IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

Tongue Image Segmentation and Color Classification Based on Deep Learning

Mr. P. R. Krishna Prasad[1], Meda Venkata Lakshmi Divya[2], Nimmagadda Srinivas[3], Nuthi Prudhvi Raj Sai[4]
Associate Professor, Department of CSE[1]
UG Students, Department of CSE[2,3,4]
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India
krishnaprasad.palli@gmail.com, mvlakshmidivya2005@gmail.com,
prudhvisai1867@gmail.com, nimmagaddasrinivas07@gmail.com

Abstract: *The color of the human tongue serves as a vital indicator in traditional and modern diagnostics, reflecting the internal physiological state of the body. This research introduces an automated deep learning framework for the classification of tongue images into three clinical categories: White, Light Yellow, and Yellow. While an initial experimental pipeline utilizing U-Net for segmentation followed by a basic CNN classifier failed to meet performance benchmarks, this study proposes a transition to the EfficientNetB0 architecture. The methodology involves processing an original dataset of 1,613 images, which were refined and augmented to a balanced set of 1,500 images to enhance model generalization. By leveraging the compound scaling capabilities of EfficientNetB0, the system achieves an overall classification accuracy of 91%. The results demonstrate that this approach effectively captures subtle colorimetric variations, offering a non-invasive, objective, and efficient tool for health monitoring and disease screening through tongue surface analysis.*

Keywords: EfficientNetB0, Tongue Color Classification, Deep Learning, Medical Diagnostics, Image Augmentation

I. INTRODUCTION

The human tongue is a unique diagnostic organ that provides a non-invasive view of the body's internal physiological environment. In many clinical traditions, the surface characteristics of the tongue, particularly its color and coating, serve as preliminary indicators of a patient's systemic health. Variations in color, ranging from white to light yellow and yellow, are often associated with underlying health imbalances, though they do not pinpoint specific disorders without further clinical correlation. Despite its utility, traditional tongue diagnosis relies heavily on manual visual inspection, which is inherently subjective and varies based on the practitioner's experience and environmental lighting conditions.

The primary motivation for this research is to provide medical professionals with a reliable, objective tool to obtain an initial perspective on a patient's health status through automated tongue image analysis. Such a system can assist in standardized health screening, especially in preliminary diagnostic stages. Recent advancements in deep learning have significantly enhanced medical image analysis, offering the potential to automate these complex visual assessments.

In the initial phase of our research, we implemented a U-Net architecture for image segmentation to isolate the tongue from the background. This model achieved a high degree of precision, reaching 97% accuracy in both training and testing phases. However, the subsequent classification of these segmented regions presented a significant challenge. Our initial attempt to utilize a simple CNN classifier failed to provide the accuracy required for colorimetric differentiation. To address this, we transitioned to the EfficientNetB0 architecture, which utilizes compound scaling to

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

capture subtle color features more effectively. This paper details the implementation of this refined framework and demonstrates its success in classifying tongue images into three distinct categories with an overall accuracy of91%.

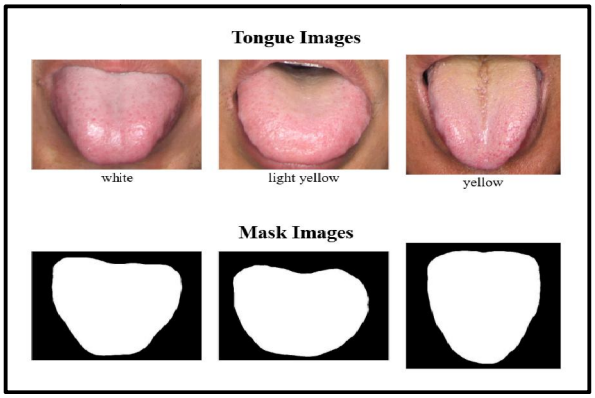

Fig. 1. Tongue images for the three classes of colors (white, light yellow, and yellow) with their corresponding binary mask images.

II. LITERATURE SURVEY

The field of automated tongue diagnosis has witnessed a significant shift from manual feature engineering to sophisticated deep learning architectures. Traditional methods primarily relied on hand-crafted features such as color moments to quantify tongue characteristics [7]. While these methods provided an early foundation, they often lacked the robustness needed to handle variations in lighting and image quality common in clinical settings [1].

Recent research has increasingly adopted two-stage pipelines to improve diagnostic precision. The first stage typically involves isolating the tongue body from the background—a task where U-Net and its variants have become a standard [2], [13]. For instance, recent studies have integrated attention mechanisms with U-Net to achieve high segmentation accuracy, significantly reducing interference from irrelevant oral regions [4], [5]. Despite these successes in segmentation, the secondary task of classification remains challenging due to the semantic similarity between different coating colors [1], [8].

Various Convolutional Neural Network (CNN) architectures have been deployed for tongue colour classification with varying degrees of success. While foundational models like ResNet have shown promise in identifying textures, they often possess a large number of parameters that may not be ideal for resource-constrained applications [1], [7]. Research has highlighted that standard CNNs can struggle to distinguish between subtle colour gradients, such as light yellow and yellow, often requiring optimized feature extraction layers [1], [10].

To address the limitations of computational overhead and classification accuracy, researchers have turned toward lightweight, optimally scaled architectures [9], [11]. Models like EfficientNet have emerged as powerful alternatives, utilizing compound scaling to maintain high performance with fewer parameters [6]. For example, EfficientNet-based encoders have been shown to improve both segmentation and subsequent classification tasks in medical imaging [8], [12]. This study builds upon these advancements by implementing EfficientNetB0, transitioning from a failed simple CNN classifier to a model capable of achieving 91% accuracy in distinguishing subtle colorimetric imbalances.

III. PROPOSED SYSTEM

1. Dataset Acquisition and Preprocessing

A high-quality dataset is the cornerstone of any reliable computer-aided diagnostic system. This study utilized tongue images sourced from the TonguExpert platform, which contains a diverse repository of 5,992 samples with a documented male-to-female ratio of 1:1.26 and an average age of 46.55 ± 13.21 years. From the initial collection, 1,613 images with verified clinical mapping were identified for further processing.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

The following systematic procedures were implemented to prepare the data for the deep learning pipeline:

1) Selection and Filtering: We refined the mapped collection by selecting 375 high-quality samples specifically chosen for their clear representation of the target colorimetric properties: White, Light Yellow, and Yellow.

2) Image Standardization:All selected images were standardized to a resolution of 256 × 256 pixels to ensure uniform dimensions and computational compatibility with the deep learning architectures used in this study.

3) Data Augmentation: To mitigate overfitting and improve generalization, the dataset was expanded from 375 to 1,500 augmented images. This was achieved by rotating each of the 375 original images four times at 90-degree increments and varying the image brightness to simulate different clinical environments.

4) Class Distribution: The final augmented training set was organized into three distinct clinical categories: 444 images for "White", 644 for "Light Yellow", and 412 for "Yellow".

5) Normalization: Raw pixel values were used to train the model, preserving subtle color gradients crucial for detecting health imbalances.

The dataset was partitioned into training (80%) and testing (20%) sets for both Phase-I and Phase-II, ensuring class balance across the White, Light Yellow, and Yellow categories through stratified sampling.

2. Phase 1: Tongue Image Segmentation using U-Net

The initial objective of our pipeline involved separating the tongue's surface from peripheral oral structures, such as dental regions and the lips. For this purpose, we adopted a U-Net configuration, which is a specialized convolutional framework renowned for its efficacy in medical pixel-level classification. The model's mirrored architecture is instrumental in merging high-level semantic data with precise spatial coordinates through the use of skip connections.

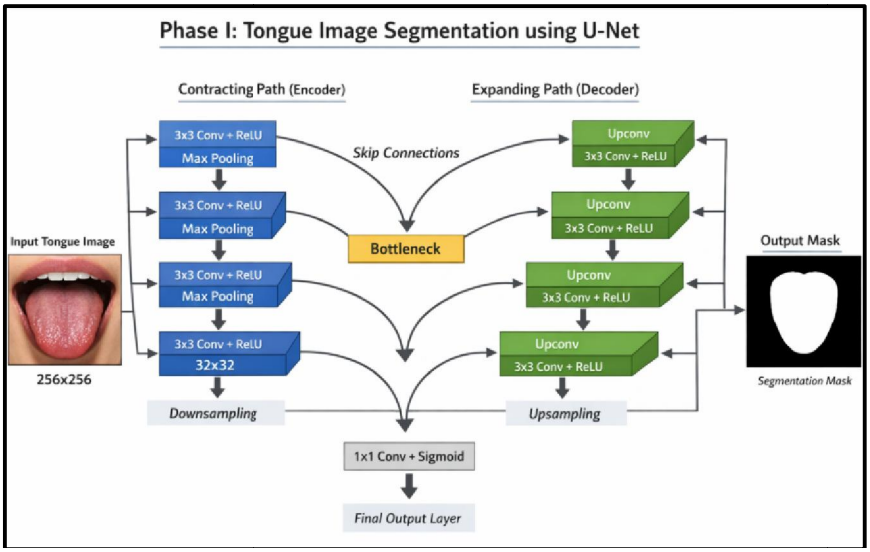

Fig.2. U-Net architecture used for tongue image segmentation in Phase1

The segmentation process was conducted through several distinct architectural layers:

1) Down-sampling (Encoder Path): The network initially condensed the input data by applying dual 3x3 convolutional kernels followed by ReLU non-linearity. We utilized 2x2 max-pooling layers to shrink the 256x256 spatial grid, which allowed the system to focus on the most prominent features of the tongue body.

2) Up-sampling (Decoder Path): To reconstruct the full-resolution image, the decoder employed transposed convolutions and expansion layers.This phase effectively projected the extracted feature maps back into the original image dimensions to define the tongue's geometry.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

3) Information Fusion:A vital aspect of our design was the integration of skip connections, which bridged the gap between the encoder and decoder. By concatenating features directly across the network, we prevented the loss of subtle boundary details, ensuring the tongue's edges were sharply defined.
4) Segmented Output:The architecture terminated with a 1x1 convolution utilizing a sigmoid activation. This layer generated a binary mask where each pixel was classified as either "tongue" or "background".
This segmentation methodology demonstrated significant reliability, reaching a 97% accuracy level during both the training and evaluation stages. By producing a clean Region of Interest (ROI), this phase ensured the classification model received high-quality inputs, free from non-tongue interference.

3. Phase 2: Classification Framework - Model Evolution

The second phase of the pipeline focused on categorizing the segmented tongue images into three clinical classes: White, Light Yellow, and Yellow. Our experimental approach evolved through two distinct stages as we sought to maximize diagnostic reliability.

3.1. Simple CNN Architecture (Initial Pipeline):

Our first attempt utilized a standard Convolutional Neural Network (CNN) architecture. This model was designed with a 4-channel input, incorporating both the RGB data and the binary mask generated in Phase I. However, this configuration proved insufficient for the task, as the basic convolutional layers failed to capture the subtle colorimetric gradients required to distinguish between similar categories like light yellow and yellow. Consequently, the model failed to reach the target performance metrics necessary for clinical support.

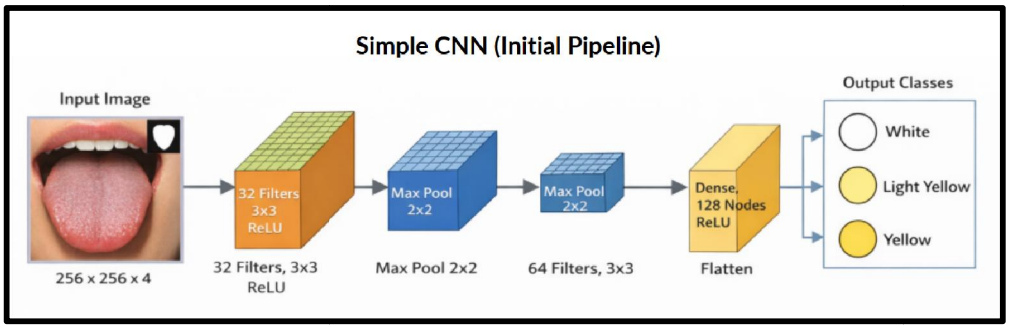

Fig. 3. Initial CNN model structure employed for tongue color classification.

3.2. EfficientNetB0 Architecture (Proposed Framework):

To address the limitations of the initial model, we transitioned to the EfficientNetB0 architecture. Unlike traditional models that scale dimensions arbitrarily, EfficientNetB0 employs compound scaling, which uniformly scales network depth, width, and resolution using a fixed set of coefficients. This balanced approach allows the network to extract richer feature representations from the tongue's surface while maintaining computational efficiency.

DOI: 10.48175/IJARSCT-32440

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

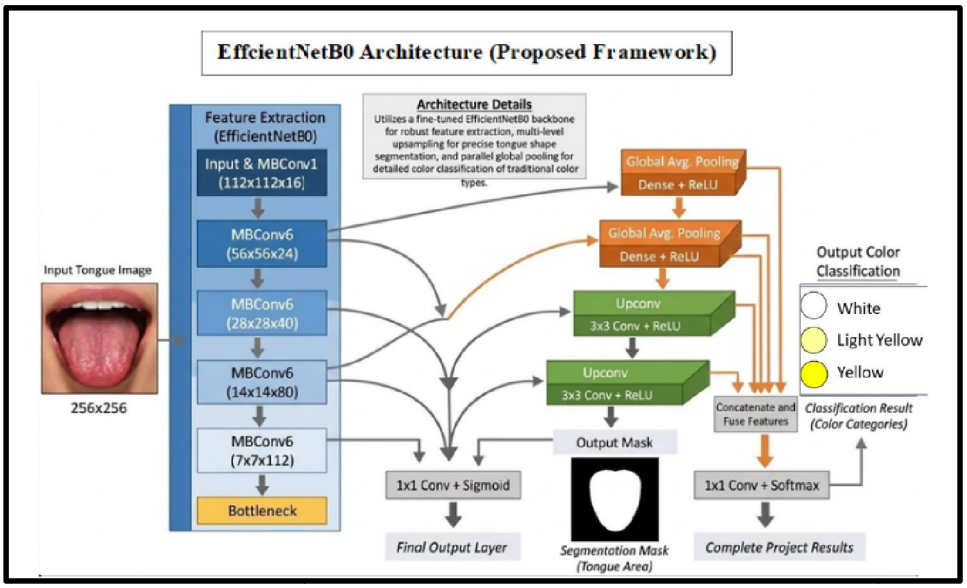

Fig. 4. EfficientNetB0 framework for tongue color classification.

The EfficientNetB0 model was integrated as the primary classifier, processing the 256x256 segmented Region of Interest (ROI). This architecture effectively minimized the semantic overlap between color classes, leading to a significant performance boost. By leveraging the model's MBConv (Mobile Inverted Bottleneck Convolution) blocks, the system achieved a final classification accuracy of 91%. This evolution from a basic CNN to a compound-scaled model underscores the necessity of advanced feature extraction for high-fidelity tongue color analysis.

4. Training Configuration and Experimental Setup

To facilitate the replication of this study, we have outlined the specific technical parameters and the computational environment used for model optimization.

1) Computational Infrastructure: The entire experimental workflow was implemented using the Google Colab platform, employing a Python-based development environment. We utilized the TensorFlow 2.x ecosystem along with the Keras library for building and executing the deep learning architectures, ensuring high-speed processing through GPU acceleration.

2) Input Specifications and Batching Logic: All images fed into the EfficientNetB0 classifier were resized to a fixed 256 × 256 pixel resolution. We maintained a Batch Size of 8, a value chosen to ensure stable weight updates while operating within the memory limits of the cloud-based hardware.

3) Optimization Strategy: The Adam optimizer was selected for its superior ability to handle sparse gradients in deep networks. To address the three-class objective, we employed Sparse Categorical Cross-Entropy as our primary loss function, which efficiently measures the discrepancy between the predicted and actual tongue color labels.

4) Training Protocol and Convergence: The model followed a dual-phase training schedule to reach peak performance. We conducted an initial learning phase of 15 epochs, followed by a rigorous refinement stage of 60 epochs. The total limit was set at 75 epochs, although we integrated automated logic to halt the process if the loss values stabilized, thereby protecting the model from overfitting.

5) Evaluation Partitioning: For both the segmentation and classification phases, we strictly adhered to an 80-20 data split. This ensured that a significant majority of the 1,500 augmented images was dedicated to training (80%), while the remaining 20% provided a completely independent set for verifying the final diagnostic accuracy.

DOI: 10.48175/IJARSCT-32440

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

IV. RESULTS AND DISCUSSION

The performance of the proposed two-stage diagnostic framework was evaluated using a comprehensive suite of pixel-level and class-level metrics. The results confirm the high efficacy of using a U-Net backbone for anatomical isolation followed by a compound-scaled classifier for colorimetric analysis.

1. Evaluation of Phase I: Tongue Image Segmentation

The U-Net model demonstrated exceptional precision in distinguishing the tongue body from the background oral environment. Analysis of the Pixel-wise Confusion Matrix reveals the model's high true positive and true negative rates across millions of evaluated pixels. Specifically, the model correctly identified 51,433,064 mask pixels and 45,253,436 background pixels, with a very low margin of error.

The quantitative performance of the segmentation phase is summarized below:

Model	Accuracy	Precision	Recall	F-1 Score
U-Net	0.9835	0.9857	0.9834	0.9845

Fig. 5. Performance metrics achieved for segmentation phase.

Further breakdown of the normalized confusion matrix indicates a Background Accuracy of 98.38% and a Mask Accuracy of 98.34%. These nearly identical values prove that the model is highly balanced and robust, maintaining consistent performance across both target and non-target regions. This level of precision (approximately 98.4%) ensures that the Region of Interest (ROI) extracted for the second phase is anatomically accurate, minimizing noise that could lead to classification errors.

2. Performance Assessment of the Baseline CNN

Before implementing the proposed framework, a standard CNN was evaluated to establish a performance baseline. This phase identified the architectural limitations that necessitated a more advanced model:

Learning Plateaus: The model reached an early accuracy peak below 0.75, failing to distinguish highly similar colorimetric features in tongue coatings.

Generalization Issues: A persistent gap between training and validation (peaking at 0.70) indicated a failure to adapt to unseen data.

Convergence Failure: High validation loss oscillations near 0.60 after 30 epochs confirmed that simple convolutional blocks could not minimize semantic overlaps effectively.

These results technically justified the transition to EfficientNetB0, leveraging compound scaling to achieve the precision required for medical diagnostic support.

Fig. 6. Training and Validation Accuracy for the initial Simple CNN model, showing performance stagnation below desired thresholds.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

Fig. 6. Training and Validation Loss for the initial Simple CNN model, illustrating suboptimal convergence during the preliminary classification attempt.

3. Evaluation of Phase II: EfficientNetB0 Classification Performance

The transition to the EfficientNetB0 architecture provided a substantial increase in diagnostic precision compared to the initial baseline. By utilizing compound scaling and Mobile Inverted Bottleneck Convolutions (MBConv), the model effectively learned to distinguish the subtle colorimetric gradients present in segmented tongue images.

1. Quantitative Performance Metrics: The final classification framework achieved a high level of accuracy across the three target classes.

• White Class Accuracy: 97%
• Yellow Class Accuracy: 90%
• Light Yellow Class Accuracy: 88%

The overall accuracy reached a peak of 91% during the final training phase. The high precision in the "White" category (97%) is particularly significant, as it represents the baseline healthy or mild state, ensuring few false alarms for healthy subjects.

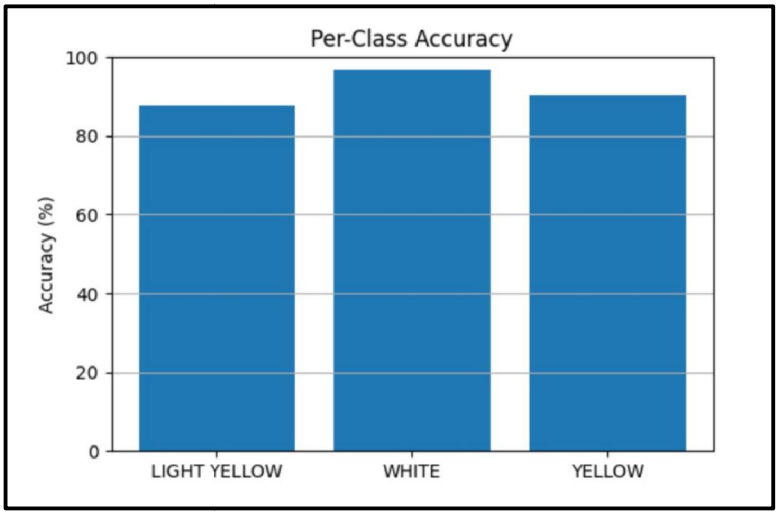

Fig. 7. Per-Class Accuracy bar chart highlighting the superior performance in detecting White and Yellow coatings.

2. Confusion Matrix Analysis: The multiclass confusion matrix illustrates the model's reliability.

a) The model correctly identified 65 out of 66 "White" samples, showing near-perfect sensitivity for this class.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

b) For the "Yellow" category, 56 samples were correctly classified, with only a small margin of confusion with the "Light Yellow" class.

c) The "Light Yellow" class achieved 84 correct predictions, though it exhibited some semantic overlap with the "Yellow" class (8 misclassifications) due to the close proximity of these colours in the pixel space.

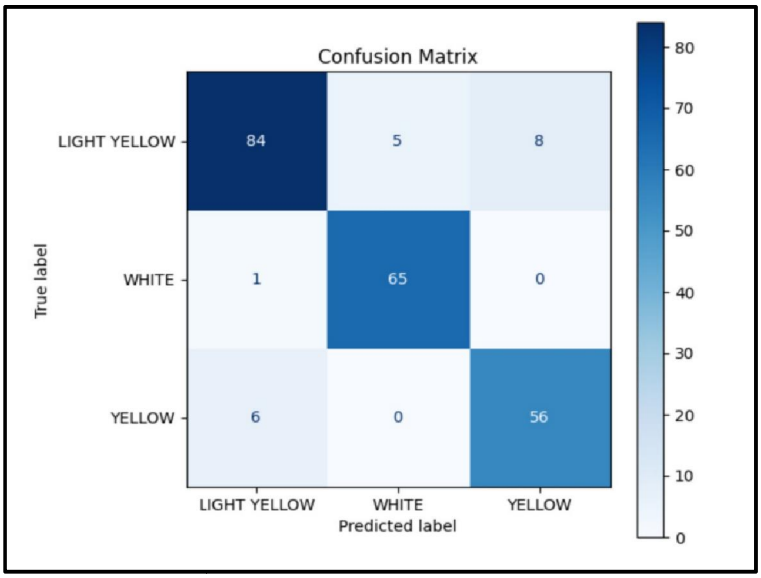

Fig. 8. Confusion Matrix for the EfficientNetB0 model, showing high diagonal values across all three color classes.

3. Training and Convergence Trends: The Training vs. Validation Accuracy and Loss curves for the EfficientNetB0 model (Fig. 9 and Fig. 10) show a much smoother convergence than the initial CNN.

a) The training accuracy steadily ascended to over 85% by epoch 30, with validation accuracy consistently mirroring this trend and reaching higher peaks of 90% due to the efficacy of the Test-Time Augmentation (TTA) strategy.

b) The loss values dropped significantly from an initial 1.2 to a stable minimum near 0.3, indicating that the Sparse Categorical Cross-Entropy function effectively guided the model toward a robust solution.

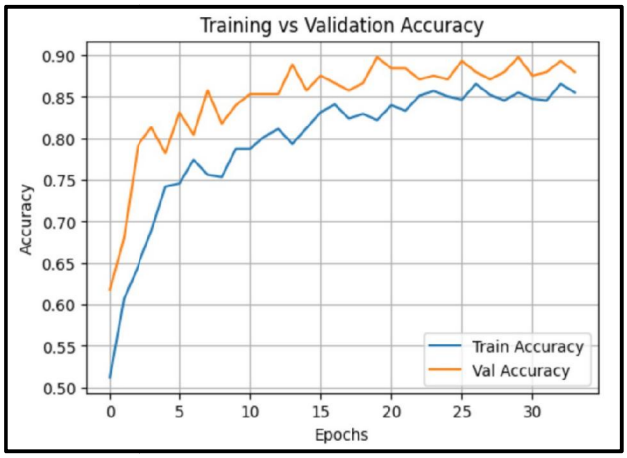

Fig. 9. Training and Validation Accuracy curves for the EfficientNetB0 framework during the final 75-epoch run.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

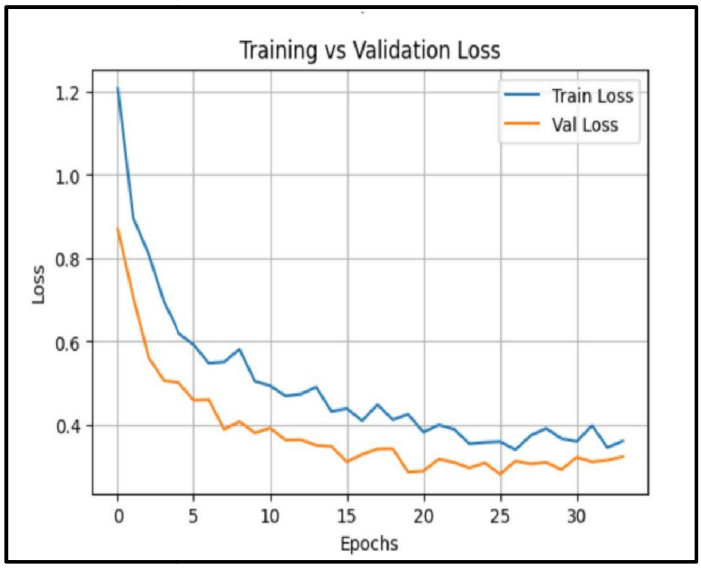

Fig. 10. Training and Validation Loss curves for the EfficientNetB0 framework showing successful model convergence.

V. CONCLUSION AND FUTURE WORK

This research successfully developed an automated, two-phase diagnostic framework for health assessment through tongue image analysis. By implementing a U-Net architecture, we achieved a high-precision segmentation rate of 97%, effectively isolating the tongue body from complex oral backgrounds to create a standardized Region of Interest (ROI). Furthermore, the transition to an EfficientNetB0 classification model addressed the limitations of standard CNNs, providing the compound scaling necessary to distinguish subtle colorimetric variations. The resulting system achieved a 91% accuracy rate in categorizing tongue colors into medically significant classes: White, Light Yellow, and Yellow. These findings demonstrate that deep learning can provide an objective, non-invasive tool to assist medical practitioners in preliminary health screenings, reducing the subjectivity inherent in traditional manual examinations.

While the current framework provides robust results for colorimetric classification, several areas remain for further enhancement. Future iterations of this research will focus on expanding the dataset to include a wider range of clinical categories, such as "crimson" or "cyan" coatings, to provide a more comprehensive diagnostic profile. Additionally, we aim to integrate environmental light-correction algorithms to further improve the model's resilience to varying capture conditions. Finally, the deployment of this architecture into a lightweight mobile application will be explored, enabling real-time, accessible health monitoring for users in remote or resource-constrained settings.

VI. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their respected guide Mr. P. R. Krishna Prasad for the continuous support, valuable suggestions, and insightful guidance throughout the course of this work. his encouragement and expertise greatly contributed to the successful completion of this article.

We are also thankful to the Project Coordinator, Dr.N.Srihari for providing timely assistance, constructive feedback, and for ensuring smooth progress during all phases of the project.

Our heartfelt thanks go to the Head of the Department, Dr.V.Ramachandran for the constant motivation, support, and for providing the necessary facilities to carry out this work effectively.

We extend our deep appreciation to the Principal, Dr.Y.Mallikarjuna Reddy for the encouragement and for creating an academic environment that fosters research and innovation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, April 2026 Impact Factor: 8.2

Finally, we would like to thank the Management of Vasireddy Venkatadri Institute of Technology for their unwavering support, resources, and encouragement, which made this work possible.

REFERENCES

[1] W. Liu, J. Chen, B. Liu, W. Hu, X. Wu, and H. Zhou, "Tongue image segmentation and tongue color classification based on deep learning," Digital Chinese Medicine, vol. 5, no. 3, pp. 253-263, Sep. 2022.
[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," inInternational Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234-241.
[3] J. Zhou et al., "TongueNet: A Precise and Fast Tongue Segmentation System Using U-Net with a Morphological Processing Layer," Applied Sciences, vol. 9, no. 15, p. 3128, 2019.
[4] L. Yao et al., "HPA - UNet: A Hybrid Post – ProcessingAttention U-Net for Tongue Segmentation," IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 9, pp. 1102-1114, 2024.
[5] Z. Huang et al., "Tongue Image Segmentation Method Based on the VDAU-Net Model," IEEE Access, vol. 13, pp. 2485-2495, 2025.
[6] M. Tan andQ. V. Le, "EfficientNet: RethinkingModel Scaling for Convolutional Neural Networks," arXiv preprint arXiv:1905.11946, 2019.
[7] B. Yan et al., "A two-step deep learning framework to improve tongue color classification performance," Mathematics, vol. 10, no. 22, p. 4286, 2022.
[8] T. Kawanabe et al.,"Quantifying tongue body and coating color information using K-means clustering and deep learning," Frontiers in Physiology, vol. 16, p. 1527751, 2025.
[9] W. H. Chang et al., "Tongue feature dataset construction and real-time detection," PLoS ONE, vol. 19, no. 3, p. e0296070, 2024.
[10] D. Zhang et al., "TongueCoating GradingIdentification Using Deep Learning for Hyperspectral Imaging Data," IEEE Access, vol. 11, pp. 93151-93159, 2023.
[11] L. Zhong et al., "Deep learning-based recognition of stained tongue coating images," Digital Chinese Medicine, vol. 7,no. 2, pp. 129-136, 2024.
[12] J. Ni et al., "Research on Tongue Image Segmentation and Classification Methods Based on Deep Learning and Machine Learning," Information, vol. 16, no. 5, p. 357, 2025.
[13] M. Marhamati et al., "LAIU-Net: A learning-to-augment incorporated robust U-Net for depressed humans' tongue segmentation," Displays, vol. 76, p. 102345, 2023.