Laryngeal Cancer Prediction via VGG16-Based Feature Extraction and Machine Learning on Narrow-Band Imaging Data

Zahraa M. Alsaif[1], Maalim A. Aljabery[1]
[1]Faculty of Computer Science & Information Technology, Computer Science Dept, University of Basrah, Basrah, Iraq
E-mail: zahraa.mahmoud@uobasrah.edu.iq, maalim.aljabery@uobasrah.edu.iq

Keywords: Laryngeal cancer, gaussian blurring technique, data augmentation, machine learning algorithms, deep learning

Received: September 14, 2025

Laryngeal cancer is a serious type of cancer that poses a significant threat to patients' lives if not detected early, due to its impact on the respiratory system. This study aims to improve an intelligent automated model capable of predicting laryngeal cancer using machine learning and deep learning algorithms. A dataset of narrowband medical images of laryngeal tissues, obtained from the Kaggle website, was obtained and divided into a 60% training set (792 images, 198 images per category), a 20% validation set (264 images, 66 images per category), and a 20% test set (264 images, 66 images per category). The dataset underwent several preprocessing steps, including the application of Gaussian filters to remove noise from the medical images, conversion from RGB to YCBCR for image enhancement using CLAHE, and the application of a sharpening filter to increase the clarity of image details and edges. Data augmentation was also used to diversify and expand the data. The VGG16 model was used to extract deep features from medical images, followed by a performance optimization phase using ANOVA-F to select the best features and a network search to fine-tune hyperparameters, before predicting the outcomes. Subsequently, machine learning algorithms KNN, SVM, and LR, as well as the deep learning model VGG16, were trained on the features extracted from the VGG16 model (256 features per image for the VGG16 classifier, 150 features per image for the KNN and SVM classifiers, and 512 features per image for the LR classifier). Finally, the results showed that the VGG16 model achieved superior performance, with an accuracy of 0.96, a resolution of 0.96, a recall of 0.96, and an F1 score of 0.96, outperforming other machine learning models, where the accuracy of the LR, SVM, and KNN models was 0.93, 0.92, and 0.90, respectively.

Povzetek: Raziskava je pokazala, da lahko modeli strojnega in globokega učenja učinkovito prepoznavajo rak grla na medicinskih slikah, pri čemer je bil model VGG16 najuspešnejši.

1 Introduction

Cancer is one of the most dangerous diseases that grows abnormal cells and spreads stubbornly, destroying the body's tissues. There are multiple treatment modalities for advanced disease, including chemotherapy, radiation therapy, and surgery in some cases, followed by adjuvant radiation therapy with or without concurrent chemotherapy [1].

Laryngeal cancer is one of the most common tumors of the neck and head, representing approximately one-quarter to one-third of neck and head tumors and 2% of all cancers in the body. Generally, laryngeal cancers grow in the vocal folds and primarily impact men [2]. Smoking, alcohol, and air pollution are the primary risk factors [3].

By measuring the effect of diet on the development of laryngeal cancer, researchers found that patients with nutritional and vitamin deficiencies are more likely to develop the disease. Most laryngeal cancers are squamous cell carcinomas. One of the main symptoms is the hoarseness of voice in patients with the disease. If mishandled, it may lead to the development of disease and may be related to airway blockage, pain, and dysphagia [2].

Using the following current diagnostic methods: indirect laryngoscopy, direct laryngoscopy, biopsy, and imaging tests, one or more of which may be performed on the patient to search for swelling, masses, or other unusual tumors, the doctor examines the neck and throat area [4].

1.1 Machine learning (ML) and deep learning (DL)

ML is the science that teaches computers new tasks and information and performs them independently without human intervention. It is a part of Artificial Intelligence (AI). Thanks to its ability to efficiently process large amounts of data, which is its main component, and it is ability to make decisions faster and more accurately compared to traditional methods, it has gained increasing popularity, especially in recent years [5].

Common diseases can be identified and diagnosed at their early stages by predicting them using ML. Recently, predictive analytics models have gained significant and pivotal importance in medicine due to the increasing volume of data in the healthcare sector. In healthcare, predictive analytics is a fundamental requirement for this sector because it significantly impacts the accuracy of disease predictions, such as laryngeal cancer, saving lives when accurate and timely predictions are made [6].

There is no doubt that DL is a widely used part of ML. It is commonly used in image analysis across various fields, including medicine. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) use radio waves and are essential medical imaging technologies for diagnosing diseases, including laryngeal cancer. DL enables doctors to diagnose diseases and provide better treatment for patients [7].

This study aims to develop a methodology for predicting laryngeal cancer, determining its progression, and selecting appropriate treatment through narrow-field imaging of laryngeal tissues, relying on deep learning and machine learning. The study utilizes the VGG16 model as a feature extractor from laryngeal images and then uses these features to build predictive models based on VGG16 itself as a classifier, along with machine learning algorithms SVM, KNN, and LR. Furthermore, ANOVA F-test and network search were employed to optimize the performance of the SVM, KNN, and LR classifiers.

2 Related works

Some relevant research has used ML and DL approaches to diagnose laryngeal cancer.

In 2024, Alzakari et al. [8] offered an automated laryngeal cancer diagnosis system by the dandelion optimizer algorithm with ensemble learning (LCD-DOAEL) on medical throat images. The Gaussian filtering technique was used to eliminate noise from medical throat images. The MobileNetv2 model was used to extract features from throat images, while the DOA model was used to get better model performance by tuning the MobileNetv2 hyperparameters. Finally, the ensemble models: Bidirectional Long Short-Term Memory (BiLSTM), Back Propagation Neural Network (BPNN), and Regularized Extreme Learning Machine (ELM) were used for classification with an achieving an accuracy of 95.07%, an F1 score of 92.4%, and an area under the curve of 94.9%, with an accuracy of 95.07%, F1-score 92.4%, and AUC 94.9%.

Although the study used the MobileNetv2 model for feature extraction, it did not specify the number of features extracted from the images. It lacked data preprocessing, relying solely on a Gaussian filter, especially given the limited dataset, and failed to employ any data augmentation techniques. Consequently, the accuracy demonstrated by the model may be exaggerated due to the possibility that the model learned specific details from the samples rather than generalizing them.

In 2023, Nuzaiha et al. [9] present an Automated Laryngeal Cancer Detection and Classification using a Dwarf Mongoose Optimization algorithm with DL techniques (ALCAD-DMODL) to detect the existence of LCA by using the DL model. Eliminate noise from medical images by the Median filtering. Furthermore, the EfficientNet B0 model was used as the feature extractor to extract features from medical images. The Dwarf Mongoose Optimization (DMO) algorithm was used to tune the hyperparameters of the EfficientNet B0 model, which was used to select the best features. A Multi-head Bidirectional Gated Recurrent Unit (MBGRU) model was applied to the classification of laryngeal cancer, with an accuracy of 93.5%, F1-score 93.5%, and AUC 95.6%.

Although the study used the EfficientNet-B0 model for feature extraction, it did not specify the number of features extracted from the images. It also lacked data preprocessing, relying solely on eliminating noise from the medical images by the median filtering, particularly given the limited dataset, and did not employ any data augmentation techniques. Consequently, the accuracy demonstrated by the model may be exaggerated, as it may have learned specific details from the samples rather than generalizing them.

In 2023, Alrowais et al. [10] present a laryngeal cancer detection and classification depending on the histopathological neck area images. The Inceptionv3 model has been used as a feature extractor procedure from medical images. Furthermore, the Aquila optimization algorithm (AOA) has been used to tuning hyperparameters were tuned using to improve the performance of a deep belief network (DBN) model used for laryngeal cancer detection and classification, therefore enhance the detection rate, achieving an accuracy of 92.4%, a precision of 92.1%, a recall of 91.8%, and an F1 score of 91.8%.

Although the study used the Inceptionv3 model for feature extraction, it did not specify the number of features extracted from the images. It also lacked traditional data preprocessing, focusing instead on improving model performance through architecture tuning without employing any data augmentation techniques, particularly given the limited dataset. Consequently, the accuracy demonstrated by the model may be exaggerated, as it may have extracted specific details from the samples rather than generalizing them, increasing the risk of overfitting.

In 2024, Li et al. [11] designed a predictive model using ML algorithms that was used to predict the 5-year survival of laryngeal squamous cell carcinoma (LSCC) patients. This was done using clinical data (such as age, gender, smoking, etc.). Preprocessing in this study was limited to removing incomplete cases and direct data encoding. The chi-square test and Spearman's correlation coefficient have been used to analyze the relationship between data characteristics. These ML algorithms were used to diagnose RF, LR, KNN, DT, AdaBoost, XGBoost, and SVM. According to the outcomes, the SVM algorithm achieves a better accuracy of 85% and an AUC of 81.2% compared to all other approaches.

Despite the use of appropriate preprocessing of the clinical dataset used in this study, the results did not show high performance, in addition to the limited dataset used, which can cause overfitting regardless of the type of data.

Table 1 summarizes an overview of the related works briefly.

Table 1: Summary of the related work

Ref.	Year	Method	Dataset Size / Type	Preprocessing	Classifiers	Accuracy	Notable Contributions
[8]	2024	Used the MobileNetv2 model for feature extraction and ensemble learning for classification.	1320 / Images.	Gaussian filtering-GF.	Ensemble learning (BiLSTM, BPNN, Regularized ELM).	95.07%.	Use of DOA with ensemble learning for enhanced laryngeal cancer image classification.
[9]	2023	Used the Efficient-B0 model for feature extraction. The Multi-Head Bidirectional Gated Recurrent Unit (MBGRU) for classification.	1320 / Images.	Noise removal by Median filtering.	The Multi-Head Bidirectional Gated Recurrent Unit (MBGRU).	93.5%.	Use of DMO with a deep learning classifier for accurate LC detection.
[10]	2023	Used InceptionV3 for feature extraction. Deep Belief Network (DBN) for classification.	1320 / Images.	Not mentioned.	Deep Belief Network (DBN).	92.04%.	Use of AOA has been used to tune hyperparameters to improve the performance of a DBN model.
[11]	2024	The chi-square test and Spearman's correlation coefficient to analyze the relationship between data characteristics. RF, LR, KNN, DT, AdaBoost, XGBoost, and SVM for prediction.	150 / Clinical dataset.	Removing incomplete cases and direct coding.	SVM.	85%.	Developed a web-based prediction platform using clinical data to estimate 5-year survival rates for LSCC patients.

3 Methodology

A comprehensive methodology was adopted to implement this research, starting with the data processing and quality improvement phase, utilizing data augmentation techniques in model training, feature extraction, and culminating in the prediction process using a carefully designed model. The work was divided into several interconnected phases to ensure the highest possible classification accuracy, with precision considered at every step of the implementation. Figure 1 illustrates the principal steps of the proposed system for predicting the stages of Laryngeal Cancer.

3.1 Dataset description

The LC (Laryngeal Cancer) dataset was obtained from the Kaggle website [12]. It is an online community and platform for data scientists and machine learning practitioners under Google LLC, serving as a massive repository for publicly available datasets. It contains 1,320 images of early-stage cancerous and benign laryngeal lesions tissue. The images (100×100 pixels) were manually generated from 33 narrow-band laryngoscopy data images collected from 33 patients with laryngeal squamous cell carcinoma (diagnosed by histopathological laboratory examination). Table 2 provides detailed information about this dataset.

Table 2: Dataset details

Name	Class	No. of Instances
Health Tissue	He	330
Hypertrophic Blood Vessels	Hbv	330
Leukoplakia	Le	330
Abnormal IPCL-like Vessel	IPCL	330
Total No. of Instances		1320

3.2 Preprocessing

Image preprocessing plays a crucial role in enhancing the quality and clarity of medical images before they are fed into ML models. It helps highlight essential features, reduce noise, and enhance contrast, thereby enabling more accurate predictions. In our proposed system, the data is reduced to noise and enhanced contrast, thereby enabling more accurate predictions. In our proposed system, the data pipeline includes a set of enhancement and augmentation techniques aimed at refining the input data. The preprocessing stages in methodology are as follows:

3.2.1 Splitting the dataset

The dataset was divided into two distinct sets: a training set, including a validation set, and a test set. Classifiers were trained using the training set, which comprised 80% of the data (the training set of 792 images, representing 60% of the data, and the validation set of 264 images, representing 20% of the data). Model performance and effectiveness were evaluated on new, previously unfamiliar data using the test set, which comprised 264 images (20% of the data).

3.2.2 Gaussian blurring technique

This function serves to reduce image noise by applying the Gaussian filter, which introduces a blurring effect. The blurring process helps retain the image's low spatial frequency components while eliminating high-frequency noise and irrelevant fine details. The Gaussian blur is mathematically represented in Equation (1), where "x" denotes the direction of filtering and σ represents the standard deviation of the Gaussian distribution. The implementation typically uses a two-pass filtering method applied separately in horizontal and vertical directions, which minimizes computational complexity and is thus often favored in practice [13]. In this work, a Gaussian filter was applied using the cv2.GaussianBlur() function from the OpenCV library, with a kernel size of 3×3 and a standard deviation (σ) of 0. The theoretical basis of the Gaussian filter is defined mathematically in Equation (1), where the Gaussian function is expressed as:

$$g(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \qquad (1)$$

However, in the actual implementation, the filter was not computed manually but rather through OpenCV's optimized built-in function.

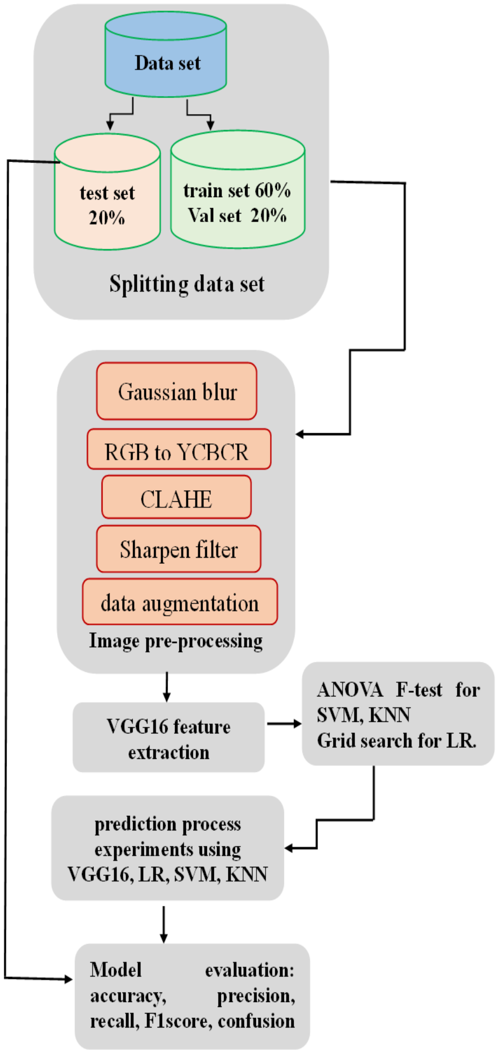


Figure 1: The proposed system for the prediction of laryngeal cancer.

3.2.3 Convert color space from RGB to ycbcr

The YCBCR color space is selected due to its proven efficiency in medical image fusion, as demonstrated in prior studies. The ability to separate luminance (Y) from chrominance components (Cb and Cr) enables the application of independent image processing techniques on each channel. This separation significantly contributes to improving the visual quality of the fused medical images [14]. In this step, cv2.cvtColor() from the OpenCV library was used for color space conversion to convert the image from BGR to YCrCb, thereby separating the luminance component (Y) from the chrominance components (CR, CB). This enables localized contrast enhancement on the Y channel using CLAHE. Fig. (1), shown as an image in RGB space, and Fig. (2), shown as components of YCBCR space.

Figure 2: Image in RGB space.

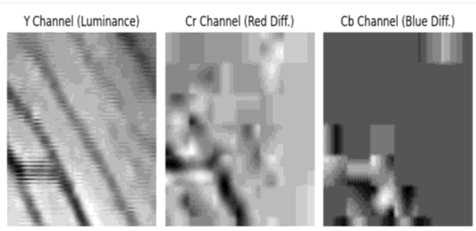


Figure 3: Components of ycbcr space.

3.2.4 Contrast limited adaptive histogram equalization (clahe)

CLAHE (Contrast Limited Adaptive Histogram Equalization) is a widely adopted technique for image enhancement that operates on histogram-based processing. It works under the assumption that local regions within an image should maintain consistency. Instead of applying a single global contrast enhancement, CLAHE divides the image into small tiles and enhances each region individually using localized histogram equalization. This results in improved contrast while avoiding over-amplification of noise, especially in medical images, where detail preservation is critical [15]. In this step, cv2.createCLAHE() from the OpenCV library was used to enhance the Y channel using CLAHE with clipLimit = 2 and tileGridSize = (8, 8), without manipulating the chrominance components CR and CB.

3.2.5 Sharpening filter

A sharpening filter was applied to enhance image clarity by emphasizing fine details and edges, which is particularly important in medical imaging. This technique is useful for improving blurry or low-resolution images by increasing the contrast at edges, where significant transitions in intensity occur [16]. Sharpening used: lied computationally using the OpenCV library with a special convolutional filter to enhance edges, where the following kernel was used: [[0, -0.5, 0],[-0.5, 3, -0.5],[0, -0.5, 0]], to enhance fine details such as tumor borders or tissue differentiation in medical images.

3.2.6 Data augmentation

Data augmentation refers to artificially increasing the size and diversity of the training dataset by applying transformations such as rotation, flipping, scaling, and shifting to the existing images. This technique helps improve the model’s generalization capability by exposing it to a wider variety of input scenarios. In medical imaging, especially with limited datasets like in laryngeal cancer classification, data augmentation is essential to reduce overfitting, enhance model robustness, and improve overall prediction accuracy by ensuring the model does not rely too heavily on specific patterns found in a small set of samples [17]. In this section, the operations flip, shear, and zoom were used to increase the data. This section was used only in the VGG16 classifier.

3.3 VGG16 feature extractor

Feature extraction represents the maximum amount of relevant information, and the image provides a complete description of a pest. Therefore, the selection of features and characteristics for objects is crucial, as they play a key role in improving the model's ability to distinguish between different classes. Objects and images are among the most prominent elements analyzed by feature extraction methodologies to extract the most notable features [18]. Features work as inputs to classifiers; in this study, the pre-trained VGG16 model was used to extract features because it has a simple and easy-to-understand architecture, which makes it easy to follow the data flow and works well as a feature extractor for limited data, compared to other models such as ResNet and EfficientNet model, which are more complex and require fine-tuning of limited data. In the VGG16 model, after passing the (224×224×3) image through 13 Conv2D layers and 5 MaxPooling layers, the data is transferred to the Global Average Pooling (GAP) layer. This layer converts the feature maps generated by the last Maxpooling layer from a 3D model into a 1D (vector) of length 512, representing a summary of the image's prominent features. This vector is then passed to the fully connected (fc1) layer, which carries the ReLU activation and serves as the primary source for extracting important features from the images. The results are then passed to the fully connected (fc2) layer, which reduces the number of features to 256, thereby lowering the dimensions and reducing computational complexity while maintaining effective content representation. Finally, the features are passed to the final fully connected (fc3) layer for classification. Although the VGG16 model was trained on a large database of regular ImageNet images, it remained effective when trained on narrow-band laryngeal images because the features it learns from the first layers (such as edges and basic shapes) are very general and useful for detecting patterns in different images, even if they are narrow bands. Therefore, the model can extract useful features from laryngeal images despite the different bands.

3.4 Best features selection and grid-search

The main goal of selecting the best features and using grid search is to improve the model's performance and thus improve the accuracy of diagnostic predictions.

Feature selection involves selecting features that are more relevant than the rest of the features within the larger set of original features. The benefits of feature selection include primarily improving predictive performance, reducing the number of features, and thus reducing the computation time of the prediction model, and improving data quality. The ANOVA F-test is a type of F-statistic that works by comparing the input feature to the target feature [19]. To select the best features, we used an ANOVA F-test to select the best 150 features out of 512 to reduce dimensions and improve the performance of SVM and KNN classifiers. Choosing the 150 features was done after the experiment, as this number showed a good balance between reducing complexity and maintaining accuracy. This is done in Python using the f_classif() function provided by the Scikit-learn library. The SelectkBest() function, based on the ANOVA test, identifies the features with the highest values as the most important. Although ANOVA (univariate) does not take into account the relationship between features, as some of the selected features may be related to each other (repetition of information), the results showed improvement. This means that the selected features were indeed distinctive and sufficient.

Grid search is a widely used hyperparameter tuning technique that improves model performance by selecting the best configuration based on evaluation metrics. This is achieved by systematically trying all possible parameter combinations to find the optimal set [20]. In this study, a grid search method is used to tune the hyperparameters of the LR classifier.

3.5 Classifiers

After completing all the previous steps, the following classifiers were applied: VGG16, LR, SVM, and KNN. These classifiers use the test dataset to test the knowledge they gained while learning from the training dataset. These classifiers are explained below:

- **VGG16 model**: the pre-trained VGG16 model is a type of convolutional neural network (CNN) with a 16-layer architecture. The input image to the network is 224 x 224 x 3, along with a fixed-size filter (3 x 3) and five max-pooling layers (2 x 2) applied across the entire network. In contrast, the two layers are fully connected to a softmax output layer at the top. It is a large network containing approximately 138 million parameters. Multiple convolutional layers are stacked to create deep neural networks that improve the ability to learn hidden features [21].
- **Logistic regression (LR):** this supervised learning method classifies data into a certain class by predicting the probability of a data point belonging to a certain class, with an output value ranging from 0 to 1. This method is commonly used in classification problems, especially in binary and multi-class classification. It is simplicity and efficiency [22].
- **Support vector machine (SVM)** is a supervised learning algorithm used for classification and regression issues. It works by finding the optimal hyperplane that separates data into distinct classes. SVM is especially effective in handling high-dimensional data, making it a suitable choice when working with complex feature vectors like those in medical image analysis [23].
- **K-nearest neighbors (KNN)** is a supervised ML algorithm used for regression and classification tasks. It classifies new data points based on similarity to existing data, typically using distance measures like euclidean distance. By majority voting among the nearest neighbors, the KNN determines the class. It is non-parametric, robust to noise, and commonly applied in statistical estimation and pattern recognition [24].

3.6 Evaluation metrics

This study used several metrics, including accuracy, AUC, precision, recall, F1 score, and confusion matrix, to evaluate the performance and efficiency of the classifiers, as listed below:

- **Confusion matrix:** it is used to assess the accuracy and performance of prediction classifiers in detail, displaying the accuracy and performance results for each class separately. The following are brief explanations of the confusion matrix elements:
  - True positive (TP): correctly predicted as a positive case of a positive.
  - True negative (TN): correctly predicted as a negative case of a negative.
  - False positive (FP): wrongly predicted as a negative case as a positive.
  - False negative (FN): wrongly predicted as a positive case as a negative.

Table 3 shows the confusion matrix for classifying four classes.

<table>
<tr><td colspan="2" rowspan="2">Confusion matrix</td><td colspan="4">Predicted class</td></tr>
<tr><td>Class 1</td><td>Class2</td><td>Class 3</td><td>Class 4</td></tr>
<tr><td rowspan="4">Actual class</td><td>Class1</td><td>TP</td><td>FN</td><td>FN</td><td>FN</td></tr>
<tr><td>Class2</td><td>FP</td><td>TN</td><td>TN</td><td>TN</td></tr>
<tr><td>Class3</td><td>FP</td><td>TN</td><td>TN</td><td>TN</td></tr>
<tr><td>Class4</td><td>FP</td><td>TN</td><td>TN</td><td>TN</td></tr>
</table>

Table 3: Confusion matrix for a classification with four classes.

Accuracy: it is the calculation of the ratio of the correct positive predictions to the total number of predictions. It is computed by the following equation:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad \textbf{(2)}$$

- **Precision metric:** the ratio of correct positive predictions to total resulting predictions. It is computed by the following equation:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

- **Sensitivity (recall):** the ratio of correct positive predictions to total positive predictions. It is computed by the following equation:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

- **F1-score:** it is a metric that combines precision and recall by calculating the harmonic mean between them. It is computed by the following equation:

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (5)$$

- **Area under the receiver operating characteristic curve (AUC-ROC):** expresses the classifier's efficiency in distinguishing between classes. The closer its value is to 1, the higher the performance in discrimination, and the further it is from 1, the weaker its performance is in distinguishing between classes.

4 Results

The result is divided into two parts. The first section presents the results obtained using the LR, SVM, and KNN algorithms, while the second section presents the results obtained using the VGG16 model.

4.1 Results using the LR, SVM, and KNN classifiers

In this section, the results of LR, SVM, and KNN classifiers are presented before and after applying performance optimization techniques such as using ANOVA F-test to select the best features and using Grid search to optimize the hyperparameters for the train set (60%) and test set (20%).

Table 4 shows the outcomes of SVM, KNN, and LR, before using improvement techniques, as the KNN, SVM, and LR, respectively, achieved an accuracy of 85 %, 87 %, and 89 %.

Table 5 shows the outcomes of SVM, KNN, LR, and LR after using optimization techniques represented by 0.92, respectively, compared to the KNN classifier, which achieved 0.90.NOVA F-test to select the best features for KNN and SVM, and grid search to adjust the hyperparameters of LR, where LR and SVM achieved accuracy results of 0.93 and 0.92, respectively, compared to the KNN classifier, which achieved 0.90.

Figure 4 (a, b, and c) shows the performance of KNN, SVM, and LR classifiers in class discrimination after using ANOVA F-test and grid search techniques based on a confusion matrix. The SVM and LR classifiers achieved better performance compared to the KNN classifier.

Figure 5 (a, b, and c) shows the ability of KNN, SVM, and LR classifiers to differentiate between classes after applying an ANOVA F-test and grid search techniques based on the ROC curve. The SVM and LR classifiers achieved superior discrimination ability, with an AUC of 0.99, compared to the KNN classifier, which achieved 0.98.

4.2 Results using the VGG16 classifiers

In this section, the outcomes for the VGG16 classifier, confusion matrices for the training and test sets, and ROC curves are presented, respectively.

Table 6 shows the results for the VGG16 classifier, which achieved an accuracy of 0.96.

Figure 6 shows the performance of the VGG16 classifier in distinguishing between classes based on the confusion matrix.

Figure (7-a) shows that the learning process of the VGG16 classifier runs smoothly and stably, demonstrating the success of the training and the effectiveness of the classifier in generalizing to new data.

Figure (7-b) shows that the Val-loss decreases at a higher rate than the loss, indicating that the model has good generalization ability and no overfitting, which means a successful training process.

Figure 8 shows the ability of the VGG16 classifier to distinguish between classes based on the ROC curve with an AUC of 1.00.

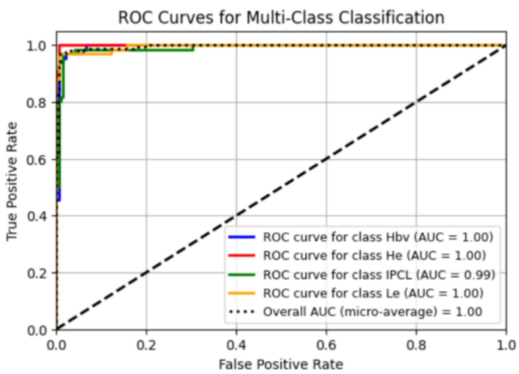


Figure 8: ROC curve of the VGG16 classifier.

Table 4: Outcomes of KNN, SVM, and LR before using ANOVA F-test and grid search.

	Train set					
	NO. of features	**Classifier**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Before using the ANOVA F-test and grid search	**512**	**KNN**	0.91	0.91	0.91	0.91
	512	**SVM**	0.91	0.91	0.91	0.91
	512	**LR**	0.88	0.88	0.88	0.88
	Test set					
	512	**KNN**	0.85	0.85	0.85	0.85
	512	**SVM**	0.87	0.87	0.87	0.87
	512	**LR**	0.86	0.86	0.86	0.86

Table 5: Outcomes of KNN, SVM, and LR after using ANOVA F-test and grid search.

Optimization techniques	**Train set**					
	NO. of features	**Classifier**	**Accuracy**	**Precision**	**Recall**	**F1-score**
ANOVA F to select features	**150**	**KNN**	0.92	0.92	0.92	0.92
	150	**SVM**	0.98	0.98	0.98	0.98
Grid Search	**512**	**LR**	0.98	0.98	0.98	0.98
	Test set					
ANOVA F to select features	**150**	**KNN**	0.90	0.90	0.90	0.90
	150	**SVM**	0.92	0.92	0.92	0.92
Grid Search	**512**	**LR**	0.93	0.93	0.93	0.93

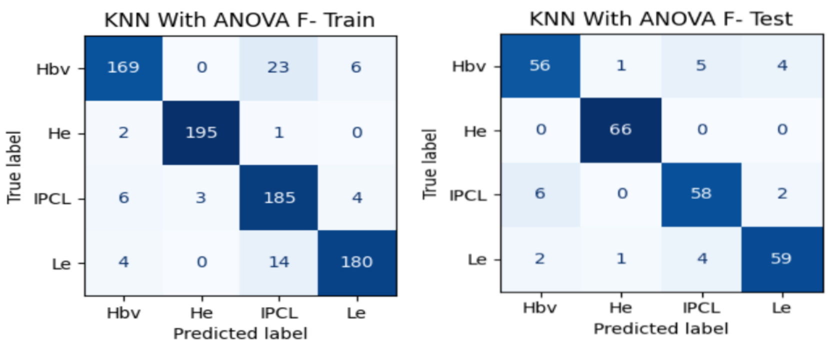


Figure 4: A: confusion matrix for the KNN classifier with anova f.

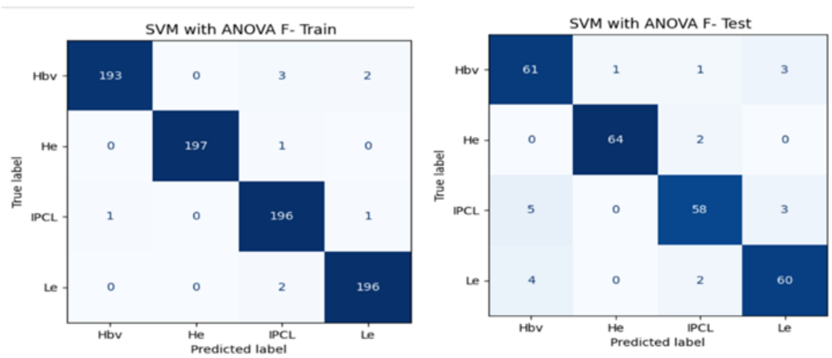


Figure 4: B: confusion matrix for the SVM classifier with anova f.

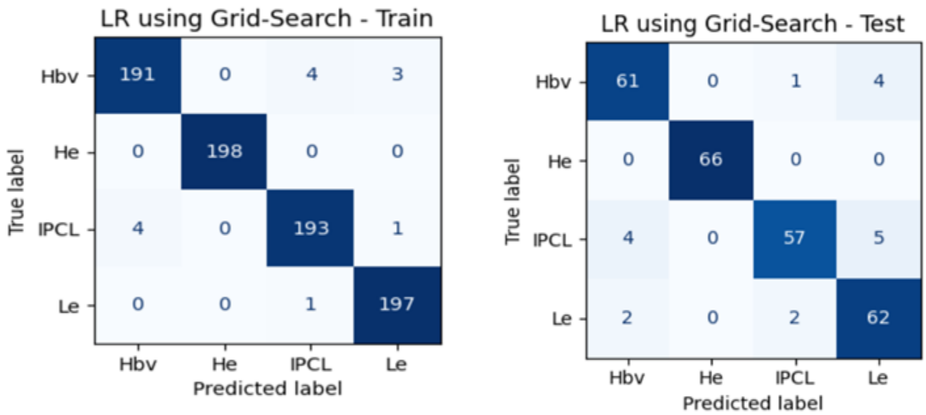
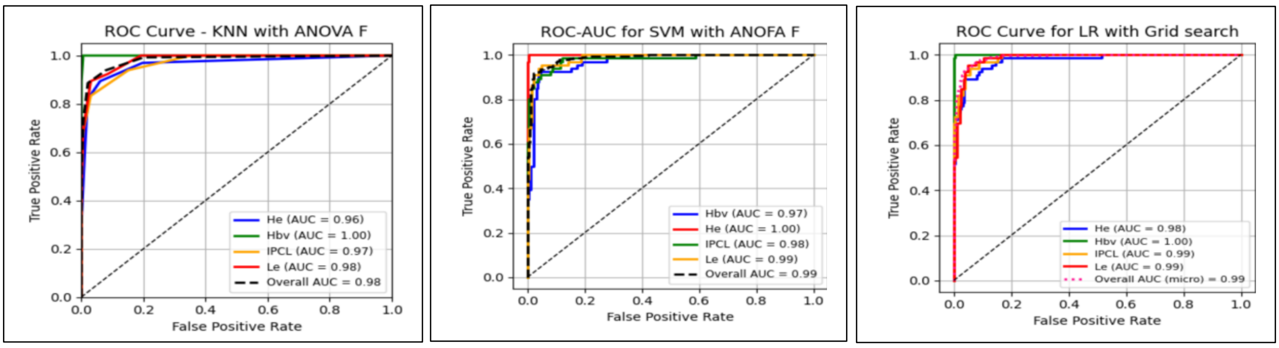


Figure 4: C: confusion matrix for the LR classifier with grid search.



(a) (b) (c)

Figure 5: (A, b, and c): ROC curves for the classifiers KNN, SVM, and LR with anova f and grid search.

Table 6: Outcome of the VGG16 classifier.

Classifier: VGG16		Precision	Recall	F1-score	Accuracy
NO. of features: 256	Train set (60%)	0.9797	0.9797	0.9797	0.9797
	Test set (20%)	0.9697	0.9697	0.9697	0.9696

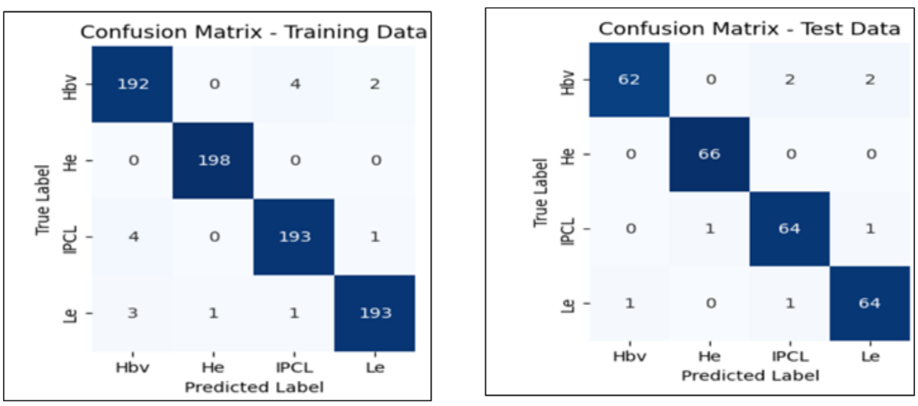


Figure 6: Confusion matrix for the VGG16 classifier.

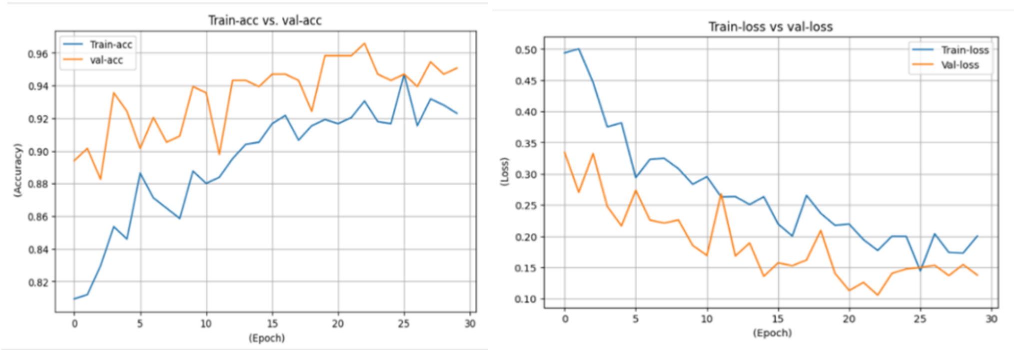


Figure (7-a) accuracy curve. Figure (7-b) loss curve.

5 Discussion

This study presents a DL-based system for predicting laryngeal cancer for early detection. ML algorithms such as LR, SVM, and KNN performed well. However, after applying an ANOVA F-test to select the best features and a Grid search to optimize the hyperparameters, the results showed a significant improvement in accuracy, especially on the test data, where the LR, SVM, and KNN algorithms achieved accuracies of 0.93, 0.92, and 0.90, respectively. In addition to the ML algorithms, the VGG16 DL model was used, which outperformed the ML algorithms with a test accuracy of 0.96.

The reason for the superiority of the VGG16 classifier using the preprocessing described above is that it is a deep learning model capable of extracting important features and learning from them automatically (combining the power of feature representation and classification at the same time), unlike the SVM, KNN, LR machine learning algorithms that rely on simple assumptions (such as linearity or distance). Therefore, F-ANOVA was used to avoid the large number of features and their uselessness, as they may cause a decrease in its performance. Thus, choosing the best features leads to a reduction in dimensions, which leads to a reduction in complexity (the fewer the features, the less the complexity) and thus reduces the occurrence of overfitting in learning and increases the speed and efficiency of training, especially with models like KNN and SVM. This is what helped us improve the results.

Despite some potential challenges in ANOVA F testing: it is univariate, where features are selected based on their individual distinguishing abilities and do not consider the correlation between features, potentially leading to the loss of important information; and it is reduced generalizability, as the model performs better on training and test data when the number of features is reduced but degrades performance on new data if important features are removed. In this study successfully removed unhelpful features; those that were most distinctive among the classes were retained, and the performance of the SVM and KNN classifiers was improved by increasing the accuracy of the KNN classifier from 0.85 to 0.90 and the SVM classifier from 0.87 to 0.92, indicating that the selected features were sufficient to effectively represent the data.

Studies [8], [9], and [10] relied on a dataset split of 80% training and 20% testing sessions. In this work, we adopted a more balanced splitting method (60% training, 20% validation, and 20% testing sessions), which contributed to a more accurate evaluation of the models' performance and helped reduce the possibility of bias in the evaluation. In addition, we used preprocessing methods such as image processing using YCBCR space, CLAHE, sharpen filter, and augmentation, so we relied on the VGG16 model as a feature extractor, which was not implemented in Studies [8], [9], and [10]. Although a clinical dataset was used in Study [11] and the SVM algorithm achieved an accuracy of 85%, it achieved even higher accuracy when used on an imaging dataset, with an accuracy of 0.92%. This contributed to the VGG16 model having better learning and generalization capabilities. The VGG16 model achieved the highest test accuracy of 0.96, compared to the results achieved by studies [8], [9], and [10], which were 95.07%, 93.5%, and 92.04%, respectively.

6 Conclusion

This study proposes an automated system for laryngeal cancer prediction and diagnosis using DL methods such as VGG16 and ML methods such as LR, SVM, and KNN. Preprocessing techniques such as Gaussian blurring, converting the RGB to YCBCR, and CLAHE were used, as well as feature extraction with VGG16, which improved the system's quality and made it suitable for analysis and predictive modeling. This research aims to assist medical professionals in diagnosing laryngeal cancer, preventing its progression, reducing diagnosis time and cost, improving treatment response, and directing treatment selection by predicting its stage. The proposed system goes through several stages. In the first stage, we divided the dataset into a 60% training set, a 20% validation set, and a 20% test set. The second stage involved preprocessing the dataset. In the third stage, features were extracted using the pre-trained VGG16 model. The fourth stage involved applying optimization techniques such as ANOVA F-test to select the best features for the SVM and KNN algorithms and using grid search to tune the hyperparameters of the LR algorithm. In addition to the evaluation phase, the system performance on test data was evaluated using precision, recall, confusion matrix, and ROC curve metrics.

One of the limitations of this study is the limited dataset, in addition to the lack of organization of images according to the identity of each patient, as the images are merged without separating the images specific to each patient. This prevents the application of performance verification techniques such as K-fold cross-validation or Leave One Patient Out (LOPO).

In this study, we focused on using only the laryngeal tissue imaging data set to predict laryngeal cancer, determine disease progression, and select immediate treatment. However, future work could expand the scope of research by developing a system to predict tumor size and survival using both imaging and clinical data sets.

References

[1] A. Sahu *et al.*, "Imaging Recommendations for Diagnosis, Staging and Management of Larynx and Hypopharynx Cancer," Feb. 01, 2023, *Georg Thieme Verlag*. doi: 10.1055/s-0042-1759504.

[2] M. Q. Ali, A. Raji, and M. S. Abdulameer, "Patterns of laryngeal cancer presentation of Iraqi patients," *Revista Latinoamericana de Hipertension*, vol. 16, no. 4, pp. 298–303, 2021, doi: 10.5281/zenodo.5812217.

[3] W. Wang *et al.*, "Creation of a machine learning-based prognostic prediction model for various subtypes of laryngeal cancer," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56687-x.

[4] A. R. Hut *et al.*, "Laryngeal Cancer in the Modern Era: Evolving Trends in Diagnosis, Treatment, and Survival Outcomes," May 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jcm14103367.

[5] H. Alaskar and T. Saba, "Machine Learning and Deep Learning: A Comparative Review," 2021, pp. 143–150. doi: 10.1007/978-981-33-6307-6_15.

[6] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Aug. 2023, doi: 10.1186/s43067-023-00108-y.

[7] A. E. Hassanien, A. T. Azar, T. Gaber, R. Bhatnagar, and M. F. Tolba, Eds., *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)*, vol. 921. in Advances in Intelligent Systems and Computing, vol. 921. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-14118-9.

[8] S. A. Alzakari, M. Maashi, S. Alahmari, M. A. Arasi, A. A. K. Alharbi, and A. Sayed, "Towards laryngeal cancer diagnosis using Dandelion Optimizer Algorithm with ensemble learning on biomedical throat region images," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-70525-0.

[9] N. Mohamed *et al.*, "Automated Laryngeal Cancer Detection and Classification Using Dwarf Mongoose Optimization Algorithm with Deep Learning," *Cancers (Basel)*, vol. 16, no. 1, Jan. 2024, doi: 10.3390/cancers16010181.

[10] F. Alrowais, K. Mahmood, S. S. Alotaibi, M. A. Hamza, R. Marzouk, and A. Mohamed, "Laryngeal Cancer Detection and Classification Using Aquila Optimization Algorithm With Deep Learning on Throat Region Images," *IEEE Access*, vol. 11, pp. 115306–115315, 2023, doi: 10.1109/ACCESS.2023.3324880.

[11] Z. Li, T. Li, P. Zhang, and X. Wang, "A practical online prediction platform to predict the survival status of laryngeal squamous cell carcinoma after 5 years," *American Journal of Otolaryngology - Head and Neck Medicine and Surgery*, vol. 45, no. 3, May 2024, doi: 10.1016/j.amjoto.2023.104209.

[12] "Laryngeal dataset." Accessed: Apr. 23, 2025. [Online]. Available: https://www.kaggle.com/datasets/mahdiehhajian/laryngeal-dataset?resource=download

[13] T. G. Devi, N. Patil, S. Rai, and C. S. Philipose, "Gaussian Blurring Technique for Detecting and Classifying Acute Lymphoblastic Leukemia Cancer Cells from Microscopic Biopsy Images," *Life*, vol. 13, no. 2, Feb. 2023, doi: 10.3390/life13020348.

[14] Q. V. Kieu, V. N. Huynh, T. P. Nghiem, O. C. Do, and G. S. Tran, "A NEW METHOD FOR MEDICAL IMAGE FUSION BASED ON GAUSSIAN BLUR FILTER AND ROBINSON COMPASS OPERATOR," *Journal of Computer Science and Cybernetics*, vol. 40, no. 2, pp. 135–146, May 2024, doi: 10.15625/1813-9663/18655.

[15] K. Nguyễn *et al.*, "Pre-processing Image using Brightening, CLAHE and RETINEX PRE-PROCESSING IMAGES USING BRIGHTENING, CLAHE AND RETINEX", doi: 10.48550/arXiv.2003.10822.

[16] W. Yan, G. Xu, Y. Du, and X. Chen, "SSVEP-EEG Feature Enhancement Method Using an Image Sharpening Filter," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 115–123, 2022, doi: 10.1109/TNSRE.2022.3142736.

[17] "(PDF) DATA AUGMENTATION METHODS IN ARTIFICIAL INTELLIGENCE." Accessed: Aug. 06, 2025. [Online]. Available: https://www.researchgate.net/publication/390549894_DATA_AUGMENTATION_METHODS_IN_ARTIFICIAL_INTELLIGENCE

[18] M. Vasantha, D. V Subbiah Bharathi, and R. Dhamodharan, “Medical Image Feature Extraction, Selection, and Classification,” 2010.

[19] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, “Analyzing the impact of feature selection on the accuracy of heart disease prediction,” Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.03239

[20] D. M. Belete and M. D. Huchaiah, “Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results,” *International Journal of Computers and Applications*, vol. 44, no. 9, pp. 875–886, 2022, doi: 10.1080/1206212X.2021.1974663.

[21] K. Kamal and H. EZ-ZAHRAOUY, “A comparison between the VGG16, VGG19 and ResNet50 architecture frameworks for classification of normal and CLAHE processed medical images,” Apr. 28, 2023. doi: 10.21203/rs.3.rs-2863523/v1.

[22] D. Sabeeh and Maalim A. Aljabery, “A Study on Cirrhosis Prediction Based on Machine Learning Techniques,” *Journal of Education for Pure Science*, vol. 14, no. 4, Dec. 2024, doi: 10.32792/jeps.v14i4.461.

[23] D. S. Ali and M. A. Aljabery, “Predicting Liver Cirrhosis Stages Using Extra Trees, Random Forest, and SVM with Data Mining Techniques,” *Informatica (Slovenia)*, vol. 48, no. 21, pp. 15–26, 2024, doi: 10.31449/inf.v48i21.6752.

[24] V. T. Hai, P. H. Chuong, and P. T. Bao, “New Approach of KNN Algorithm in Quantum Computing Based on New Design of Quantum Circuits,” *Informatica (Slovenia)*, vol. 46, no. 5, pp. 95–103, Mar. 2022, doi: 10.31449/inf.v46i5.3608.