The Seybold Report

ISSN 1533-9211

DOI 10.5281/zenodo.6637102

IDENTIFICATION OF OROPHARYNGEAL CANCER USING MACHINE LEARNING MODEL

KUMAR R[1], Dr S PAZHANIRAJAN[2] and Dr S UMAMAHESWARAN[3]

1. Research Scholar, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Chidambaram. Assistant Professor, MVJ College of Engineering, Bangalore. Email: rkumarmecse@gmail.com
2. Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Chidambaram.
3. Professor, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore.

ABSTRACT

Oropharyngeal cancer is a main worldwide health problem accounting for **606,520** deaths in 2020 and it is most predominant in middle- and low-income nations. The most important purpose is being carried out this research is to find the Oropharyngeal Cancer Lesions affected region in the tongue images. The combined diagnostic framework with hybrid features selection techniques is utilised in this study to discover the traits that assist the most to the detection of Oropharyngeal cancer, reducing the amount of features obtained from a range of patient information indirectly. Using hybrid feature selection, twenty- five qualities were reduced to 14 features. Support Vector Machine.Following that, four classifiers were employed to forecast the identification of patients with Oropharyngeal cancer: Updatable Multilayer Perceptron, Nave Bayes, and K- Nearest Neighbors Furthermore, after adding feature subset choice with SMOTE during preprocessing stages, the SVM surpasses other machine learning algorithms, according to the findings. Using the initial data gathered in this study, a hybrid classifier algorithm was assessed to detect Oropharyngeal cancer lesions and features like color, texture, and geometry were extracted. Our initial findings establish support vector machine has the probable to challenge this stimulating task.

Keyword Head: GVF algorithm, Oropharyngeal Cancer Lesions, hybrid classifier algorithm, support vector machine.

1. INTRODUCTION

Oropharyngeal Cancer is because either the pancreas does not produce adequate insulin, or the cells of the body fail to respond properly to the insulin generated. Oropharyngeal Cancer Lesions type 1 - a disease that causes in autoimmune annihilation of insulin-creating beta cells of the pancreas. Oropharyngeal Cancer Lesions kind 2 - a metabolic disease that is described by high blood sugar levels in the framework of relative insulin and insulin resistance defect.

Basically, Oropharyngeal Cancer Lesions is a grouping of metabolic diseases where high blood glucose levels over an extended period. This high blood glucose leads to the signs and symptoms of increased urination, improved hunger, and improved thirst. Organic, Oropharyngeal Cancer can cause many complications. Tongue images have been seized by utilizing a specifically constructed in-house device taking color correction into account. Every image was segmented to find its front pixels. The three groupings of features like geometry, texture, and color have been extracted from the tongue foreground image with the appropriate pixels located. The image processing primarily deals with image classification, feature

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

extraction, image segmentation, image improvement, image acquisition. Oropharyngeal Cancer Lesions is a significant health issue in the 21st century. This characterizes a huge economic responsibility to health care representatives and authorities. Based on the statistics from the World Health Organization, in world deaths caused by Oropharyngeal Cancer will reach about a few million people in 2030. In recent years, technologies have been developed in the medical field to detect Oropharyngeal Cancer Lesions. Even the patients in the death condition also get cured by development in the medical field. Imaging has become an essential component in the field of biomedical research and clinical practice. It helped doctors in critical surgeries.

By utilizing a variety of image processing techniques, feature extraction, segmentation, and image classification is utilized to identify Oropharyngeal Cancer Lesions using the tongue. Gradient Vector Flow segmentation is providing accurate boundary, region, and pixel-based segmentation in the tongue image. Feature extraction is utilized to extract the geometry, texture, and color from the tongue image. Moreover, the support vector machine, minimum distance, and Bayes classifier are utilized to categorize whether the tongue image is Oropharyngeal cancer or healthy. Oropharyngeal Cancer Lesions have been detected from various techniques utilizing classification and segmentation feature extraction. Especially, there were a significant number of attempts that depend on segmentation to detect Oropharyngeal Cancer Lesions. To better understand detecting Oropharyngeal Cancer Lesions, it is beneficial to examine and analyze the existing systems. Consequently, recent methodologies and procedures in the area of detecting Oropharyngeal Cancer Lesions have been discussed.

Machine learning was a branch of AI technology that employs statistical approaches to enable computers to "train" from expert knowledge. ML techniques employ process strategies to obtain data directly from knowledge rather than relying on a pre-programmed equation. Unsupervised machine learning learns a network on existing output and input information to predict upcoming outcomes, whereas supervised learning teaches a system on unknown output and input information to predict upcoming results. Unsupervised learning, on the other hand, seeks out hidden trends or patterns in data.

In the face of concern, supervised machine learning creates a model that provides assumptions based on argumentation. A supervised learning method analyzes a set of information and activities to generate result data and builds a method to forecast how the system will react to incoming input. To create a predictive model, the supervised approach employs regression and classification algorithms. Individual acts are predicted by regression methods, whereas continuum actions are predicted using classification approaches. Decision trees, SVM, Nave Bayes, K-nearest neighbour, neural networks, and logistic regression are some of the most prominent classification techniques.

Detection of Oropharyngeal Cancer Lesions Using Hybrid Classifiers

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

2. RELATED WORK:

Xingzheng Wang et al. [2013] presented a geometric dispersion level of human tongue colour for analytical extraction of features, as well as describing three tongue colour characteristics. Tongue colour gamut is a technique for predicting a large range of colours. They offer a one-class SVM approach for a colour range descriptor, and they show how to color-allocate specific tongue attributes to ensure efficiency.

Automatic detection of microaneurysm was introduced by AkaraSopharak et al. [2011]. On non-dilated pupil pictures and low-contrast retina, the set of architectural operatives is used to detect the microaneurysm. To improve quality, pre- processing is done first to detect microaneurysms. False detection might be caused by exudates and vessels, as identified in the second phase. Before a microaneurysm may be detected, vessels will be removed from the image. Finally, the microaneurysm was discovered on shaky photos. It does not measure the severity of the condition; instead, this only reveals the microaneurysm.

To identify malignant cells in breast cells, Mahfuzah Mustafa et al. [2014] suggested using a Gradient Linear Flow Snake technique. GVF is used to segment breast tissue in order to find a particular cancer location. To remove superfluous noise from the mammography picture, a Gaussian Low Pass Filter is used. After that, the Snake method connects to the malignant region.

Mehdivatankhah et al. [2014] proposed fully computerised categorization MRI scans of the brain which can determine if a person is sick or well. Wavelet transform is used to retrieve the feature from MRI pictures. Reduction of features PCA is used to reduce the dimensions, which increases the computing complexity and expense. SVM and Cuckoo are used in a hybrid technique to classify if a person was sick or well.

Wenshu Li et al. [2009] predicted a new tongue feature extraction method based on a higher tier set approach. Initially, the tongue's shape was altered in the HSV colour space, and a way for enhancing contrast between the tongue and other parts of the picture was exhibited. The tongue shape contour restriction is described by an energy value between parametric form and the developing curve model in an enhanced thresholding technique. This method yields the desired outcome.

WangmengZuo et al. [2004] suggested integrating the segmentation method with polar edge sensor for computerised tongue separation. Firstly, a polar edge sensor was meant to remove the tongue body's edge efficiently. It is suggested that you use an adaptive threshold edge bithreshold technique. Finally, to separate the mouth body from the tongue picture, an initial and active shape approach are provided.

The SVM classifier was suggested by Bob Zhang et al. [2014] for recognising Oropharyngeal Cancer Lesions and Nonproliferative Oropharyngeal Cancer. Bi-Elliptical Deformable Contour was used to divide the tongue picture. The tongue front picture is used to extract the geometry, structure, and colour properties of the tongue. The SVM classifier may be used to categorise this.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

The major goal of this project is to avoid collecting samples of blood by using the injection approach and to enhance the segmentation technique using the GVF Snake methodology to remove artificial edges.

3. MODULE DESCRIPTION

A thorough description of the dataset and ML methods utilised in this work can be found in this section. To begin the mouth cancer level predictive model, it is necessary to gain a better understanding of medical terminology and processes from dentists. As a result, a discussion with some dentists was held to clarify Oropharyngeal cancer ideas. Recognizing Oropharyngeal Cancer, using Tongue Characteristics (Color, Geometry, and Texture) in Current Work is focused on these 3 ideas. These methods are effective in identifying Oropharyngeal cancer lesions based on tongue characteristics. The proposed study is carried out in order to improve categorization accuracy.

Random Forest

The random forest classifier builds many decision trees after selecting a random subset of data. It's also the most often utilised algorithm for obtaining reliable data. For regression and classification problems, random forest could be employed. It's a machine learning algorithm that's supervised. Random forest has a number of advantages, including the ability to be used for both regression and classification issues, which encompass the majority of contemporary machine learning techniques. It works in a similar way to a decision tree and employs the bagging method. Bagging is the process of combining the creation of models with the improvement of the output outcomes. To forecast the outcome, a random forest includes two or multiple decision trees. Rather than looking for features while partitioning the nodes, it introduces randomness to the trees. This selects the right characteristic among the random subset's features. It might lead to a better design in the end. As a result, Random Forest splits a node into random subgroups of the attributes. Another benefit of random forest was that calculating the relevance of close features in the prediction phase was simple.

K-Nearest Neighbor (KNN)

Because of its simplicity and efficiency, k-NN is an extensively used pattern categorization algorithm. It is acceptable for both small and large datasets. For more difficult situations, it provides accurate findings. KNN is a regression and classification statistical method that is commonly used in industry to solve classification problems. KNN takes into account three factors: ease of interpretation, computation time, and prediction strength. The KNN algorithm was widely used for understanding and calculating large amounts of data in a short amount of time.

Support Vector Machine (SVM)

SVM was ML approach that produces accuracy while consuming minimal processing resources. SVM could be used for categorization as well as regression. Its primary goal, however, is to develop categorization systems. It's done by looking for hyper-planes in a set of attributes that classifies the datasets. Many hyper-planes could be used to distinguish data

ISSN 1533-9211

DOI 10.5281/zenodo.6637102

points. The data points' properties on either side of the hyper-planes belong to separate classes.

Decision Tree

A decision tree was a type of structural diagram that is used to provide answers to problems depending on particular criteria. A Decision Tree was a type of supervised machine learning method that is commonly employed in categorization tasks. A tree can be used in a variety of ways in life, and it is also used in deep learning by encompassing both regression and classification trees, commonly called as CART. A decision tree was a type of flowchart that is structured. Every internal node represents a test on an element, every branch a test outcome, and every leaf of the end node a classifier. The root node was the node at the very top of a tree. In decision processes, a decision tree could be used to describe decisions and decision making.

Multilayer perceptron

MLP is an ANN process that creates a collection of outcomes from a set of input data. An MLP was made up of numerous levels of input nodes connected by a directed network between the output and input layers, with nodes travelling in a single direction. By training the information, a predictive network is constructed using the backpropagation neural method. MLP was a machine learning supervised learning model. This represents a deep learning method because there are numerous layers of neurons. MLP was commonly utilized in voice recognition, machine language translation, and image computing. It is typically utilised for supervised models and neurobiology, but it may also be used for parallel dispersion processing.

Logistic Regression

After linear regression, logistic regression was the most well machine learning computation. Logistic regression and Linear regression are comparable from a variety of perspectives. Regardless, the most significant distinction is in what they are used for. For predicting/forecasting values, linear regression methods are used, however logistic regression can be used for classification activities.

Synthetic Minority over Sampling Technique (SMOTE).

SMOTE is a well-known strategy for constructing a classifier for the imbalance dataset. Uneven distribution of underlying output classes makes up an imbalance dataset. SMOTE is often used to solve classification problems in datasets with imbalances. When dealing with an imbalanced dataset, the SMOTE preprocessing method is regarded one of the most dependable techniques. Since its release, a number of SMOTE variants have been developed and implemented in attempt to improve the present SMOTE technique's reliability and adaptability in various contexts. In the field of machine learning and data mining, SMOTE is regarded as one of the most potent preprocessing approaches. The goal of SMOTE is to do interpolation within data of minority class samples in order to enhance their numbers. It also aids in classification generalisation. In SMOTE, fake instances are used to oversample the

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

minority class. These examples were created using the feature space of the minority class. The number of neighbours is decided based on the requirements for sampling. Using these neighbours, a line is created with the data points from the minority class. When dealing with unbalanced datasets, SMOTE is a particularly successful approach. In training examples, it tends to balance the number of majority and minority class instants.

Tongue Capture Device

A special in-house device is designed to capture the tongue images. A CCD camera with two D65 fluorescent lamps evenly spaced across the lens to provide even illumination. CIE recommends a 45° angle between emergent light and incident light. Patients who want to reveal their tongue to the camera should rest their chin on a chinrest while the image is being captured. The photographs, which range in size from 257 x 189 pixels to 443 x 355 pixels, will be recorded in JPEG format and colour adjusted to remove any irregularities in colour images caused by device reliance and lighting changes. BioMed Chinese Medicine Repository gathered the tongue images..

Preprocessing

The image is enhanced before it is offered as an input to the other procedures during preprocessing. Preprocessing usually entails removing noise, enhancing, dividing sections, and so on. To generate a new illumination value in the output image, preprocessing techniques use a small residential area of a pixel in the input image. Filtration is a term used to describe certain types of preprocessing procedures.

Local Preprocessing methods are divided into two groups based on the processing goal: flattening conceals noise or other tiny fluctuations in the image, which is analogous to the removal of high frequencies in the spectral domain. Leveling also creates sharp edges that provide important information about the image. Gradient operators will be based on the image function's local variations. Gradient operators are used to highlight such areas in an image. In the pre-processing stage, noise is removed and procedures are improved.

Color Correction

Color correction is a technique that uses filters or colour gels in television, cinematography, stage lighting, photography, and other professions. The goal is to adjust the light's overall colour; typically, the light's colour is computed using a colour temperature scale and a green-magenta axis perpendicular to the colour temperature axis.

Without colour correcting gels, a scene could have been a mix of several colours. Color correction gels in the front of light sources can be used to match the colour of the different lighting sources. When exhibited in a theatre or on television, combined lighting can create an unappealing appearance. Gels can also be used to make a scene look more realistic by simulating the natural blend of colour temperatures. This application is beneficial, especially when precise lighting is required. Color gels can also be used to colour light for a more creative impact.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

Segmentation Gradient Vector Flow Snake Technique

The GVF Snake approach is expected to play a key role in tongue image segmentation. The gradient vectors of a binary edge or a Gray level map are dispersed by the GVF Snake method. GVF has the capacity to transfer the deformed model into boundary concavities and has a large picture capture by utilising several dimensional photographs.

Snakes with GVF (Gradient vector flow) are an extension of the well-known active or snakes' contours approach. The difference between GVF and typical snakes is that the finals converge to boundary concavity and do not require modification in close vicinity to the boundary. The new snake v is a two-dimensional dynamic contour parametrically defined as v(s) = [y(s), x(s)], where s G [0, 1] decreases the energy function:

$$E = \int (E_{int}\left(v(s)\right) + E_{image}\left(v(s)\right) + E_{con}\left(v(s)\right)ds$$

On equation 1, the GVF snake extension uses a GVF field as a limitation energy. Pressure forces snakes (balloon snakes), multi-resolution snakes, and distance possibilities are some of the other constraint energy challenges. The image gradient can be used to determine the gradient of a picture in a specific direction. In the proposed effort, the processes involved in identifying Oropharyngeal Cancer are:

- GVF-calculates gradient
- Sobel operator determines gradient by convolving filter mask with a matrix comprising of image pixels.
- Gradient thickness signifies distance (in pixels) among two points, the disparity in intensity which characterizes the value of gradient.
- Iterations are the numbers of iteration done during GVF computation.
- Smoothing factors should be set in accordance with the quantity of noise present within the image. The greater the noise the larger the value Smoothing factor is the normalization factor regulating the exchange among the first and the second integral term.
- Time length is computed for every iteration.

Feature Extraction

Feature extraction is the process of extracting features from the foreground tongue image using color, texture, and geometry. The variety of colors has been

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

extracted by utilizing color feature extraction by means of tongue color gamut. In texture feature, eight blocks were divided with the help of a 2-D Log Gabor filter. Geometry features are calculated for the mathematical formula.

Tongue Color

All potential colors were characterized by tongue color gamut that is displayed on the surface of the tongue and endures within the red limit. The Color Feature Extraction is given in Figure 1

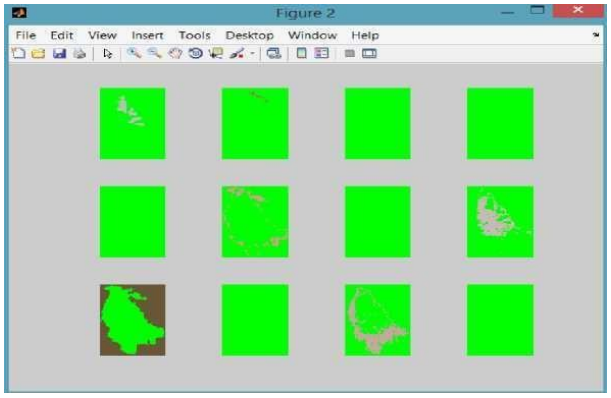


Figure 1 Color Feature Extraction

Tongue Texture

To distinguish the 9 tongue texture features, the 8 blocks of texture values purposely located on the top of the tongue are used, along with the added mean of all 8 blocks. The figure 2 shows how to extract texture characteristics from the tongue.

Regions further than the tongue boundary should be included in big blocks, which would intersect with other blocks.

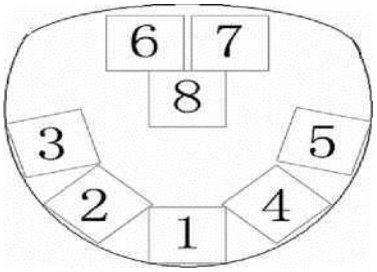


Figure 2 Location of the eight texture blocks on the tongue.

Smaller block sizes will aid to avoid colliding, but they will not be able to cover all eight regions as well. The blocks will be calculated automatically by using a segmented binary tongue front image to determine the tongue's centre. Following that, the tongue's edges are

ISSN 1533-9211

DOI 10.5281/zenodo.6637102

formed, and similar portions are calculated from the tongue's centre to place the eight blocks. Block 1 is at the tip, Blocks 7 and 6 are at the root, Blocks 5 and 6, Blocks 2 and 3, and Block 8 are on both sides, and Block 8 is in the middle. The entire tongue block is then calculated using the 2D Log Gabor Filter.

The Tongue Log Gabor Filter is the technique that is utilized to extract the blocks. Eight blocks positioned on the surface of the tongue to characterize the texture of tongue images. The Texture Feature Extraction is specified in Figure 3.

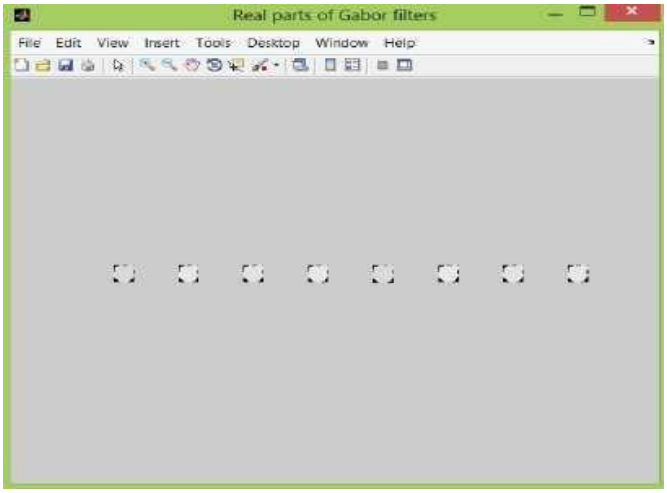


Figure 3 Texture Feature Extractions

Tongue Geometry Features

The geometry features retrieved from tongue photos are defined in the following section. These characteristics are influenced by measurements, regions, ranges, and ratios.

Triangle area ratio (tar) is defined as the percentage of ta to a::

Tar=ta/a

Triangle area (ta) is the area of a triangle that is defined within the foreground of the tongue.

Ratio of square area: The proportion of sa to an is known as the SAR

sar = sa/a

The area of a square described within the front of the tongue using reduced half- distance z is known as square area (sa).

sa = $4z^2$

The ratio of ca to an is known as the circle area ratio (car).

Car=ca/a. The area of a circle within the foreground of the tongue using less significant half-

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

distance z, where r = z, is called circle area (ca).

ca = nr^2

sar = sa/a

The number of pixels in the tongue foreground is used to define the area (a) of a tongue.

The proportion of cd to l is known as the centre distance ratio (cdr).

Cdr=cd/l

Length-Width Ratio: The length-width ratio is a measurement of how long something is compared to how wide it is. The ratio of a tongue's length to its width is called lw.

Iw = l/w

Hybrid Classifier

In this study, a hybrid classifier has been formed by combining three classifiers such as minimum distance classifier and Bayes classifier, support vector machine. The class projections of those classifiers have been merged to enhance the classification precision of the forecasted method.

- The class was classified by a minimum distance classifier which is based on the image features and forecasts the relations among class and features.
- For each class, the Bayes classifier is used to determine the next possible outcome. The forecast's result is the class with the highest posterior probability.
- Using Support vector machine image features are segregated into classed by applying the hyperplane on the data points.

Three classifiers are used exclusively in this study to predict which group the feature belongs to. The results of the hybrid classifier show that the image taken is unhealthy, confirming Oropharyngeal cancer. When all of the classifiers assign an unhealthy class to a feature When both classifiers agree that a taken image's feature belongs in the healthy category, the image is considered healthy.

4. IMPLEMENTATION TOOLS

Datasets

The Oropharyngeal cancer data collection, oversampling approach (SMOTE), and feature selection algorithms utilised in this study are described in this section. BioMed Chinese Medicine Repository gathered the information. Because it was a preprocessed dataset for the repository directly, processing this dataset was a breeze. Despite the fact that the dataset had been preprocessed, it was simple to choose the features for the prediction procedure. Twenty-

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

five attributes were reduced to 14 features using hybrid feature selection. Support Vector Machine is a type of support vector machine. Following that, four classifiers were used to predict the identification of Oropharyngeal cancer patients: K-Nearest Neighbors, Nave Bayes, and Updatable Multilayer Perceptron According to the data, the SVM outperforms other machine learning algorithms after including feature subset selections with SMOTE during preprocessing stages. A hybrid classifier algorithm was evaluated using the initial data collected in this study to detect Oropharyngeal cancer lesions, and features such as colour, texture, and geometry were retrieved.

4.1 Experimental results

In this paper, the existing and proposed work detect Oropharyngeal Cancer Lesions on tongue features using Bi-Elliptical Deformable Contour (BEDC), Support Vector Machine, and Gradient Vector Flow (GVF), Hybrid Classifier. The Procedure of proposed and existing work includes the subsequent steps:

Step 1: The Chinese Medicine Repository's BioMed Central database was used to compile the information.

Step 2: In previous research, BEDC (Bi-Elliptical Deformable Contour) was used to segment tongue pictures. It also distinguishes between pixels in the background and those in the foreground. The image of a patient with Oropharyngeal cancer lesions is used as input. The segmented image using the BEDC approach is shown in Figure 4.

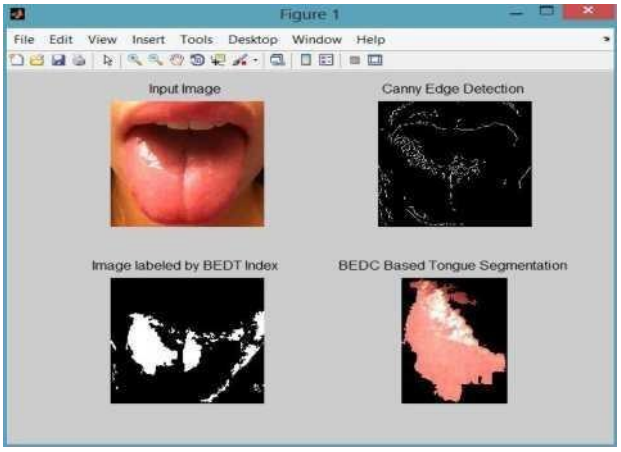


Figure 4 BEDC Segmented Image

Step3: Features such as geometry, texture, and color are extracted after segmentation. The values of characteristics are then computed using standard deviation and mean. Table 1, Table 2, and Table 3 demonstrate the mean shape, texture, and hues of Healthy and Oropharyngeal Cancer, as well as their standard deviations.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

Table 1 Geometry Features for Oropharyngeal Cancer Lesions and Healthy Images (BEDC)

FEATURES	OROPHARYNGEAL LESIONS	HEALTHY
Area	8976	54678
Center distance	79.3	130.87
Center distance ratio	1.2	1.2
Circle area	19874	34567
Circle area ratio	3.2357	2.3456
Length	176	320
Length-width ratio	2.3456	1.234
MEAN	76854	30984
Smaller half distance	67	130.43
Square area	56789	323456
Square area ratio	65784	6.3432
STD	14378	70983
Triangle area	7895	309872
Triangle area ratio	1.8738	1.2345
Width	134	324

Table 2 Color Features for Oropharyngeal Cancer Lesions and Healthy Images (BEDC)

COLORS'	Oropharyngeal Lesions	H
C	1.2345	1
R	1.09723	1.2673
B	1	1
P	1	1
DR	1	1
LR	1.773	1.6773
LP	1	1.5663
LB	1.7553	1.6773
BK	1.863	1.7783
GY	1	1
W	1.234	1.7823
Y	1.5773	1
MEAN	1.367	1.6788
STD	1.6773	1.234

ISSN 1533-9211

DOI 10.5281/zenodo.6637102

Table 3 Texture Features for Oropharyngeal cancer Lesions and Healthy Images (BEDC)

LOCKS	B-1	B-2	B-3	B-4	B-5	B-6	B-7	B-8	MEAN	STD
Oropharyngeal Lesions	6.598	5.997	5.811	5.57	5.346	5.215	5.112	5.002	5.5813	0.5351
H	8.097	7.296	6.934	6.627	6.289	6.148	6.057	5.927	6.6718	0.7411

LOCKS	B-1	B-2	B-3	B-4	B-5	B-6	B-7	B-8	MEAN	STD
Oropharyngeal Lesions	6.598	5.997	5.811	5.570	5.346	5.215	5.112	5.002	5.5813	0.5351
H	8.097	7.296	6.934	6.627	6.289	6.148	6.057	5.927	6.6718	0.7411

The extracted structures are now supplied as input for SVM (Support Vector Machine) classification. The features are linked to the training features, and the result is produced. The result is then classified as Oropharyngeal Cancer or healthy using SVM.

Step 4: The segmentation procedure is used in projected work to segment the image using GVF (Gradient Vector Flow). GVF segments the full tongue image from the provided input. Figures 5 and 6 show the original tongue image that was used for segmentation and the GVF Snake segmented image.

Step 5: Geometry, texture, and colour are extracted from a segmented tongue image. Then, using standard deviation and mean, the values of characteristics are calculated. Tables 4, 5, and 6 demonstrate the mean geometry, texture, and hues of Oropharyngeal Cancer and Healthy, as well as their standard deviations.

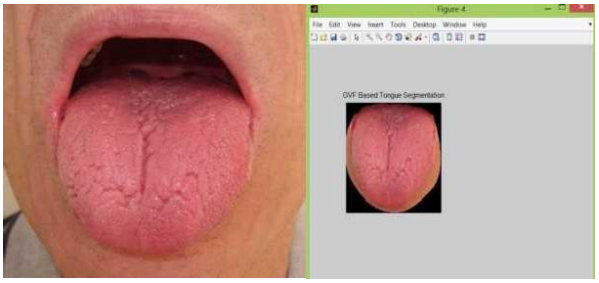


Figure 5 Original Image **Figure 6 GVF Segmented Image**

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

Table 4 Color Features for Oropharyngeal Cancer Lesions and Healthy Images (GVF)

COLORS	C	R	B	P	DR	LR	LP	LB	BK	GY	W	Y	MEA	STD
Oropharyngeal Lesions	0.0094	0.1358	0	0	0.1171	0.3112	0	0..0499	0.2311	0.0001	0.0755	0.0698	0.999	0.101
Healthy	0	0.0298	0	0	0.0014	0.4854	0.002	0.0196	0.1884	0	0.2725	0.0009	0.083	0.1547

Table 5 Texture Features for Oropharyngeal Cancer Lesions and Healthy Images (GVF)

BLOCKS	B-1	B-2	B-3	B-4	B-5	B-6	B-7	B-8	MEAN	STD
Oropharyngeal Lesions	1.926	1.859	1.798	1.769	1.739	1.710	1.671	1.635	1.7633	0.0965
Healthy	2.032	1.917	1.868	1.837	1.794	1.733	1.706	1.660	1.8183	0.1217

Table 6 Geometry features for Oropharyngeal Cancer Lesions and Healthy Images (GVF)

FEATURES	Oropharyngeal Lesions	HEALTHY
Width	302	252
Length	271	251
Length-width ratio	0.8974	0.996
Smaller half-distance	135.5	125.5
Center distance	135.5	125.5
Center distance ratio	0.5	0.5
Area	62932	45682
Circle area	57700	49500
Circle area ratio	0.9161	1.0826
Square area	293764	252004
Square area ratio	4.668	5.5165
Triangle area	40921	23766
Triangle area ratio	0.6502	0.6923
MEAN	35089	29198
STD	81268	69498

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

The image is classified as strong or Oropharyngeal Cancer Lesions using a hybrid classifier (Support Vector Machine, Bayes classifier, and Minimum distance). Based on the projections of three classifiers, the recommended work improves recognition of healthy and Oropharyngeal Cancer Lesions images.

Step 6: The categorization is done on both the planned and existing data sets that have been trained and tested.

Step 7: Finally, the values of Specificity and Sensitivity are computed to determine accuracy.

5. PERFORMANCE EVALUATION

To determine and analyse our suggested approach Gradient Vector Flow to identify Oropharyngeal Cancer Lesions using tongue features, we used a variety of evaluation measures. The total performance of the hybrid classifier is calculated using metric values such as Average accuracy (AC), Specificity (SP), and Sensitivity (SE). The formulas are listed in Table 7. The sensitivity is the percentage of positive cases that are well exposed by the test, while the specificity is the proportion of negative cases that are properly detected by the test. The number of correctly classified experiments determines classification accuracy.

Table 7 Evaluation Measures

Measures	**Formula**
Average Accuracy	(SE+SP)/2
Specificity	SP=TN/(TN+FP)
Sensitivity	SE=TP/(TP+FN)

Evaluation Metrics

In order to assess the accuracy in classification, sensitivity, and specificity values are calculated for Bi-Elliptical Deformable Contour and Gradient Vector Flow.

Sensitivity is the quality or condition of being sensitive. Specificity is the ability of the test to correctly identify those without the disease. Specificity is the quantity or state of being specific. Sensitivity is the ability of a test to properly identify those with the disease. In simple terms, high sensitivity implies that an algorithm returned more relevant results.

Confusion Matrix

The confusion matrix is evaluated in order to make a judgement that the classifier might make. Table 8 shows a confusion matrix.

Where,

- TN (True Negative) is the number of properly classified healthy people.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

- The number of Oropharyngeal Cancer Lesions misclassified as healthy is expressed as FP (False Positive).

- FN (False Negative) is the number of healthy people who have been misclassified as Oropharyngeal Cancer Lesions, while TP (True Positive) denotes the number of Oropharyngeal Cancer Lesions that have been correctly classified.

Table 8 Confusion matrix

TYPES		**PREDICTED**	
		Healthy	**Oropharyngeal Cancer Lesions**
True	**Healthy**	TP	FP
	Oropharyngeal Cancer Lesions	FN	TN

The confusion matrix for the existing algorithms (BEDC, Feature Extraction (Color, Texture and Geometry) and SVM) and proposed algorithm (GVF Snake Technique, Feature Extraction (Geometry, Texture and Color) and Hybrid Classifier (SVM, Minimum Distance and Bayes Classifier) are given below. For 10 images, based on the Predicted records and True records, the FN, FP, TN, TP values are calculated. Table 9 depicts the confusion matrix for the proposed work, using the GVF snake technique. Table 10 depicts the confusion matrix for the existing work, BEDC

Table 9 confusion matrix for the proposed work, using the GVF snake technique.

METHOD	**ACTUAL**	**PREDICTED**	
		DM(Positive)	Healthy(Negative)
Deep learning	Oropharyngeal Lesions (Positive)	9(TP)	1(FP)
	Healthy(Negative)	3(FN)	7(TN)

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

Table 10 Confusion Matrix for BEDC

TECHNIQUE	ACTUAL	PREDICTED	
		DM(Positive)	Healthy(Negative)
Hybrid classifier	Oropharyngeal Lesions (Positive)	7(TP)	3(FP)
	Healthy(Negative)	2(FN)	8(TN)

Result Analysis

Machine Learning methods such as SVM,Decision tree, KNN, Random forest,Naive Bayes,MLP, and Logistic regression were coded in Python to create the prediction tool. As a result, the Oropharyngeal data estimation result is presented in Figure 7.

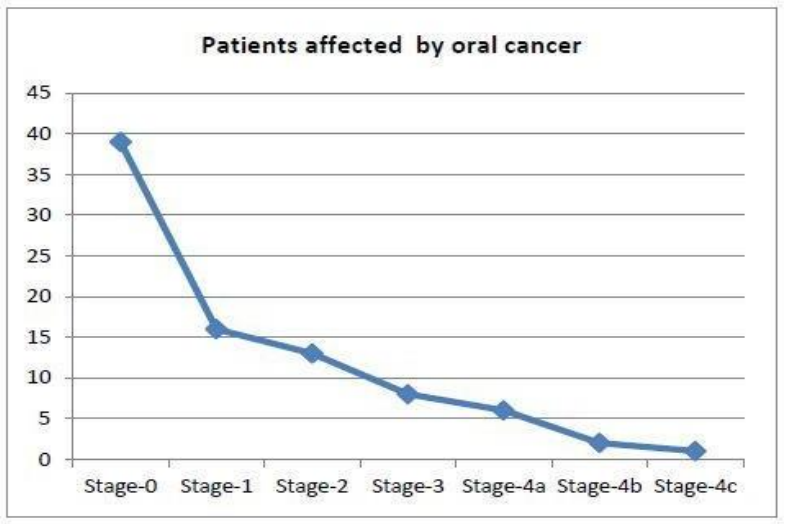


Figure 7 Prediction result of dataset.

By using three layers of cross-validation on efficiency, the prediction outcome was analysed for accuracy outcomes of several machine learning techniques. In percentage, Table 11 explains the 3 folds of cross-validation findings. The accuracy results of several machine learning techniques are shown in fig 9.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6637102

Table 11: Cross-Validation of Different Machine Learning Algorithms in Three Folds

Classifiers	Fold-1(%)	Fold-2(%)	Fold-3(%)
Decision Tree	85.567	85.567	84.876
SVM	82.759	82.759	82.553
KNN	82.759	82.759	80.255
Random Forest	79.310	79.310	81.404
Logistic reg.	68.966	68.966	59.236
MLP	75.862	75.862	69.704

The percentage of 3 folds of cross-validations is shown in table 11. This makes it easier to figure out how accurate certain algorithms are.

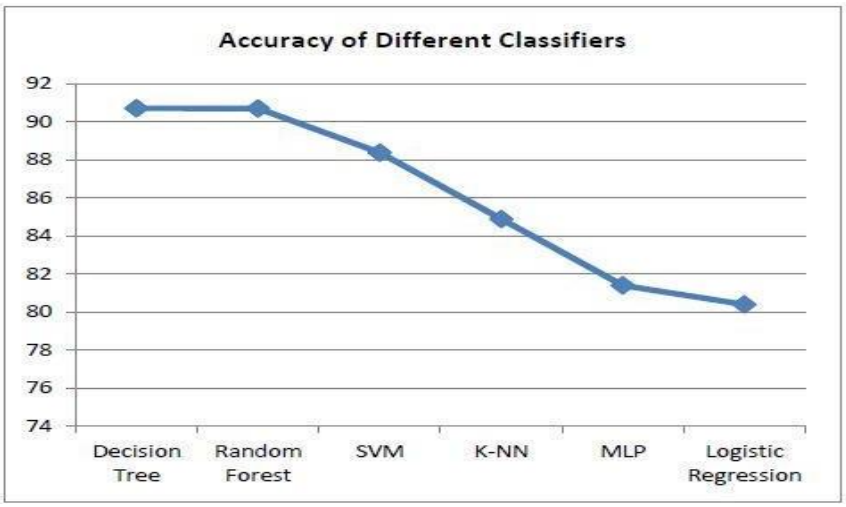
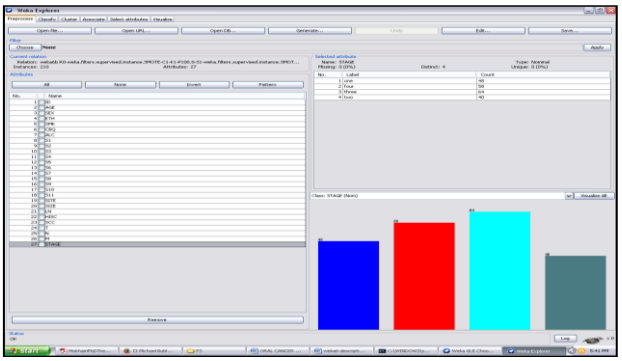


Figure 8. Different classifiers' accuracy results

The graphical portrayal of various average accuracy of distinct ML methodologies is shown in Figure 8, where the Decision Tree is 90.688 percent accurate, Random Forest is 91 percent accurate, SVM is 88 percent accurate, KNN is 85 percent percent accurate, MLP is 81 percent accurate, and Logistic Regression is 80 percent accurate. It is evident that the accuracy rates of the Decision Tree and Random Forest methods are identical.

ISSN 1533-9211

DOI 10.5281/zenodo.6555875



Evaluation Methods

Balance Data set

The original OC data set were categorized into four classes. There were 58 instances of the majority class (stage four), 16 for stage three and stage one and two falls under the category of minority class with the number of instances less than 10. In this study, for the training set 10-fold cross-validation is used. The minority class is over-sampled at 100%, 200%, 300%, and 400% of its original size. Table 12 shows the result of re-sample an imbalance OC data set using SMOTE. The result after over sampling showed the number of instances is a re-sample to 210instead of 82 instances

Table 12 Balanced class distribution for OC by applying SMOTE

Class Name	# of Instances	%	# of Instances with SMOTE	%
One	3	3.66	48	22.86
Two	5	6.09	40	19.05
Three	16	19.51	64	30.48
Four	58	70.73	58	27.62
Total	82		210	

Table 12 shows the class distribution of each minority class ofOC data set, stage one (22.86%) and two (19.05%) are almost balance as majority class, stage three (30.48%) and four (27.62%). Figure 9 shows the balance OC data set using SMOTE in WEKA software

Figure 9: Balance OC data set using SMOTE in WEKA software

ISSN 1533-9211

DOI 10.5281/zenodo.6555875

Optimum Features Selected

After loading the data set, the FS algorithms are applied to find the most significant features of the data set. It started with all features selected (FS0), cfsSubSetEval with Best First Forward (FS1), InfoGain Variable Evaluator combined Sequential Backward Selection or known as Linear Forward Selection with Floating Forward Selection (IGSBFS) (FS2), Correlation Variable Evaluator with Ranker (FS3), and hybrid FS3 with CfsSubset Evaluator with Linear Forward Selection (FS4). Table 13 shows the details of results for each FS method.

Table 13 Selected Attributes with Features Selection methods

FS	**Method**	**Selected attributes**
FS0	No selected feature	25 attributes
FS1	Cfs SubSet Eval Best First Forward	2,3,8,9,15,16,17,18,19,20,21,2 2,23,24 (14 attributes)
FS2	Correlation Attribute Eval Ranker	20,23,21,22,16,19,24,8,2,15,7,17,3,18,5,1,13,9,11,6,25,10,14,4,12 (25 attributes). Remove 11 attributes
FS3	Cfs Subset Eval Linear Forward Selection(forward)	20,23,21,22,16,19,24,8,2,15,17,3,18,9 (14 attributes)

The experiment of FS using WEKA software started with 25 features and 210 instances. It ended at FS4 with 14 optimal features namely 2, 3, 8, 9, 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24.

Accuracy Classification Performance

The performance measure of accuracy is considered in order to evaluate the efficiency of the FS methods. The measures are compiled by the following unit: Classification Accuracy (%) = (TP+TN) / (TP + FP + FN +TN). In this study, the evaluations are conducted in WEKA with 10 fold cross validation. Four different machine learning algorithms are used to classify the OC data set with four FS methods:

- Updateable Naive Bayes (NB). This is the updateable version of Naïve Bayes and using

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6555875

estimator classes.

- Multilayer Perceptron (MLP). A Classifier that uses backpropagation network to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time.
- SMO-Poly Kernel (E-1.0) (SVM). This implementation globally replaces all missing values and transforms nominal variables into binary ones. It also normalizes all features by default.
- K-Nearest neighbors classifier (lazy.IBk). K-nearest neighbors classifier can select appropriate value of K based on cross-validation. It can also do the distance weighting.

Table 14 shows the result for the classifier without oversampling method, SMOTE. It started with select all features of OC data set, 25 features. Next feature selection phase, FS2 is also carrying on with 25 features. Finally, a classifier with 14 selected features from FS3 is generated. Using oversampling (SMOTE), the results for three FS methods with four classifiers show that the features selected by the integrated diagnostic model contributed to improved accuracy of the entire classification algorithm used for the OC data sets.

Table 15 demonstrates that FS with SMOTE outperforms FS without the implementation of SMOTE. The accuracy of OC data set for FS3 improves from 87.80% to 94.76% for NB,90.24% to 95.24% for MLP, 86.59% to 96.20% for

SVM and 76.83% to 91.43% for KNN. Findings from Table 16 are also shown that the highest classification accuracy performance using SVM algorithm, with accuracy of 96.19% with 14 optimal features selection namely 2, 3, 8, 9,15, 16, 17,

18, 19, 20, 21, 22, 23 and 24. The empirical comparison between fives FS methods for the entire classifier algorithm is as well performed as graph comparison as Fig

10. It shows the optimal features set from FS3 contribute the highest accuracy performance.

TABLE 14. Performance Accuracy for Three Selected Features Selection On Oc Data Set Without Smote

Classification Accuracy Without SMOTE (%)			
Algorithm	FS0	FS2	FS3
NB	85.37 14.63	75.61 24.39	87.80 12.20
MLP	76.83 23.17	79.27 20.73	90.24 9.76
SVM	62.20 37.80	62.20 37.80	86.59 13.41
KNN	75.61 24.39	75.61 24.39	76.83 23.17

TABLE 15. Performance Accuracy for Three Selected Features Selection On Oc Data Set With

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6555875

Smote

Classification Accuracy Without SMOTE (%)			
Algorithm	FS0	FS2	FS3
NB	91.90 8.10	91.91 8.10	94.76 5.24
MLP	94.29 5.71	93.81 6.19	95.24 4.76
SVM	93.33 6.67	93.33 6.67	96.20 3.80
KNN	86.19 13.81	86.19 13.81	91.43 8.57

TABLE 16. Performance Accuracy for Five Features Selections On Oc Data Set

Algorithm	No. of features	Accuracy (%)			
		NB	MLP	SVM	KNN
FS0	25	91.90	94.23	93.33	86.19
FS1	14	94.76	94.76	92.38	90.95
FS2	25	91.90	93.81	93.33	86.19
FS3	14	94.24	95.24	96.19	91.43
FS4	14	94.76	94.76	92.38	90.95

ISSN 1533-9211

DOI 10.5281/zenodo.6555875

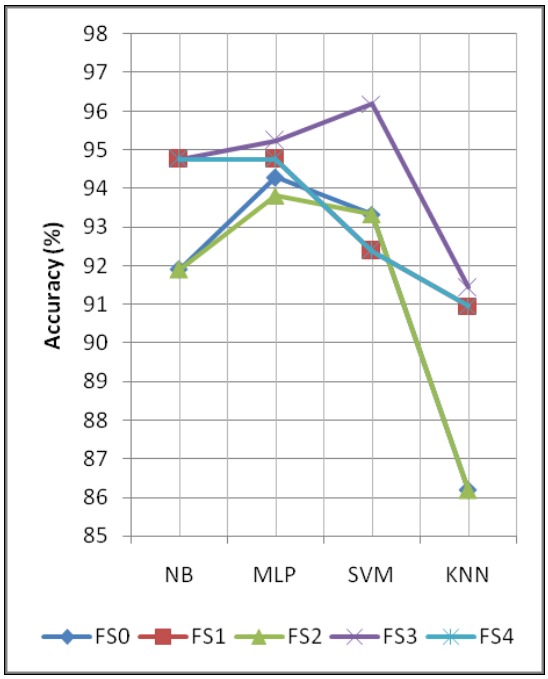


Fig. 10. Performance accuracy comparison between the fives features selection methods with NB, MLP, SVM and KNN algorithm.

CONCLUSION

The main purpose of performing this analysis is to identify the Oropharyngeal Cancer Lesions affected region in the tongue images. Image processing is a safe and time- consuming tool to detect Oropharyngeal Cancer Lesions in an effective and precise manner. In this research, Oropharyngeal Cancer Lesions is noticed by means of advanced methods of image processing like Hybrid classifier, Feature extraction and Gradient vector flow algorithm which select the images in partly. The current work utilized the GVF algorithm to detect Oropharyngeal Cancer Lesions using features involved in tongue images. Feature extraction and hybrid classifier methods are involved in this research for detecting Oropharyngeal Cancer Lesions. The main work of this research is to detect Oropharyngeal Cancer Lesions accurately using tongue features. By using a hybrid classifier all the tongue images are get classified and provide the result appropriately.

REFERENCES

[1] Bob Zhang , B. V. K. Vijaya Kumar and David Zhang, ''Detecting Oropharyngeal Cancer Lesions and Non-proliferative Oropharyngeal Cancer Using Tongue Color, Texture, and Geometry Features", IEEE Transactions on Bio-Medical Engineering, vol. 61, no.2, Feb-2014.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6555875

[2] W. Zuo, K. Wang, D. Zhang, and H. Zhang, "Combination of Polar Edge Detection and Active Contour Model for Automated Tongue Segmentation," in Proceedings of the Third International Conference on Image and Graphics, pp. 270-273, 2004.

[3] W. Li, S. Hu, H.Li, and S. Wang, "A novel segmentation of tongue image," International Journal of Functional Informatics and Personalized Medicine, vol. 2, no.3, pp.315-324, 2009.

[4] Mehdi vatankhah, ImanAttarzadeh,"Proposing an Efficient Method to Classify MRI Images Based on Data Mining Techniques",International journal of Computer Science &Network Solutions,Volume 2, No.8, Aug.2014. (http://www.ijcsns.com)

[5] Mahfuzah Mustafa, NurAzwa Omar Rashid, and RosdiyanaSamad, "Breast Cancer Segmentation Based on GVF Snake," IEEE Conference on Bio-Medical Engineering and Sciences 8-10 Dec 2014.

[6] AkaraSopharak, BunyaritUyyanonvara, Sarah Barman and Tom Williamson," Automatic Microaneurysm Detection from Non-dilated Oropharyngeal Cancer Retinal Images", Proceedings of the World Congress on Engineering, Vol II, July2011.

[7] Bo Pang and David Zhang," The Bi-Elliptical Deformable Contour and its application to automated tongue segmentation", IEEE Trans.Med.Imag., vol.24, no.8, pp 946- 956, Aug- 2005.

[8] Xingzheng Wang, Bob Zhang, Zhimin Yang, Haoqian Wang, and David Zhang, "Statistical Analysis of Tongue Images for Feature Extraction and Diagnostics",IEEE Transactions On Image Processing, Vol. 22, No. 12, December 2013.

[9] Kainuma, Mosaburo, et al. "The association between objective tongue color and endoscopic findings: results from the Kyushu and Okinawa population study (KOPS)." *BMC complementary and alternative medicine* 15.1 (2015): 372.

[10] Kawanabe, Tadaaki, et al. "Quantification of tongue colour using machine learning in Kampo medicine." *European Journal of Integrative Medicine* 8.6 (2016): 932-941.

[11] Meng, Dan, et al. "Tongue images classification based on constrained high dispersal network." *Evidence-Based Complementary and Alternative Medicine* 2017 (2017).

[12] Liu, Zhi, et al. "Automated tongue segmentation in hyperspectral images for medicine." *Applied Optics* 46.34 (2007): 8328-8334.

[13] Elnakib, Ahmed, et al. "Medical image segmentation: a brief survey." *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*. Springer, New York, NY, 2011. 1-39.

[14] Huang, Bo, et al. "Tongue shape classification by geometric features." *Information Sciences* 180.2 (2010): 312-324.

[15] Zhang, Bob, BVK Vijaya Kumar, and David Zhang. "Detecting diabetes mellitus and nonproliferative diabetic retinopathy using tongue color, texture, and geometry features." *IEEE transactions on biomedical engineering* 61.2 (2013): 491-501.

[16] Kim, J., et al. "A digital tongue imaging system for tongue coating evaluation in patients with Oropharyngeal malodour." *Oropharyngeal Diseases* 15.8 (2009): 565-569.

[17] Kainuma, Mosaburo, et al. "The association between objective tongue color and endoscopic findings: results from the Kyushu and Okinawa population study (KOPS)." *BMC complementary and alternative medicine* 15.1 (2015): 372.