**Binary Classification of Benign and Malignant Laryngeal Lesions Using Transfer Learning, and Ensemble Methods**

Md.Shaharear Kabir Rabby, Md Alfaz Zaman Akash, Zahidul Islam Turja, Md. Rabby Rayhan

\*Ahmed Wasif Reza1

Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh

**Emails:** shaharearrabby711@gmail.com, info.alfazakash@gmail.com, zahidulturja@gmail.com, rabbyrayhan56@gmail.com

**Abstract:** This paper presents a reliable classification approach for Laryngeal lesions using pre-trained models and sophisticated combining methods. The dataset includes 11,144 images from 210 patients, with 69% benign and 31% malignant lesions. The data preprocessing included standardization, augmentation, and normalization. When implemented in Google Colab, the stacked model was formed by combining ResNet50Vs2, MobileVs3, and EfficientNetV2, with a test data accuracy of 93.76%. Future work should be focused on diversifying the dataset and collaborating with medical professionals to refine the classification approach continuously.

**Keywords:** Larynx Cancer Classification, Transfer Learning, Ensemble Learning, Meta Modeling.

**1. Introduction**

Ensuring the well-being of our mouths and teeth is vital for overall health. The World Health Organization emphasizes that oral health extends beyond just our smiles, playing a substantial role in our overall health and life quality. The mouth, home to various bacteria, acts as a passage to both our respiratory and digestive systems, offering insights into potential underlying diseases. For example, diseases like gum disease (periodontitis) can lead to infections affecting vital organs such as the heart and kidneys. Essentially, giving importance to oral health constitutes a worthwhile investment in our overall well-being and quality of life [1].

In recent times, endoscopic imaging modalities have become standard procedures for efficiently screening individuals to potentially diagnose laryngeal cancer at its early stages in various clinical settings. These screenings are beneficial before clinical surgery for the histological examination of tissues during optical surgery [2]. Some strategies, like combining contact endoscopy (CE) with narrow-band imaging (NBI), can enhance and enlarge the visualization of morphological changes, including real-time 3D alignment of vocal folds and sub-epithelial blood vessels.

The advanced progress in feature engineering, along with the evolution of Machine Learning (ML) and Deep Learning (DL) techniques within the medical field, offers various avenues to support clinicians and address challenges encountered in clinical settings. Numerous computer-based methods have been applied to images for this purpose. These approaches aim to aid otolaryngologists by furnishing additional information on cancer stage and vascular tree characteristics [3]. In the domain of Laryngeal image analysis, a combination of Convolutional Neural Networks (CNN) incorporating texture and frequency-domain-based features, alongside a set of manually crafted texture and first-order statistical features, has been suggested for the classification of benign and malignant cancer.

**2. Literature Review**

In our research, we have reviewed various papers related to our field and examined the methods and conclusions of other researchers. Our goal is to understand the challenges, techniques, and concerns in the field of research methods.

Laryngeal carcinoma is a common and life-threatening disease. It can be classified as benign, malignant, primary, or secondary. Manual classification is inefficient, and early detection is critical. Neural networks can be used to identify and classify these tumors.

In [4], authors propose a neural network approach that includes pre-processing and segmentation, as well as feature extraction and classification. Hybrid classifiers are used to improve the accuracy of neural network classification.

In [5], the authors propose a deep learning technique for the simultaneous detection of benign vocal cord tumors in laparoscopic images and classification into cysts, granulomas, leukoplakia, nodules, and polyps. The technique is intended for simplified home-based self-prescreening to detect early-stage tumors around the vocal cord. Four CNN models were implemented, trained, validated, and tested using 2183 laryngoscopic images. The experimental results showed that Yolo V4 had the highest F1 score for all tumor types and the lowest false-negative rate for each type. The embedded-operated Yolo V4 model demonstrated an equivalent F1 score to the computer-operated Yolo-4 model. The authors conclude that the proposed deep-learning-based home screening techniques have the potential to improve the long-term survival of patients with vocal cord tumors by aiding in the early detection of tumors around the vocal cord.

The authors of [3] proposed a deep convolutional neural network (DCNN)-based fully automatic CE-NBI endoscopic image method for the classification of laryngeal lesions. To obtain the best architecture for the classification task, the study used a pre-trained ResNet50 model with fine-tuning and a cut-off-layer technique. A total of 72 models were trained and tested using the gathered data in three experiments aimed at identifying the most appropriate model.

According to the findings, Model 5 was chosen for additional testing because it demonstrated excellent accuracy, sensitivity, and specificity in Experiment 1. Model 5 showed a high degree of performance in distinguishing benign from malignant lesions in CE-NBI images. The research findings indicate that the suggested methodology can provide otolaryngologists with an impartial evaluation.

To distinguish between benign and malignant laryngeal lesions, the authors [6] of this study compared multi-observer manual classification with automatic classification of CE+NBI images. They made use of 1632 high-resolution CE+NBI images with distinct vascular patterns, as well as a dataset comprising 68 patients. While the automatic classification employed an algorithm based on vascular patterns, the manual classification involved more and less experienced otolaryngologists visually evaluating the images. The automatic classification showed promise in mitigating the problem of misclassifying laryngeal lesions, as evidenced by its high sensitivity and specificity.

The authors [7] of this paper studied the diagnostic performance of narrowband imaging (NBI) in detecting laryngeal cancer using the Ni classification. Using a statistical analysis software package, the authors discovered that the Ni classification, when applied to NBI, has a sensitivity of 95.0% and a specificity of 83.3% when it comes to the diagnosis of laryngeal cancer. Higher Ni grades were strongly correlated with more advanced disease, according to the final results of the method, which involved comparing histopathological grades to their respective Ni classification**.**

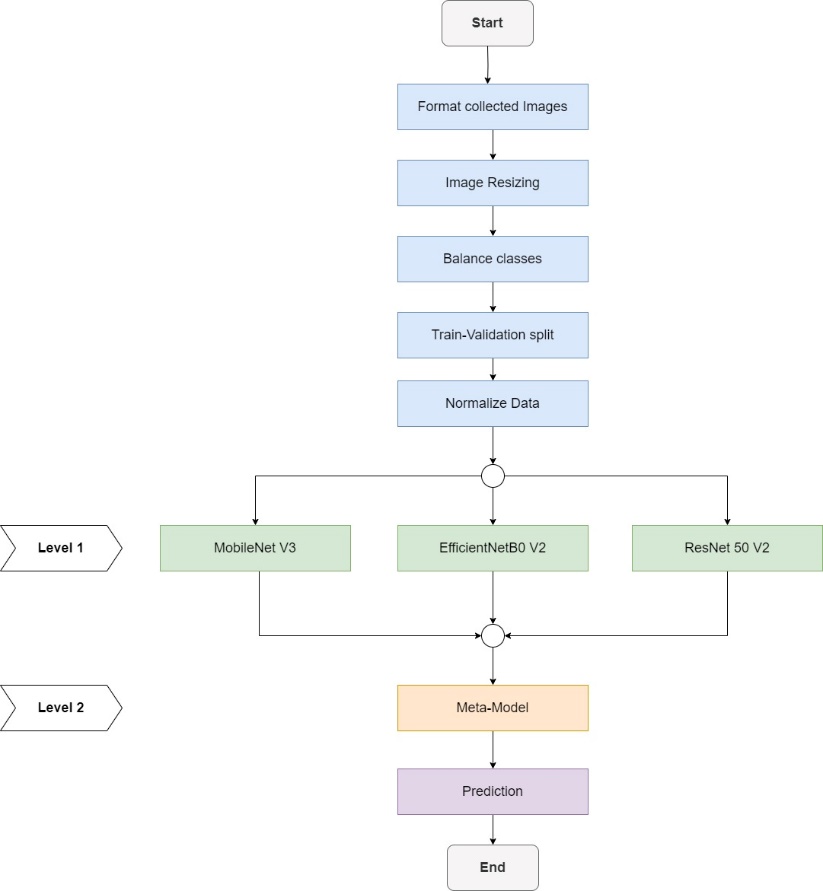
In [8], the authors introduced a convolutional neural network (CNN) along with an effective image segmentation technique, achieving an impressive overall accuracy of 98.12%.

The authors of [9] proposed a hybrid model integrating convolutional neural networks (CNNs) and long short-term memory recurrent neural networks (LSTM RNN) for the classification of four benign and four malignant breast cancer subtypes. Their model, evaluated on the BreakHis dataset containing 2480 benign and 5429 malignant cancer images at varying magnifications, achieved a notable accuracy of 92.5% for multi-class classification.

In [10], authors explored texture feature extraction for larynx images, representing images as hilly surfaces to identify different paths between points. The introduced CyEfF features demonstrated a maximum classification accuracy of 88.2%, improving the GF classification accuracy by 3 to 12 percent. These studies collectively highlight the effectiveness of advanced neural network architectures and innovative feature extraction techniques in achieving high accuracy across various medical imaging applications.

**3. System Architecture and Design**

The proposed Larynx Cancer Classifying method’s design procedure and architecture is described in detail in this section.

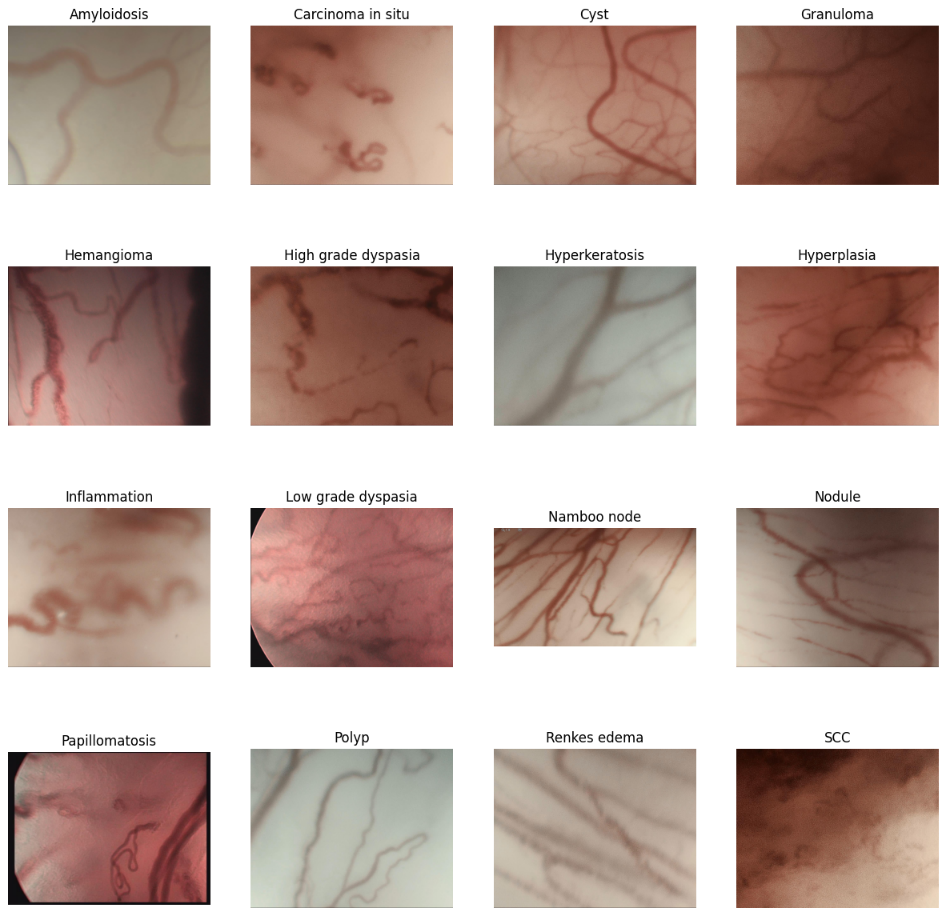


**Figure 1**: Workflow of the proposed system

The workflow of the proposed framework is represented in Figure 1. This framework can be divided into three main parts, data-preprocessing, training pre-trained models, and building a meta-model using the base model.

**3.1 Dataset Description**

The dataset employed in this research originates from Contact Endoscopy – Narrow Band Imaging (CE-NBI) [4] data set for laryngeal lesion assessment.



**Figure 2:** Images of 16 sub-categories in alphabetical order

Comprising a total of 11,144 images, the dataset spans two primary categories: Benign and Malignant. These images were collected from a diverse group of 210 different patients over the period from January 1, 2015, to December 31, 2021, in Germany.

The Benign class encompasses 7,657 images, representing lesions from 150 distinct patients. This category is further subdivided into 13 sub-categories, providing a subtle representation of various benign laryngeal lesions. Notably, the Benign class constitutes 69% of the total dataset, underscoring the prevalence and diversity of benign lesions captured in the dataset.

In contrast, the Malignant class comprises 3,487 images originating from 60 different patients. This class is divided into 3 sub-categories, capturing distinctive features of malignant laryngeal lesions. The Malignant class accounts for the remaining 31% of the dataset, reflecting the focus on malignant lesions within the collected CE-NBI data.

The dataset exhibits a rich diversity in patient demographics, with lesions collected from a varied pool of 210 patients. The inclusion of both benign and malignant cases from multiple patients enhances the dataset's representativeness and ensures a comprehensive evaluation of the proposed classification method.

This dataset, sourced from a six-year period, provides a temporal dimension to the analysis, capturing potential variations and trends in laryngeal lesions over the specified timeframe.

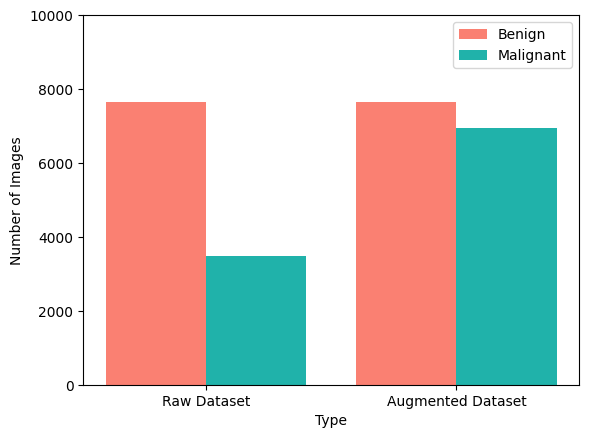
**3.2 Data Preprocessing**

In this subsection, we detail the data processing steps undertaken to prepare the dataset for model training. Initially, the dataset, consisting of 11,144 images, was categorized into two primary classes: Benign and Malignant. These classes further comprised 13 and 3 subclasses, respectively. To facilitate a binary classification task, the problem was redefined, combining the subclasses within each primary class.

To address variations in image sizes, a crucial step was taken to standardize all images to a consistent size. While the original dataset exhibited diverse image dimensions, ranging from (868, 540, 3) to (1280, 1280, 3), we uniformly resized all images to (224, 224, 3). This standardization ensures uniformity across the dataset and facilitates the training of machine learning models, which typically require fixed input dimensions.

Recognizing the inherent class imbalance, with the Benign class containing 2.2 times more data than the Malignant class, a data augmentation technique was applied specifically to the Malignant class. This step aimed to handle biases and create a more balanced dataset. Post-augmentation, the Malignant class comprised 6,948 images, bringing it in close alignment with the size of the Benign class.

To assess model performance accurately, the dataset was partitioned into training, validation, and test sets in a ratio of 7:1:2, respectively. This distribution ensures a representative mix of data for model training, validation, and evaluation, contributing to robust model generalization.



**Figure 3:** Comparison between two classes before and after augmentation

Prior to model training, the dataset underwent normalization to standardize pixel values. Normalization is a critical preprocessing step that enhances model convergence by bringing data within a consistent numerical range. In this case, pixel values were scaled to a range suitable for effective training.

These meticulous data processing steps collectively contribute to the creation of a well-prepared dataset, setting the stage for subsequent model training and evaluation.

**4. Implementation and Evaluation**

The implementation process, experimental setup and evaluation of this research is described in detail in this section.

**4.1 Hardware and Software Setup**

The proposed system was implemented in the Google Colab notebook environment, leveraging the advantages of a cloud-based, collaborative platform. The free version of Google Colab was employed, offering a convenient and accessible space for model development and training. To expedite the training process and enhance computational efficiency, we were allocated the Tesla T4 GPU, a high-performance graphics processing unit known for its capabilities in accelerating deep learning tasks.

**4.2 Implementation**

In addressing the binary classification task, our study involved the training of three pre-trained models: MobileNetV3, EfficientNetB0V2, and ResNet50V2. To initiate the training, we loaded each pre-trained model and exposed them to both raw and augmented datasets. The inclusion of both datasets in the training pipeline aimed to facilitate a comprehensive comparison of model performance under varied data conditions, ensuring robustness and adaptability.

To mitigate overfitting during the training process, we implemented early stopping techniques. Specifically, we monitored the validation loss with a patience of 3 epochs, halting the training if there was no improvement in the validation loss for three consecutive epochs. This strategic approach prevented the models from learning noise in the training data and enhanced their ability to generalize to unseen data.

Additionally, to fine-tune the learning process, we dynamically adjusted the learning rate during training. A reduction of 20% in the learning rate was triggered whenever the validation loss showed signs of reaching a standstill. This adaptive learning rate strategy aimed to navigate the model through the optimization landscape more effectively, allowing for finer convergence and improved overall performance.

After the individual training of MobileNetV3, EfficientNetB0V2, and ResNet50V2, we delved into advanced model combining techniques for further enhancement. Specifically, we explored ensemble learning and meta-modeling. Ensemble learning involved the aggregation of predictions from the three individual models, harnessing their collective intelligence. Simultaneously, meta-modeling leveraged the insights from the individual models to create a higher-level model, capturing subtle patterns and improving overall accuracy. These post-training techniques were instrumental in extracting the full potential of our pre-trained models, yielding superior results compared to individual models.

**4.3 Experimental Result**

Table 1 provides a comprehensive overview of the performance metrics for all models created in our study, evaluated on both the raw dataset and the augmented dataset. Notably, ResNet50V2 emerged as the top-performing pre-trained model, exhibiting exceptional predictive capabilities with an impressive accuracy of 95.16% on the training set. Furthermore, ResNet50V2 demonstrated robust generalization to unseen data, achieving accuracy rates of 90.8% and 91.17% on the validation and testing datasets, respectively.

However, the performance of ResNet50V2 was surpassed by the ensemble of these three models. The combination of ResNet50V2 with MobileNetV3 and EfficientNetB0V2, resulted in an ensemble that achieved even higher accuracy rates, showcasing the effectiveness of model aggregation for improved performance.

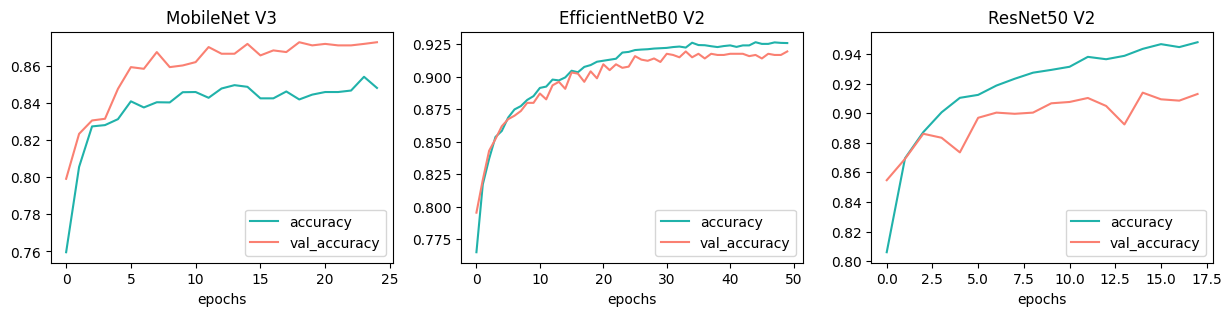
**Table 1:** Comprehensive overview of the performance metrices for each model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Dataset** | **Training  Loss** | **Training  Accuracy (%)** | **Validation Loss** | **Validation Accuracy (%)** | **Test Accuracy (%)** |
| MobileNet V3 | Original | 0.3306 | 84.8 | 0.2971 | 87.26 | 88.03 |
| MobileNet V3 | Augmented | 0.176 | 93.60 | 0.2108 | 92.45 | 91.36 |
| EfficientNetB0 V2 | Original | 0.2022 | 92.57 | 0.2166 | 91.92 | 91.08 |
| EfficientNetB0 V2 | Augmented | 0.2077 | 91.74 | 0.2286 | 91.5 | 89.78 |
| ResNet 50 V2 | Original | 0.1535 | 94.8 | 0.2075 | 91.3 | 91.12 |
| ResNet 50 V2 | Augmented | 0.1462 | 95.16 | 0.2159 | 90.85 | 91.17 |
| Ensemble | Original |  |  |  |  | 92.87 |
| Ensemble | Augmented |  |  |  |  | 92.18 |
| Meta-Model | Original | 0.1126 | 95.95 | 0.1655 | 93.19 | 93.5 |
| Meta-Model | Augmented | 0.1027 | 96.35 | 0.1661 | 93.55 | 93.76 |

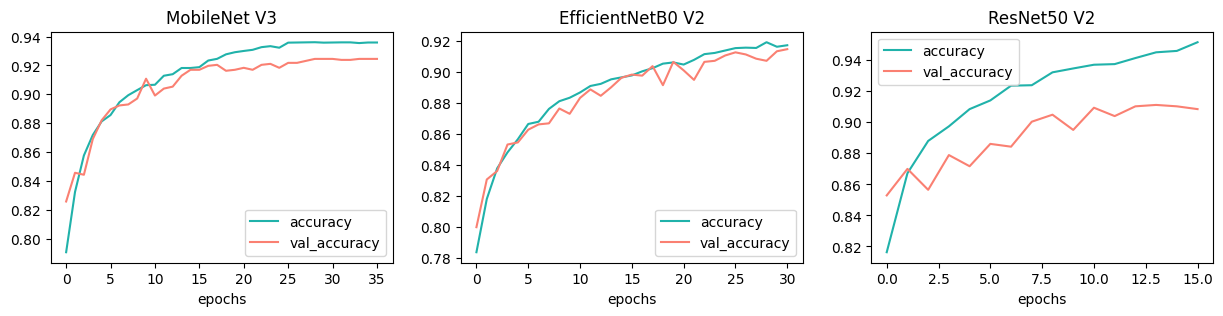
Remarkably, the stacked model, an advanced meta-modeling technique, outperformed both the individual pre-trained models and their ensemble. The stacked model attained an impressive accuracy of 93.76% on test set during training on the augmented dataset. This underscores the significance of leveraging meta-modeling to harness the complementary strengths of individual models, ultimately enhancing the classification performance.

**4.4 Performance Evaluation**

Following our earlier discussions, it is evident that the Stacked model, combining MobileNetV3, EfficientNetB0V2, and ResNet50V2, demonstrated superior performance on unseen data.



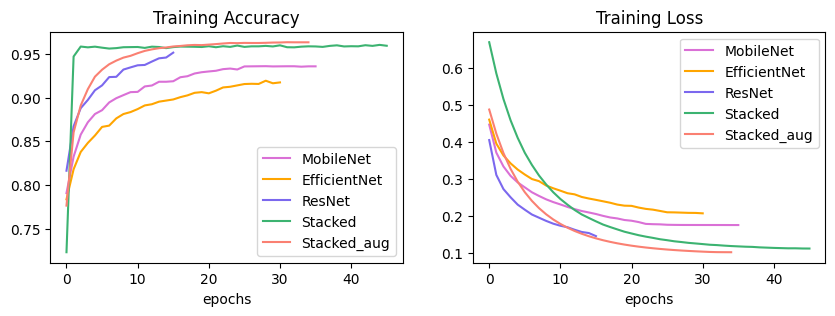
**Figure 4:** Training accuracy curve of the base models on original dataset



**Figure 5:** Training accuracy of the base models on augmented dataset

Specifically, when trained on the balanced augmented dataset, the Stacked model achieved an impressive accuracy of 93.76% on the test data, showcasing its efficacy in laryngeal lesion classification. Upon closer examination of the accuracy graphs depicting the training of individual base models, notable observations emerge. EfficientNetB0V2 exhibited robust performance on both the balanced and biased datasets, showcasing its versatility in handling different data distributions. MobileNetV3, while initially demonstrating promising results on the original dataset, exhibited substantial improvement when trained on the augmented dataset. Conversely, ResNet50V2 faced challenges during validation, indicating a potential risk of overfitting. Despite these concerns, the subsequent analysis in Table 1 underscores the model's ability, as ResNet50V2 outperformed others on the test dataset, justifying its inclusion in the stacked model.

While the accuracy and loss curves provide valuable insights into the training dynamics of individual models, a crucial decision-making factor was the performance on the test dataset. Despite validation concerns, ResNet50V2 demonstrated superior predictive capabilities on unseen data, making it a valuable addition to the ensemble of models within the Stacked architecture. This strategic decision leveraged the strengths of each model to achieve optimal performance on the test set.



**Figure 6:** Accuracy and loss curves of 5 models

Figure 6 illustrates the accuracy and loss curves for all models trained. The visual representation affirms the superior performance of the Stacked model compared to individual base models. This comprehensive comparison reinforces the efficacy of model stacking as a technique for enhancing classification accuracy in laryngeal lesion assessment.

**5. Conclusion**

In summary, our research introduces a robust classification method for laryngeal lesion assessment, employing pre-trained models and advanced combining techniques. Leveraging a carefully prepared dataset and strategic data preprocessing, we trained individual models, with ResNet50V2 exhibiting superior accuracy on the test set. The stacked model, combining MobileNetV3, EfficientNetB0V2, and ResNet50V2, emerged as the top performer, achieving a remarkable 93.76% accuracy on the test data. The dataset, though extensive, may not encompass all laryngeal conditions, potentially impacting model generalization. Additionally, reliance on pre-trained models assumes universal feature representation. Future improvements should focus on dataset diversification, exploring advanced pre-trained models, and integrating interpretability techniques. Collaborations with medical professionals for richer data and continuous refinement of the classification approach will be crucial for advancing the effectiveness and applicability of our method.

**References:**

1. R. R. Tobias, L. C. De Jesus, M. E. Mital, S. C. Lauguico, A. A. Bandala, R. R. P. Vicerra, and E. P. Dadios, "Throat Detection and Health Classification Using Neural Network," 2019 International Conference on Contemporary Computing and Informatics (IC3I), pp. 38-43, 2019, doi 10.1109/IC3I48456.2019.9055535.
2. P. K. Sahoo, S. Mishra, R. Panigrahi, A. K. Bhoi, and P. Barsocchi, "An Improvised Deep-Learning-Based Mask R-CNN Model for Laryngeal Cancer Detection Using CT Images," *Sensors*, vol. 22, no. 22, p. 8834, Nov. 2022. [Online]. Available: <https://www.mdpi.com/1424-8220/22/22/8834>.
3. M. A. Alshehri, A. A. Almazyad, A. A. Alshammari, and A. A. Almazyad, "Deep Convolution Neural Network for Laryngeal Cancer Classification on Contact Endoscopy-Narrow Band Imaging," *J. Med. Syst.*, vol. 45, no. 1, p. 7, Dec. 2021.
4. N. Esmaeili, N. Davaris, A. Boese, et al., "Contact Endoscopy – Narrow Band Imaging (CE-NBI) data set for laryngeal lesion assessment," *Sci Data*, vol. 10, p. 733, 2023. DOI: 10.1038/s41597-023-02629-7.
5. G.H. Kim, Y.J. Hwang, H. Lee, et al., "Convolutional neural network-based vocal cord tumor classification technique for home-based self-prescreening purpose," *BioMed Eng OnLine*, vol. 22, no. 81, 2023. DOI: 10.1186/s12938-023-01139-2.
6. Esmaeili, Nazila, Illanes, Alfredo, Boese, Axel, Davaris, Nikolaos, Arens, Christoph, Navab, Nassir and Friebe, Michael. "Manual versus Automatic Classification of Laryngeal Lesions based on Vascular Patterns in CE+NBI Images" *Current Directions in Biomedical Engineering*, vol. 6, no. 3, 2020, pp. 70-73. <https://doi.org/10.1515/cdbme-2020-3018>
7. Ahmadzada, S., Tseros, E., Sritharan, N., et al. (2020). “The value of narrowband imaging using the Ni classification in the diagnosis of laryngeal cancer,” *Laryngoscope Investigative Otolaryngology*, 5(6), 665–671.
8. R. Bhattacharjee, K. Suganya Devi and S. Vijaykanth, "Detecting Laryngeal Cancer Lesions From Endoscopy Images Using Deep Ensemble Model," *2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, Karaikal, India, 2023, pp. 1-6, doi: 10.1109/IConSCEPT57958.2023.10170113.
9. Srikantamurthy, M.M., Rallabandi, V.P.S., Dudekula, D.B. et al. “Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning,” *BMC Med Imaging* 23, 19 (2023). <https://doi.org/10.1186/s12880-023-00964-0>
10. Esmaeili, N.; Boese, A.; Davaris, N.; Arens, C.; Navab, N.; Friebe, M.; Illanes, “A. Cyclist Effort Features: A Novel Technique for Image Texture Characterization Applied to Larynx Cancer Classification in Contact Endoscopy—Narrow Band Imaging,” Diagnostics **2021**, 11, 432. https://doi.org/10.3390/diagnostics11030432