray MF, et al. Pulmonary nodule characterization: a comparison of conventional with quantitative and visual semi quantitative analysis using contrast enhancement maps. *Eur J Radio*. 2006;59:244-252.
8. Al-Kadi OS. Texture measures combination for improved meningioma classification of histological images. *Pattern Recognit*. 2010;43:2043-2053.
9. Krishnan MMR, Shah P, Choudhary A, Chakraborty C, et al. Textural characterization of histopathological images for oral sub-mucous fibrosis detection. *Tissue Cell*. 2011;43:318-330.
10. Krishnan MMR, Venkatraghavan V, Acharya UR, et al. Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm. *Micron*. 2012;43:352-364.
11. Krishnan MMR, Pal M, Paul RR, Chakraborty C, et al. Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of Oral submucous fibrosis. *J Med Syst*. 2012;36:1745-1756.
12. Krishnan MMR, Pal M, Paul RR, Chakraborty C, et al. Hybrid segmentation, characterization and classification of basal cell nuclei from histopathological images of normal oral mucosa and oral submucous fibrosis. *Expert Syst Appl*. 2012;39:1062-1077.
13. Das DK, Chakraborty C, Sawaimoon S, Maiti AK, Chatterjee S. Automated identification of keratinization and keratin pearl area from in situ oral histological images. *Tissue Cell*. 2015;47:349-358.
14. Klatt J, Gerich CE, Grobe A, et al. Fractal dimension of time-resolved autofluorescence discriminates tumour from healthy tissues in the oral cavity. *J Craniomaxillofac Surg*. 2014;42(6):852-854.
15. Prabhakar SK, Rajaguru H. Performance analysis of linear layer neural networks for oral cancer classification. *6th ICT International Student Project Conference (ICT-ISPC)*. Skudai, Malaysia: IEEE; 2017 INSPEC Accession Number: 17303242.
16. Latithamani K, Punitha A. Detection of oral cancer using deep neural based adaptive fuzzy system in data mining techniques. *Int J Recent Technol Eng*. 2019;7(5S3):397-404.
17. Tetarbe A, Choudhury T, Toe TT, Rawat S. Oral cancer detection using data mining tool. *3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. Tumkur, India: IEEE; 2017:35-39.
18. Kumar R, Srivastava R, Srivastava S. Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features. *J Med Eng*. 2015;2015:1-14.
19. Krishnan MMR, Chakraborty C, Paul RR. Quantitative analysis of sub-epithelial connective tissue cell population of oral submucous fibrosis using support vector machine. *J Med Imaging Health Inform*. 2011;1:4-12.
20. Krishnan MMR, Shah P, Chakraborty C, Ray AK, et al. Statistical analysis of textural features for improved classification of oral histopathological images. *J Med Syst*. 2012;36:865-881.
21. Hiremath PS, Iranna HY. Fuzzy rule based classification of microscopic images of squamous cell carcinoma of esophagus. *Int J Comp Appl*. 2011;25(11):30-33.
22. Lewis JS Jr, Ali S, Luo J, Thorstad WL, Madabhushi A. A quantitative histomorphometric classifier (QuHbIC) identifies aggressive versus indolent p16-positive oropharyngeal squamous cell carcinoma. *Am J Surg Pathol*. 2014;38:128-137.
23. Jeyaraj PR, Nadar ERS. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *J Cancer Res Clin Oncol*. 2019;145(4):829-837.
24. Halicek M, Lu G, Little JV, Wang X, Patel M, et al. Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *J Biomed Opt*. 2017;22(6):60503.
25. Jaremenko C, Maier A, Stefan S, et al. Chapter 82: Classification of confocal laser endomicroscopic images of the oral cavity to distinguish pathological from healthy tissue. *Bildverarbeitung für Die Medizin*. Berlin, Heidelberg: Springer Vieweg; 2015:479-485.
26. Marc A, Knipfer C, Oetter N, et al. Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. *Sci Rep*. 2017;7:11979.
27. Chakraborty M, Mukhopadhyay S, Dasgupta A, et al. A new paradigm of oral cancer detection using digital infrared thermal imaging. *Proceedings of SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis, 97853I*. SPIE Digital Library, CA: SPIE Digital Library; 2016.
28. Hameed KAS, Banumathi A, Ulaganathan G. Cell nuclei classification and immunohistochemical scoring of oral cancer tissue images: machine-learning approach. *Asian J Res Soc Sci Human*. 2016;6(10):732-747.
29. Rahman TY, Mahanta LB, Das AK, Sarma JD. Towards digital diagnosis of oral cancer: a study on optimum preferences of histopathological techniques and features. *Med Legal Update*. 2020;20(3):232-238.
30. Shrivastava S, Shakya R. Study on histological grading of Oral squamous cell carcinoma and its co-relationship with regional lymph node metastasis. *Int J Sci Res*. 2019;8(2):40-43.
31. Rahman TY. A histopathological image repository of normal epithelium of oral cavity and oral squamous cell carcinoma. New York, NY:

Elsevier Inc; 2019; *Mendeley Data, v1*. Retrieved from http://dx.doi.org/10.17632/ftmp4cvtmb.1
32. Rahman TY, Mahanta LB, Das AK, Sarma JD. Histopathological imaging database for oral cancer analysis. *Data Brief*. 2020;29:105114.
33. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62-66.
34. Śluzek A. Improving performances of MSER features in matching and retrieval tasks. In: Hua G, Jégou H, eds. *Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science*. Vol 9915. Cham, Switzerland: Springer; 2016.
35. Amoda N, Kulkarni RK. Efficient image segmentation using watershed transform. *Int J Comput Sci Technol*. 2013;4(2):214-218.
36. Malik F, Baharudin B. The statistical quantized histogram texture features analysis for image retrieval based on median and Laplacian filters in the DCT domain. *Int Arab J Info Tech*. 2013;10(6):616-624.
37. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cyber*. 1973; SMC-3;(6):610-621.
38. Girisha AB, Chandrashekhar MC, Kurian MZ. Texture feature extraction of video frames using GLCM. *IJETT*. 2013;4(6):2718-2721.
39. Tang X. Texture information in run-length matrices. *IEEE Trans Image Process*. 1998;7(11):1602-1609.
40. Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognit*. 1996;29:51-59.
41. Silva C, Bouwmans T, Frelicot C. An eXtended center-symmetric local binary pattern for background modeling and subtraction in videos. *VISAPP 2015*. Berlin, Germany: Scitepress Digital Library, Science and Technology Publications, Lda; 2015.
42. Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. *Syst Man Cybernet*. 1978;8(6):460, 4309999-473. https://doi.org/10.1109/TSMC.1978.
43. Patel MJ, Gamit CN. A review on feature extraction techniques in content based image retrieval. *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. Chennai, India: IEEE; 2016. https://doi.org/10.1109/WiSPNET.2016.7566544.
44. Dalal N, Triggs B. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) INSPEC*. San Diego, CA: IEEE; 2005 Accession Number: 8588935.
45. Rao H, Shi X, Rodrigue AK, et al. Feature selection based on artificial bee colony and gradient boosting decision tree. *Appl Soft Comput*. 2019;74:634-642.
46. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. Burlington, MA: Morgan Kaufmann; 2005.
47. Burges J, Christopher C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowl Discov*. 1998;2:121-167.
48. Huang CL, Liao HC, Chen MC. Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Syst Appl*. 2008;34:578-587.
49. El-Naqa I, Yang Y, Wernick MN, et al. A support vector machine approach for detection of microcalcifications. *IEEE Trans Med Imaging*. 2002;21:1552-1563.
50. Vapnik V. *Statistical Learning Theory*. 2nd ed. New York, NY: Wiley; 1998.
51. Hardle WK, Simar L. *Applied Multivariate Statistical Analysis*. 4th ed. Berlin, Heidelberg: Springer Science & BusinessMedia; 2012.
52. Dobson AJ, Barnett A. *An Introduction to Generalized Linear Models*. New York, NY: CRC Press; 2008 Chapter 8.
53. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27:861-874.
54. Duncan JS, Ayache N. Medical image analysis: Progress over two decades and the challenges ahead. *IEEE Trans Pattern Anal Mach Intell*. 2000;22:85-106.
55. Rahman TY, Mahanta LB, Das AK, Sarma JD. Automated oral squamous cell carcinoma identification using shape, texture and color features of whole image strips. *Tissue Cell*. 2020;63:101322.

How to cite this article: Rahman TY, Mahanta LB, Choudhury H, Das AK, Sarma JD. Study of morphological and textural features for classification of oral squamous cell carcinoma by traditional machine learning techniques. *Cancer Reports*. 2020;3:e1293. https://doi.org/10.1002/cnr2.1293