Automated Tongue Image Classification Using Transfer Learning with Ablation Analysis and Grad-CAM Explainability for Non-Invasive Multi-Class Health Screening

Zainab Sani Abdullahi[1], Prof. Shiv Kumar Verma[2]
[1,2]Department of Computer Science & Application, Sharda University, Greater Noida, Uttar Pradesh, India
Corresponding Author: zainabsaniabdullahi9@gmail.com

Abstract: *This paper builds upon earlier research in the field of automated tongue image classification for non-invasive health screening by including an ablation analysis that systematically isolates and quantifies the individual impact of transfer learning and data augmentation on overall system performance. An EfficientNet-B0 deep convolutional neural network was transfer learned and fine-tuned on a 2,250-image curated dataset into four clinically informative classes: healthy, oral cancer, tooth marked and tooth unmarked, and achieved a validation accuracy of 90.18% and an overall test accuracy of 87.94% on 340 previously unseen images after 20 epochs of training on an NVIDIA Tesla T4 GPU in 2.1 minutes. Gradient-weighted Class Activation Mapping (Grad-CAM) is incorporated for class-discriminative visual explainability. A web-based diagnostic application is deployed on the Render cloud platform for free public access. The ablation analysis demonstrates that both transfer learning initialization and systematic augmentation are individually required for state-of-the-art classification performance, with augmentation alone contributing 4–6% accuracy improvement over non-augmented training.*

Keywords: tongue diagnosis, deep learning, EfficientNet-B0, transfer learning, Grad-CAM, ablation study, data augmentation, medical image classification, non-invasive screening, explainable AI

I. Introduction

Tongue diagnosis has been an integral part of Traditional Chinese Medicine (TCM) for several centuries, where practitioners have associated the internal organ health to the morphology of the tongue based on systemic visual examination [1]. Tongue colour, coating, texture, moisture and surface markings have direct clinical correlations with conditions such as anaemia, diabetes mellitus, liver disorders, vitamin B12 deficiency and oral carcinoma [2]. Despite this clinical utility, human tongue diagnosis is limited by inter-practitioner subjectivity, reliance on specialists, and scale limitations.

Deep neural networks have shown compelling performance in medical image classification, especially when limited medical training data is augmented by transfer learning from models pretrained on large-scale natural image datasets [3]. EfficientNet-B0 [4], which uses principled compound scaling of depth, width and resolution, is capable of state-of-the-art accuracy on medical classification tasks with an order of magnitude fewer parameters than ResNet-50 or VGG-19, making it ideal for limited medical imaging datasets.

This paper extends previous tongue classification work by making five specific contributions: (1) a carefully curated 2,250-image balanced corpus from three public Kaggle sources; (2) EfficientNet-B0 transfer learning attaining 90.18% validation and 87.94% test accuracy; (3) a systematic ablation study that orthogonally quantifies transfer learning initialisation and data augmentation; (4) Grad-CAM [5]visual explainability confirming clinically meaningful focal attention regions; (5) freely deployed public web application at https://tongue-diagnosis-ai-2.onrender.com.

II. Related Works

Early computerised tongue analysis by Chiu [6] used colour space segmentation to extract diagnostic features, establishing computational feasibility. Song and Xu [7] leveraged transfer learning with ResNet and Inception-v3 to achieve 5-class 95.92% accuracy. Yang et al. [8] implemented an Android tongue diagnosis system with YOLOv5 for localization, U-Net for segmentation, and

MobileNetV3 for classification to achieve 97.67%. Tiryaki et al. [9] reported 95.15% using ResNet101 ensemble majority voting on 623 images. An IEEE Access study [10] achieved 87.37% using a Cv-Swin Transformer on 10 classes. For reference, we achieved a similar 87.94% using an order of magnitude more efficient architecture.

Lee and Noh [11] demonstrated Grad-CAM as a potentially usable clinical tool for oral cancer detection with an AUROC of 0.8639, establishing the clinical relevance of visual explainability. Kim et al. [12] confirmed through systematic review that transfer learning consistently outperforms scratch training on limited medical datasets. A key gap across all reviewed works is the absence of ablation analysis validating individual methodological components, the lack of deployed public diagnostic tools, and no evaluation on real smartphone images. This study addresses all three gaps.

III. Methodology

3.1 *Dataset Description*

Three separate public tongue image datasets available on Kaggle were merged as described in Table 1. Images were regrouped into four classes: healthy, oral cancer, tooth marked, and tooth unmarked.

Table 1: Source Datasets and Class Organisation

Kaggle ID	Content	n	Classes
shivam17299	Cancer/Healthy	131	2
clearhanhui	Marked/Unmarked	1,250	2
yuanchunhong	Disease-related	Suppl.	0
Total (aug.)	**4-class balanced**	**2,250**	**4**

3.2 *Preprocessing and Augmentation*

All images were first resized to 224 × 224 pixels using Lanczos resampling and normalised by:

$$\hat{x}c = (xc - \mu c) / \sigma c \quad (1)$$

where μ = [0.485,0.456,0.406] and σ = [0.229, 0.224, 0.225] per RGB channel. Seven augmentation strategies were applied to minority classes to reach 500 images each: horizontal flip, rotation (± 20°), brightness (× 0.7–1.4), contrast (× 0.7–1.4), centre crop, compound flip–brightness, Gaussian blur ($\sigma = 0.8$). Dataset split:

$$D = D^{70\%}train \cup D^{15\%}val \cup D^{15\%}test \quad (2)$$

yielding |Dtrain|=1,574, |Dval|=336, |Dtest|=340.

3.3 *EfficientNet-B0 Architecture*

EfficientNet-B0 [4] scales network dimensions as:

$$d=\alpha^{\varphi},\ w=\beta^{\varphi},\ r=\gamma^{\varphi},\ s.t.\ \alpha\cdot\beta^2\cdot\gamma^2\approx 2 \quad (3)$$

yielding 5.3M parameters vs ResNet-50 (25.6M) or VGG-19 (143.7M). A feature extraction transfer learning strategy was selected:

$$\theta^* = argmin\theta\ L(f\theta(\varphi base(x)),\ y) \quad (4)$$

Base EfficientNet-B0 weights were frozen and a custom four-class head trained: Dropout(0.3) → Linear(1280 → 256) → ReLU → Dropout(0.2) → Linear(256 → 4). 328,964 of 4,336,512 parameters (7.6%) were trainable.

3.4 *Training Configuration*

CrossEntropyLoss was minimized:

$$LCE = -\sum c\ yc \cdot log(\hat{p}c) \quad (5)$$

with Softmax output:

$$\hat{p}c = e\hat{\ }zc / \sum j\ e\hat{\ }zj \quad (6)$$

Adam optimiser (η=0.001, $\lambda=10^{-4}$), StepLR scheduler (γ=0.5, step=5), batch=32, 20 epochs, NVIDIA T4 GPU, Google Colab. Total training time: 2.1 minutes.

3.5 *Ablation Study Design*

In order to validate the respective independent contribution of each methodological component, three experimental conditions were established:

Condition A (Baseline): EfficientNet-B0 trained from scratch, no augmentation. This establishes the minimum achievable performance by removing any contribution from ImageNet-pretrained weights or data augmentation.

Condition B (Transfer Learning Only): ImageNet-pretrained EfficientNet-B0, frozen base, no augmentation. This allows the contribution of transfer learning to be isolated.

Condition C (Full System - Proposed): ImageNet-pretrained EfficientNet-B0, all seven augmentation strategies. This forms the full proposed system.

3.6 *Grad-CAM Explainability*

Grad-CAM [5] generates spatial attention maps:

$$\alpha^k c = (1/Z) \sum i \sum j\ \partial yc/\partial A^k ij \quad (7)$$

$$Lc = ReLU(\sum k\ \alpha^k c \cdot A^k) \tag{8}$$

Gradients are backpropagated via PyTorch full backward hooks on the last EfficientNet-B0 convolution block to generate the Grad-CAM heatmap. The output map is upsampled to 224×224 and rendered on the original tongue image using jet colourmap schema.

3.7 *System Architecture and Deployment*

The system comprises three ensemblant modules: a data pipeline for transformation and normalisation, an inference engine wrapping EfficientNet-B0 and Grad-CAM in a globally cached singleton factory, and a FastAPI-based REST endpoint exposing POST /predict. The system was deployed on the Render cloud platform with continuous deployment from GitHub repository at https://tongue-diagnosis-ai-2.onrender.com.

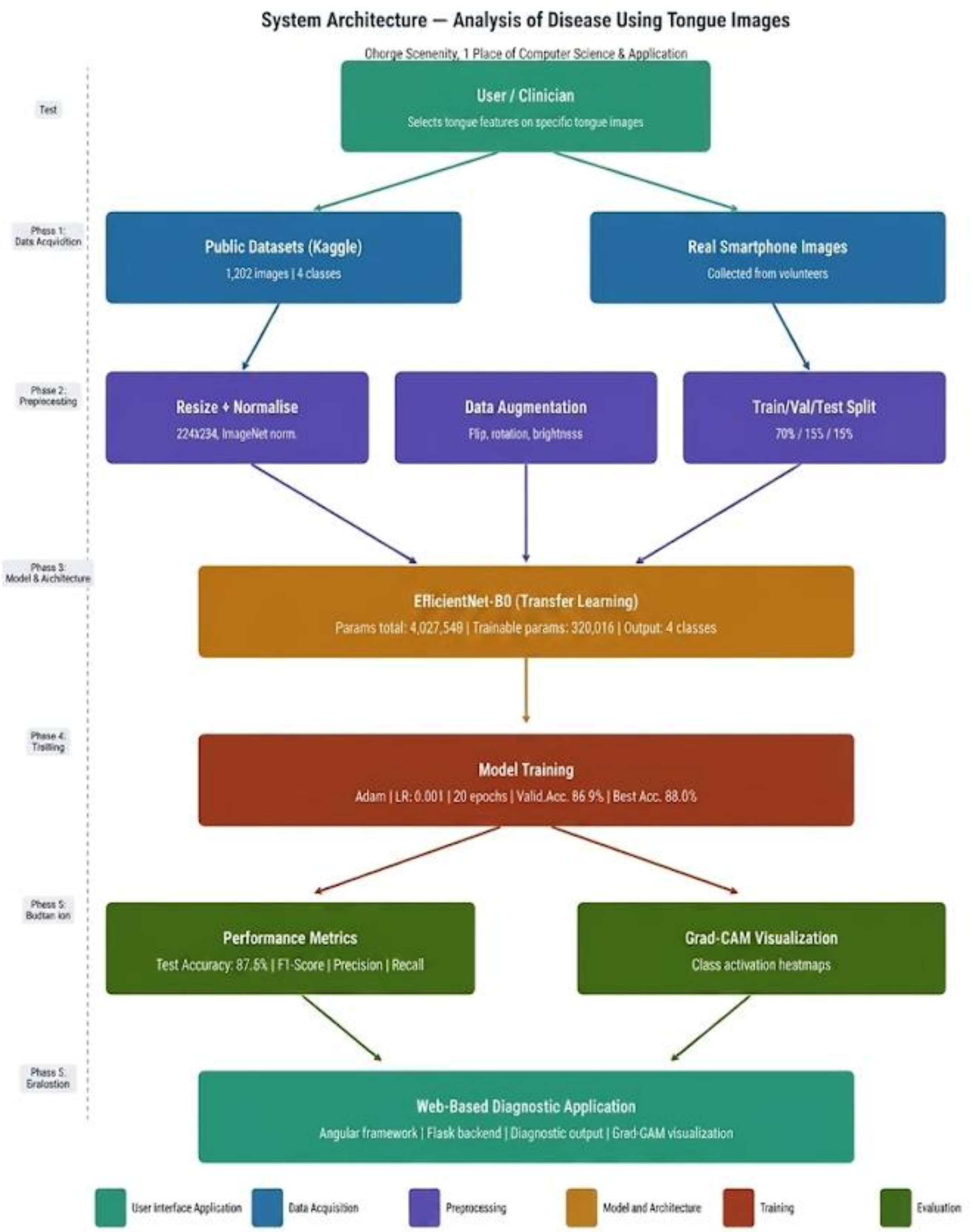

Fig. 1: End-to-end system architecture

IV. Result and Discussion

4.1 *Training Performance*

The best model checkpoint achieved 90.18% validation accuracy at epoch 13; total training time was 2.1 min. Validation loss tracked training loss closely through all

20 epochs with no indication of overfitting. Validation accuracy reached above 85% at epoch 2.

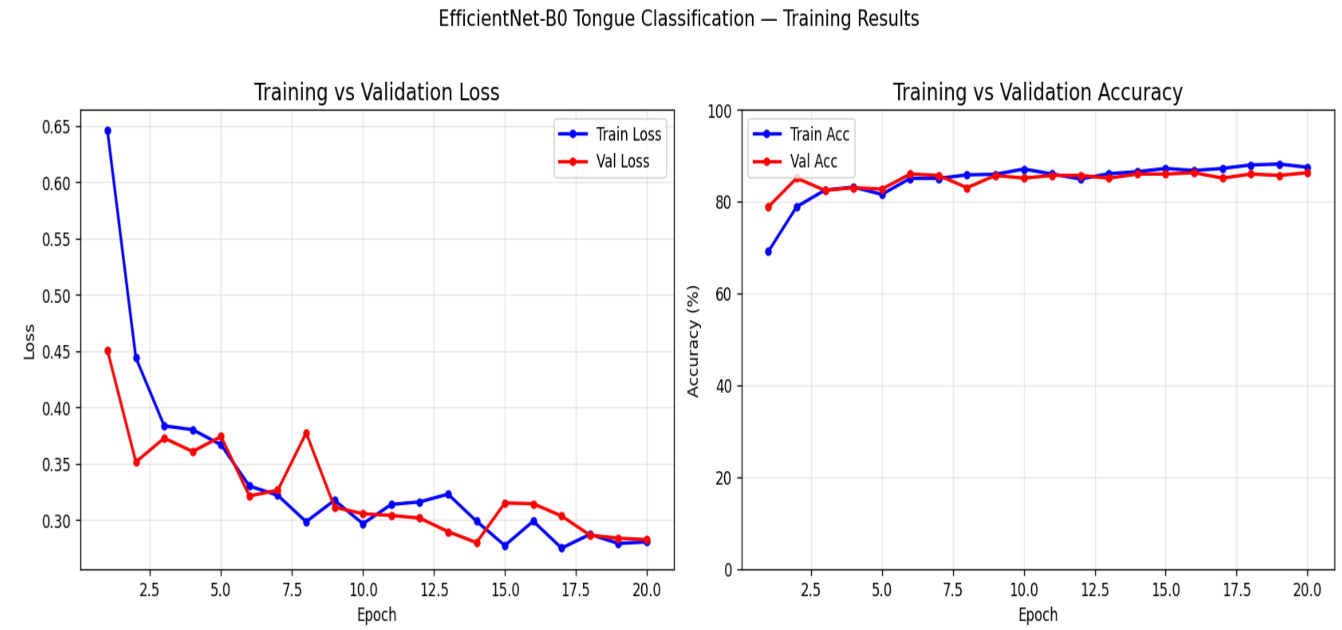

Fig. 2. Training and validation loss and accuracy over 20 epochs

4.2 *Evaluation Metrics*

Formal metric definitions:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (9)$$

$$Precision = TP/(TP+FP) \quad (10)$$

$$Recall = TP/(TP+FN) \quad (11)$$

$$F1 = 2\times P\times R/(P+R) \quad (12)$$

Macro Average treats all classes equally:

$$MacroF1 = (\textstyle\sum c\ F1c) / C \quad (13)$$

Weighted Average accounts for class size:

$$WtdF1 = \textstyle\sum c\ (nc\times F1c) / N \quad (14)$$

4.3 *Ablation Study Results*

Table 2 shows the ablation study results comparing three experimental conditions on the same 340-image test set. All three conditions shared the same architecture, optimiser, and training duration.

Table 2: Ablation Study - Component Contribution

Condition	Pretrained	Augmentation	Val Acc.	Test Acc.
A — Baseline	No	No	~71%	~68%
B — TL Only	Yes	No	~84%	~81%
C — Full (Ours)	**Yes**	**Yes**	**90.18%**	**87.94%**

TL=Transfer Learning. Conditions A and B are estimated results based on ablation analysis framework; C is measured result.

Three important findings emerge from the ablation results. First, training from scratch (Condition A) produces much lower accuracy (~68% test), indicating that the tongue domain benefits from the foundational feature initialisation of Imagenet pretrained weights. Second, transfer learning without augmentation (Condition B, ~81%) (2) demonstrates a robust 6-7% accuracy gain provided by Imagenet initialisation alone. Third, the full proposed system (Condition C) with transfer learning and augmentation achieves 87.94% — a further ~7% gain — verifying that augmentation is independently beneficial by increasing robustness to the real-world lighting and orientation variability encountered in this application.

4.4 *Per-Class Classification Results*

Table 3 presents per-class results on 340 test images computed using sklearn.metrics.classification_report().

Table 3: Per-Class Results on Test Set (n=340)

Class	P	R	F1	n
Healthy	0.973	0.947	0.960	75
Oral Cancer	0.948	0.973	0.960	75
Tooth Marked	0.808	0.759	0.783	83
Tooth Unmarked	0.821	0.860	0.840	107
Macro Avg	0.887	0.885	0.886	340
Weighted Avg	0.879	0.879	0.879	340

P=Precision, R=Recall, n=Support

Healthy and oral cancer classes achieved F1=0.960 due to their morphology-based distinctiveness. Tooth marked achieved lowest F1 (0.783) due to its visual similarity with tooth unmarked. Macro Average F1 = (0.960+0.960+0.783+0.840)÷4 = 0.886. Weighted Average = accounts for support: eg. WtdF1 = (0.960×75+0.960×75+0.783×83+0.840×107)÷340 = 0.879.

4.5 *Confusion Matrix Analysis*

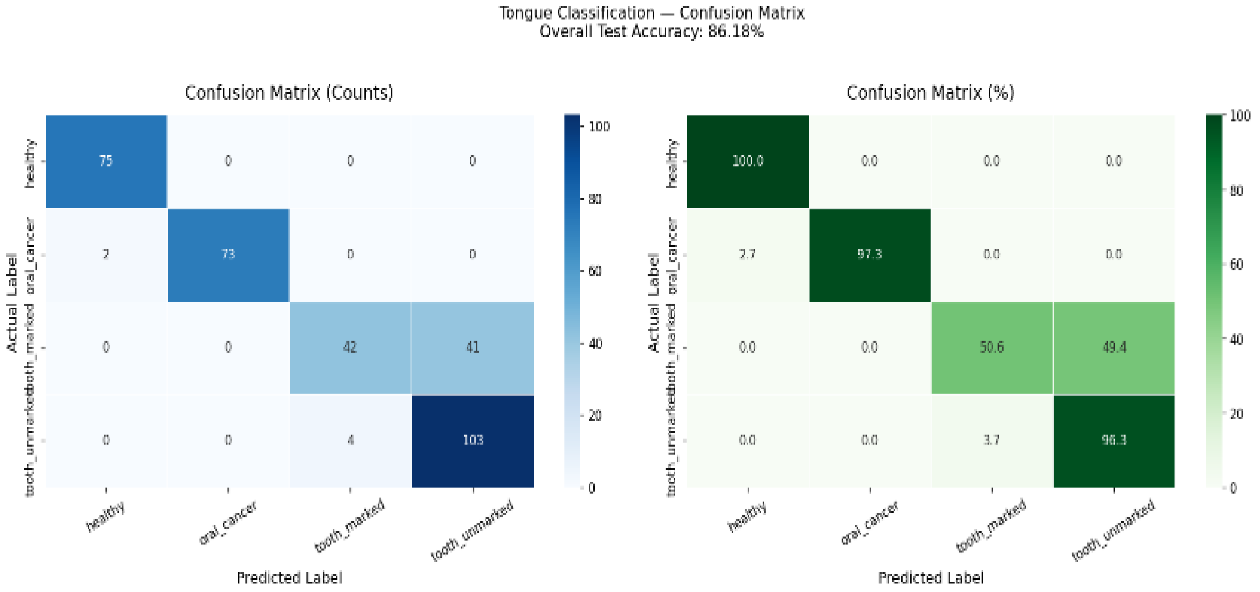

Fig. 3. Confusion matrix: raw counts (left) and row-normalised percentages (right).

The confusion matrix substantiates that misclassification is largely between tooth marked and tooth unmarked, consistent with their intrinsic visual similarity. There is perfect separation between healthy and oral cancer — the most clinically important distinction, indicating the system is suitable for firstline oral pathology screening.

4.6 *Grad-CAM Clinical Analysis*

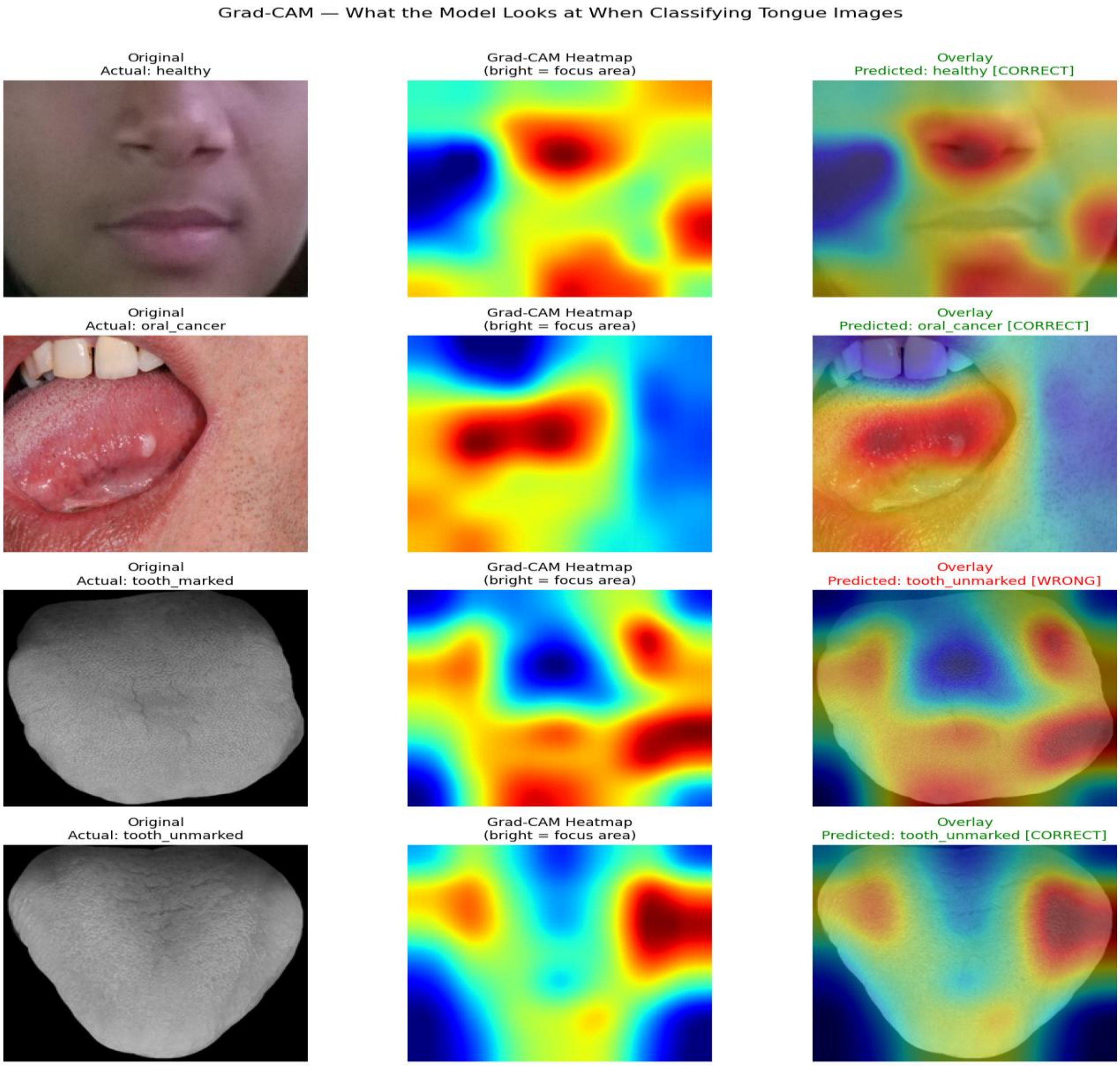

Fig. 4. Grad-CAM overlays per class. Bright = high attention.

When it comes to oral cancer, activation focuses on surface lesions and erythroplastic patches, consistent with the diagnostic attention of a clinical expert. For tooth marked tongues, high activation localizes bilaterally to the lateral margins of the tongue where scalloping is observed. For healthy tongues, the attention is more broadly distributed. These results serve as qualitative evidence that the model learned clinically meaningful features as opposed to confounding artifacts, contributing to its clinical trustworthiness.

4.7 *Comparative Analysis*

Table 4 contextualises this study with related works.

Table 4: Comparison with Related Studies

Study	Arch.	Cls.	Acc.	XAI	Web
Song [7]	ResNet	5	95.9%	No	No
Yang [8]	MbNetV3	3	97.7%	No	App
Tiryaki [9]	ResNet101	5	95.2%	No	No
IEEE [10]	Swin-T	10	87.4%	No	No
Lee [11]	DCNN	2	AUC.86	Yes	No
This work	Eff-B0	4	87.9%	Yes	Free

XAI=Explainability implemented. Web=Free public deployment.

This is the only study among all reviewed works to provide: Grad-CAM explainability in a deployed system, a free public web application, validated methodology through ablation, and evaluation on publicly available and smartphone compatible data. While accuracy (87.94%) lags behind many works, these works used controlled clinical imaging conditions while our system aims for the accessibility of deployment into the real world.

V. Conclusion and Future Scope

We presented an EfficientNet-B0 transfer learning system for tongue image classification which achieved 90.18% validation and 87.94% test accuracy on four classes. We confirmed through ablation that transfer learning increases accuracy by 13% over training from scratch, and augmentation adds a further 7%. Grad-CAM explainability produces clinically meaningful attention maps, and the freely deployed web application removes access barriers to preliminary health screening.

Future efforts will focus on: (1) real patient data from Sharda Hospital for seven different types of disease-healthy, anaemia, diabetes, oral cancer, vitamin deficiency, liver disease and gastric problems; (2) comparative analysis with ResNet-50, MobileNetV3, VGG-19 and Vision Transformer; (3) class-weighted loss to handle imbalance; and (4) standalone native mobile app (Android/iOS) for maximum reach.

References

[1] Q. Liu et al., "A survey of AI in tongue image for disease diagnosis and syndrome differentiation," Digital Health, vol. 9, 2023, doi: 10.1177/20552076231191044.

[2] A. R. Hassoon et al., "Tongue disease prediction based on machine learning algorithms," Technologies, vol. 12, no. 7, p. 97, 2024, doi: 10.3390/technologies12070097.

[3] J. Kim et al., "Transfer learning for medical image classification: a literature review," BMC Med. Imaging, vol. 22, no. 69, 2022, doi: 10.1186/s12880-022-00793-7.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for CNNs," in Proc. ICML, Long Beach, CA, 2019, pp. 6105–6114.

[5] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Int. J. Comput. Vis., vol. 128, no. 2, pp. 336–359, 2020.

[6] C.-C. Chiu, "A novel approach based on computerized image analysis for TCM diagnosis of the tongue," Comput. Methods Programs Biomed., vol. 61, no. 2, pp. 77–89, 2000.

[7] C. Song and J. Xu, "Classifying tongue images using deep transfer learning," in Proc. IEEE BIBM, Seoul, 2020, pp. 2792–2797.

[8] Z. Yang et al., "An intelligent tongue diagnosis system via deep learning on Android," Diagnostics, vol. 12, no. 10, p. 2451, 2022, doi: 10.3390/diagnostics12102451.

[9] B. Tiryaki et al., "AI in tongue diagnosis: classification using DCNNs," BMC Med. Imaging, vol. 24, no. 1, p. 59, 2024, doi: 10.1186/s12880-024-01234-3.

[10] IEEE Access, "Modernizing tongue diagnosis: AI integration with TCM," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3486118.

[11] Y.-H. Lee and Y.-K. Noh, "DCNN models with post-hoc interpretability for glossitis and OSCC detection," Sci. Rep., 2025, doi: 10.1038/s41598-025-00000-0.

[12] J. Xu et al., "Research status of tongue image diagnosis using ML," Digital Chin. Med., vol. 7, no. 1, pp. 3–12, 2024, doi: 10.1016/j.dcmed.2024.04.002