

Modal Deep Learning for Tongue Diagnosis

Niparat Boonkun*, Adisak Sangsongfa*,
Phayung Meesad**, and Noppadol Amdee*

Received: May 13, 2025
Revised: August 22, 2025
Accepted: August 22, 2025

* Corresponding Author: Niparat Boonkun, E-mail: niparatboo@mcru.ac.th

DOI: 10.14416/j.it.2026.v2.004

Abstract

Tongue diagnosis is a key component of traditional medical systems and increasingly relevant in intelligent healthcare. This study proposes a multi-modal approach integrating visual and textual data to classify tongue images into three diagnostic categories: Normal, HF Tip, and Abnormal. We compare the performance of three deep learning architectures: Convolutional Neural Network (CNN) for image-based classification, Long Short-Term Memory (LSTM) for text-only processing, and a hybrid CNN+LSTM model for multi-modal fusion. A 10-fold cross-validation was conducted on a curated dataset of 1,273 tongue images and corresponding medical descriptions. Results demonstrate that the fusion model significantly outperforms unimodal models with an accuracy of 98.11%, indicating the effectiveness of combining image and language features in automated diagnosis systems.

Keywords: Tongue Diagnosis, Deep Learning, Multi-modal Learning, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Medical Image and Text Fusion.

1. Introduction

Tongue diagnosis has long been a vital non-invasive diagnostic tool in traditional medical systems, particularly within Traditional Chinese Medicine (TCM) and its derivatives practiced across Asia, including Thailand. The condition of the tongue's palette reflects different pathophysiological processes such as organ disharmony, poor blood circulation, and inflammation affecting most of the body; hence, it is a pointer for the suitability of holistic and integrative health approaches [1]-[2].

Advanced digital imaging, machine learning, and natural language processing (NLP) have driven the remarkable change in traditional language diagnosis. The technologies of this type can carry out pattern recognition task automatically, remove the subjectivity in observations, integrate the metadata of clinical research, thus, the diagnosis can be more precise and definitive, reducing mistakes and quickly carrying out the treatment regimen. Multi-modal learning, which combines visual data (tongue images) with textual descriptions (clinical annotations or physician notes), has emerged as a powerful approach in medical AI [3].

These innovations are especially pertinent in Thailand, where traditional and modern medicine often coexist. Thailand's demographic diversity, cultural receptiveness to integrative health solutions, and widespread use of mobile health platforms make it a compelling case for deploying AI-based diagnostic tools in community healthcare and telemedicine environments [4].

The existing literature on image-based tongue diagnosis using convolutional neural networks (CNNs) and text-based classification using recurrent or transformer-based models has largely remained unimodal. In most cases, image-only methods achieve a diagnostic accuracy of 93-96%, while text-only models usually do not go beyond 85% due to the brevity and variability of clinical narratives. On the other hand, there is a limited number of studies that compare the performance of a unified multi-modal framework under the same experimental settings. Such separation of different modalities results in the incomplete overlap of the potential synergy between structured visual features and unstructured textual cues. Thus, the paper is a point to study and close the gap, which is

* Muban Chom Bueng Rajabhat University.

** King Mongkut's University of Technology North Bangkok, Thailand.

done by comparing and evaluating three deep learning methods—CNN (image only), LSTM (text only), and a fused CNN+LSTM model (image + text)—for the task of tongue condition classification into diagnostic categories, i.e. setting a stronger baseline and a novel contribution to automated tongue diagnosis research area.

2. Literature Review

The computerization of the study and interpretation of the tongue has turned this qualitative approach into a quantitative application. It has thus made it possible to have direct and understandable assessments. The steps for this conversion have been mainly established by developing image processing and pattern recognition technologies that have provided the scientific basis for comprehensive tongue analysis. Early innovations, such as the RGB-based segmentation system [5], have played a crucial role in this evolution by providing a standardized approach to clinical evaluation. Creating an Asian tongue image database further highlights the importance of demographic-specific analysis, particularly in diverse populations like Thailand [6].

Studies specific to Thailand have provided essential input regarding the frequency and nature of tongue disorders in different populations. For example, the multicenter survey of Dhanuthai et al. [7] described the lateral border of the tongue as the commonest site of oral mucosal lesions and noted that the reactive and inflammatory categories were the most prevalent types. The results of the study by Jankittivong et al. [8] have supported the mentioned point since they figured out that over 83% of the dental patients had some oral lesions and that coated tongues were the predominant group. Moreover, Aittwarapoj et al. [9] have offered a clinicopathological analysis in a Thai population which emphasized that oral potentially malignant disorders and squamous cell carcinoma were among the key clinical issues in tongue diagnosis. Epidemiological studies are an excellent example of the significance of the population-specific datasets in the enhancement of the diagnostic models.

Yet, research conducted outside of Thailand has also made a great contribution to this field. One such instances is that of Li et al. [10] who showed by combining tongue, face, and sublingual images together with clinical text the detection of diabetes becomes more accurate, thereby demonstrating the usefulness of the multimodal strategy. Correspondingly, the work of Jiang et al. [11] referred to as deep learning application in China's multi-label tongue image analysis reporting as a result a remarkably high accuracy but with a generalization issue inherent in the dataset homogeneity. Thirunavukkarasu et al. [12] dealt with the thermal-visible tongue image fusion, which enhanced the efficiency but required very expensive and specialized equipment and, therefore, was limited to scale. Such instances delineate the high points and setbacks of international efforts.

The comparison of the Thai and non-Thai studies not only spotlights the issues that population-specific data and the implementation of AI in Thailand are two strong points of this country but also that the studies done elsewhere can help to fuse the strategies and discover new modalities. However, limitations are still present in both contexts: most models are based on the training of narrow datasets, have no external validation, and in addition, the explainability feature provided is minimal. Our research work is a continuation of all these international projects where we have methodically carried out experiments on multimodal deep learning with a handpicked dataset with the aim of achieving a good balance between the research principles and the contextual suitability.

Beyond visual analysis, textual data such as clinical narratives and lifestyle factors (e.g., smoking and alcohol use) have been shown to correlate with specific tongue conditions. Jankittivong et al. [8] linked such factors to common lesions, emphasizing the diagnostic value of context-rich textual descriptions. Meanwhile, AI-driven platforms in Thai healthcare have successfully applied natural language processing (NLP) for clinical text analysis. For instance, Panyasorn et al. [13] employed Thai NLP techniques to assess hospitalization appropriateness with high precision, illustrating the broader applicability of AI in processing clinical text data.

In several studies, visual and textual data convergence in multi-modal analysis has shown superior diagnostic performance. Li et al. [10] integrated tongue, facial, and sublingual images with clinical text data to diagnose diabetes mellitus and impaired glucose regulation, achieving enhanced performance through modality fusion. Likewise, deep learning models such as CNNs have been used to classify tongue images into disease categories accurately. Thirunavukkarasu et al. [12] applied SVM classifiers to fused image data, attaining 88.12% accuracy in distinguishing diabetic and non-diabetic subjects.

Another promising avenue is thermal-visible image fusion, which combines blood flow and inflammation indicators from thermal images with morphological features from visible spectrum images. Liu et al. [14] employed discrete wavelet transform (DWT) fusion rules and deep learning classifiers to improve diagnostic precision. While effective, such methods require specialized equipment and complex preprocessing, which may limit widespread adoption.

Based on bibliometric analyses [15]-[16], the country of Thailand has become a leader in the healthcare AI field in the Southeast Asia region. The main factors contributing to this were identified as the supportive policies of the state and the large number of AI-specific research centers. The studies showcase the emergence of a solid national pattern of the use of AI in image-based diagnostics, natural language processing, and the implementation of AI technologies in public health systems, which was mainly driven by the policies. In addition, participatory disease detection systems, for instance, those being a part of the DoctorMe application [17], not only demonstrate the capability of Thailand in broadening early diagnosis but also show the possibility of reaching farther populations by utilizing AI-powered mobile health platforms.

The literature also explores the utility of data visualization and decision support systems in translating complex multi-modal data into actionable insights. For example, Mitranont et al. [18]-[19] created visualization tools and AI-powered systems to direct healthcare officials in utilizing empirical clinical data (HCID). These methods are vital in that they help AI models

maximize the clinical usefulness through interpretability and trust of the practitioner.

Still, all those improvements notwithstanding, every method has its downsides. Aside from being definite and duplicable, analysis based on images may be devoid of the clinical setting. However, the more semantically rich a medium, the more analysis of a text can have instances of subjectivity and a lack of consistency. Merging like-minded elements will remove such difficulties, and the most comprehensive diagnosing method will be proposed. Here, the deep learning network that particularly uses CNNs is an excellent tool for extracting the image's features -- while this is not the case for the usual machine learning models that offer interpretability with less efficiency in terms of computation. The methodology type relies on the diagnosis's difficulty level and the available resources for implementation.

Challenges, especially those related to data standardization, algorithmic fairness, and explanatory models, need to be resolved now, as that is the only way to assure that AI-driven healthcare remains equitable and transparent. Future research will focus on self-supervised learning, attention-based fusion architectures, and cross-modal transformers, potentially providing more gains in multi-modal diagnostic performance.

In conclusion, literature establishes a strong foundation for applying multi-modal learning to tongue diagnosis. However, systematic image and text data integration using deep neural architecture remains limited. This study addresses that gap by evaluating and comparing CNN, LSTM, and CNN+LSTM models on a labeled dataset, contributing new insights into designing intelligent, data-driven diagnostic systems for use in Thai and global healthcare contexts.

3. Research Methodology

This section introduces the stepwise process for developing and evaluating a multimodal tongue diagnosis system that uses images and texts as data. This approach involves preparing the dataset, preprocessing, designing the model, training steps, and evaluating parameters.

3.1 Dataset Composition

A dataset of 1,273 tongue images was carefully selected in cooperation with Ratchaburi Hospital under an institutional review protocol. All participants gave their informed consent, and the images were taken using standardized digital imaging equipment in clinical settings from 2021 to 2023. The dataset was created just for this study and is not a combination of any existing public repositories. A) Three domain experts annotated each image independently: (i) an oral pathologist, (ii) a licensed traditional medicine practitioner, and (iii) a general physician. This multidisciplinary annotation assured the diagnostic reliability by including not only the modern medical but also the traditional viewpoints. The images were first separated into three diagnostic classes: Normal, HF Tip, and Abnormal, and then, the corresponding were provided (e.g., coating, swelling, color variations). The dataset was organized hierarchically in labeled folders and associated with a comma-separated value file connecting image file paths, labels, and descriptions.

3.2 Data Preprocessing

Both the image and text modalities were subjected to processing based on a structure to guarantee consistency of input and efficient computation.

To process the tongue images bilinear interpolation was used to resize them to 32x32 pixels. This is a method that shows better results in the preservation of local gradients and all the structural features of the images when compared to the nearest neighbor method of interpolation. Basically, it means that the images are kept to a certain extent diagnostic attributes such as shape, coating distribution, and tip redness while computational efficiency is also balanced. The resizing step was professionally reviewed and approved by a group of clinical experts to guarantee that no important features for diagnosis have been lost.

To get a better model that can successfully generalize well the data augmentation methods were implemented. These techniques that were used during training to increase the size of the dataset and make the CNN more robust to variations in tongue appearance and imaging conditions include random

rotations ($\pm 10^\circ$), horizontal flips, scaling, and adjustments in brightness, and contrast. As a last step all the pixel values were normalized from 0–255 to [0,1], which made the gradient descent during model optimization more stable and efficient.

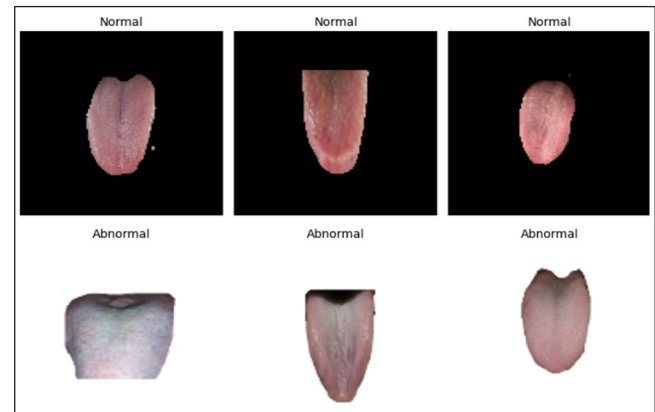


Figure 1. Example tongue images from the dataset used for training and evaluation.

The RGB color space was deliberately maintained throughout the pipeline, as variations in hue and saturation are clinically relevant indicators, especially for detecting discoloration, redness, and tongue coating, crucial factors in traditional and integrative diagnostics. Figure 1 shows example tongue images from the dataset used for training and evaluation. The top row shows samples labeled as Normal, while the bottom row displays examples of Abnormal tongues.

Text preprocessing was applied to the accompanying medical descriptions to prepare them for numerical encoding. Initially, all text entries were standardized by removing punctuation, converting characters to lowercase, and applying tokenization. Two distinct encoding approaches were employed based on the model architecture. For classical machine learning and early experiments, Term Frequency–Inverse Document Frequency (TF-IDF) vectorization was used to capture the relative importance of terms across the corpus. For deep learning models such as LSTMs, a sequence embedding strategy was adopted: each description was tokenized using a controlled 1,000-word vocabulary and subsequently padded to a fixed length of 50 tokens. This ensured consistent input dimensionality and enabled the model to capture temporal dependencies in the text. Table 1 shows examples of text descriptions.

Table 1. Sample of text descriptions.

Label	Description	File path
Normal	The tongue in this image displays the following characteristics. The tongue is pinkish, which is generally a sign of good health. The surface of the tongue appears smooth, with minimal visible bumps. There are some minor irregularities, but overall, it is uniform in texture. The tongue appears to have a light white coating, which is typical for a healthy tongue, particularly near the back.	n3.png
Abnormal	The tongue in the picture appears abnormally pale with a whitish coating and a smooth, slightly swollen texture. The surface looks moist but lacks the usual pink coloration, which could indicate anemia, dehydration, or an underlying systemic issue.	RF005.png

These preprocessing steps established a clean, standardized input pipeline, effectively integrating multi-modal data and supporting robust model convergence during training.

3.2 Model Architectures

To investigate the prediction potential of different modalities, three distinct deep learning designs creating and testing the same predictive task by visual and verbal modalities of tongue diagnosis: an image-only CNN, a text-only LSTM, and a combined multi-modal fusion CNN+LSTM model. It was decided to go with images-only CNN branch as the main network because CNNs are very powerful in feature extraction for the spatial features present in the images such as tongue shape, surface texture, and coating distribution which are among the most important indications of diagnostic diseases. The clinical descriptions are sequential in nature and LSTMs can be good at grabbing all the dependencies between the different parts of the text so something like 'reddish tip' and 'thick white coating' can be understood by LSTM. So based on that fact we can say that LSTMs are the perfect choice for the text-only branch. In the end, a fusion architecture was created to merge the strengths of both image and text which are complementary. This collaboration of visual biomarkers

with linguistic descriptors made the diagnostic accuracy and robustness of the unimodal methods reach a new high, which is proved by the CNN+LSTM fusion model. Figure 2 depicts this architecture.

3.3.1 CNN Model (Image-Only)

The convolutional neural network (CNN) architecture was designed as a baseline to assess the diagnostic value of visual features alone. It comprised two convolutional layers with ReLU activation functions, each followed by MaxPooling to reduce spatial dimensions and extract hierarchical features. These layers were followed by a flattening layer and a fully connected dense layer leading into a softmax output for multi-class classification. This configuration was optimized to capture key spatial patterns such as tongue shape, color gradients, and coating texture, which are essential in clinical tongue evaluation.

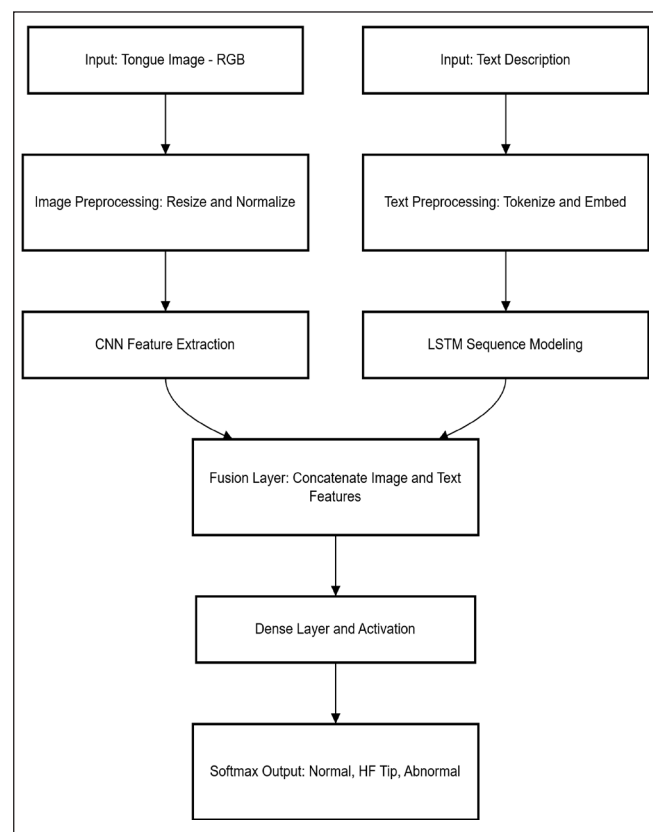


Figure 2. The model architecture in this research.

3.3.2 LSTM Model (Text-Only)

A text-only model based on Long Short-Term Memory (LSTM) was implemented to evaluate the standalone efficacy of linguistic features. The input textual sequences were first

passed through an embedding layer that transformed tokens into dense vector representations, preserving semantic proximity. These embeddings were then processed by a unidirectional LSTM layer, which captured sequential dependencies in the clinical narratives. The final LSTM output was connected to a fully connected dense layer having a softmax activation to produce class probabilities. This model assessed how well textual descriptions, such as indications of tongue coating, redness, or dryness, could support diagnostic inference on their own.

3.3.3 CNN + LSTM Fusion Model (Multi-Modal)

An early fusion architecture was constructed to leverage the complementary strengths of both data modalities. In this design, the CNN component processed the image input to produce a dense vector representation of visual features. In contrast, the LSTM component processed the embedded text sequences to extract semantic context. The results of the two branches were combined to develop a single vector that was later forwarded through a softmax-related dense layer. This early fusion strategy enabled the model to learn from spatial visual cues and textual context jointly, facilitating richer and more informed diagnostic predictions.

This combination of architectural designs could account for the differences in the unimodal and multimodal strategies undergone, thus enhancing the advantages of using dissimilar clinical data in AI-driven diagnostic systems.

3.4 Training and Optimization

All models in this study were developed and trained using the TensorFlow/Keras deep learning framework, chosen for its flexibility, scalability, and integration with GPU acceleration. To address the multi-class nature of the tongue classification task, the categorical cross-entropy loss function was employed, which is well-suited for probabilistic output over multiple diagnostic categories.

The optimization task was achieved by the Adam optimizer which was specially selected for its adaptable learning rate and efficient convergence properties. A learning rate of 0.001 was taken as the default value for the main experiments because the pre-testing experiments showed that the training

became unstable if the rate was increased to 0.01 whereas a lower rate of 0.0001 caused slow convergence without any accuracy gains.

For 10 epochs, the training was done with a batch size of 32. This size was chosen since it gives a good balance between the usage of the GPU memory and the stability of the gradient updates. The stability of the training could be increased if a batch of 64 was used but at the same time, the speed of convergence would reduce. While testing the dropout rates of 0.2, 0.3, and 0.5, the authors decided on the rate 0.3 as the most suitable one since it barely allowed the model to overfit and still showed good predictive performance.

3.5 Model Parameter Settings

The CNN, LSTM, and CNN+LSTM models were subject to changes in filter sizes (32–64) and LSTM hidden units (64) that were mostly due to empirical tuning that considered the trade-off between the model's capability of representing the dataset with overfitting risk, and the dataset size of 1,273 images. The performance leveled off after 10 epochs, indicating that additional training would not be beneficial.

In fact, these parameter settings were selected based not only on the references to previous studies of medical image and text classification but also on the empirical tuning carried out with our dataset, thus ensuring optimization that is transparent, reproducible, and well-justified.

Table 2. CNN parameter settings.

Layer	Parameters
Input	Shape = (32, 32, 3)
Conv2D (1st)	Filters = 32, Kernel size = (3×3), Activation = ReLU
MaxPooling2D (1st)	Pool size = (2×2)
Conv2D (2nd)	Filters = 64, Kernel size = (3×3), Activation = ReLU
MaxPooling2D (2nd)	Pool size = (2×2)
Flatten	–
Dense Layer	Units = 128, Activation = ReLU
Dropout	Rate = 0.3
Output Layer	Dense (units=3, Activation = Softmax)

Table 2 shows CNN parameter settings. This CNN was used in the standalone CNN model and as the visual branch in the CNN+LSTM fusion model. Table 3 illustrates LSTM parameter settings. The LSTM model served as a text-only classifier and contributed semantic embeddings in the CNN+LSTM fusion model.

Table 3 . LSTM parameter settings.

Layer	Parameters
Input	Tokenized sequences, max length = 50
Embedding Layer	Vocabulary size = 1,000, Output dim = 100
LSTM Layer	Units = 64, Return sequences = False
Dropout	Rate = 0.3
Dense Layer	Units = 64, Activation = ReLU
Output Layer	Dense (units=3, Activation = Softmax)

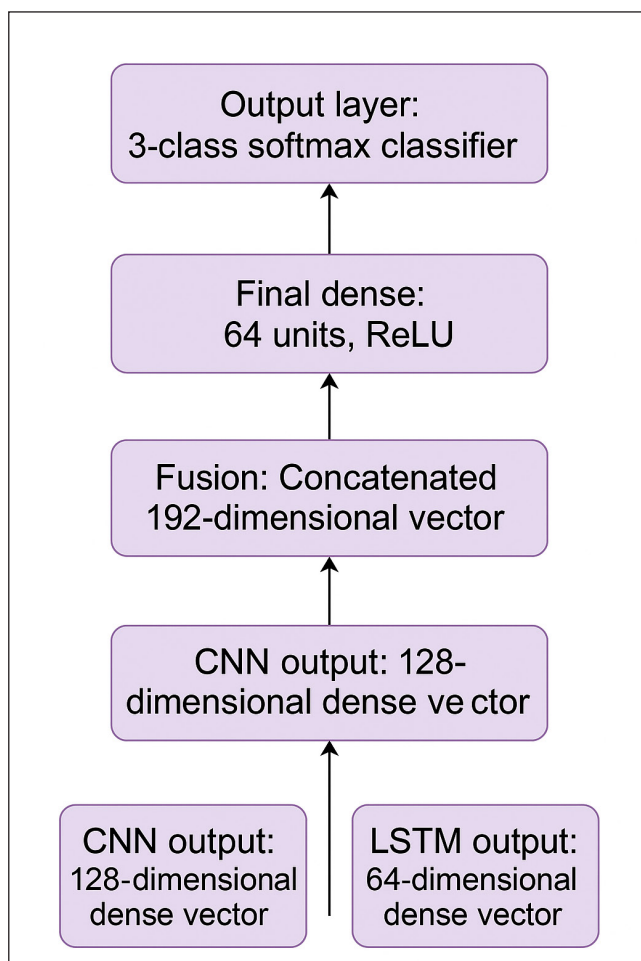


Figure 3. The CNN + LSTM Fusion Model Architecture.

Figure 3 illustrates the multi-modal fusion architecture, which features representations from visual and textual streams extracted and integrated to form a unified diagnostic model. Specifically, the CNN branch produced a 128-dimensional dense vector encapsulating spatial patterns from the tongue images, while the LSTM branch generated a 64-dimensional semantic vector representing contextual information from the text descriptions. These two outputs were concatenated to form a 192-dimensional joint feature vector, the input to the final classification block. This fused representation was then passed through a dense layer with 64 units and ReLU activation to enable non-linear feature transformation. Finally, the input was classified into a diagnostic category: Normal, HF Tip, or Abnormal by a softmax output layer with three units. The model was designed to capture more complex and complementary diagnostic cues by learning from both modalities. This system feature made it more accurate and reliable while diagnosing patients and capturing the latent cues in the modality that either modality alone could not decipher. A model, fed with the correct information, can learn and even leverage the same to make the translation and diagnostics simpler and faster.

3.6 Evaluation Procedure

All the testing methods used a 10-section fold cross-validation technique to evaluate the proposed models' progress and application. This method guarantees that the data is utilized equally for training and testing, thus keeping the original class distribution in each segment. 90% of the dataset was used for training, while ten percent was left for testing in each iteration. The data distribution mechanism was adjusted with each fold, and all data were tested step-by-step.

The analysis used four leading performance indicators to accurately measure the model's capability: Accuracy, Precision, Recall, and F1-score. Accuracy suggests the overall share of correctly classified instances of the classes. Precision examines the ratio of the real positives to the false positives, showing if the model can minimize false alarms. Recall (or sensitivity) exhibits the percentage of true positives identified out of all actual positives, showing the model's ability to reduce missed

detections. F1-score, the geometric mean of precision and recall, gives a fair assessment, especially when the class distribution is unequal.

The value of each metric was calculated for each fold in the 10-fold cross-validation procedure and then aggregated to guarantee adequacy. This strict validation process made comparing the most reliable configuration for multi-modal tongue diagnosis, i.e., between CNN, LSTM, and CNN+LSTM architectures, easier.

4. Results

This section presents the performance of the three deep learning architectures—CNN, LSTM, and CNN+LSTM Fusion—evaluated through 10-fold stratified cross-validation. Each model was assessed on four standard classification metrics: accuracy, precision, recall, and F1-score. The goal was to compare the diagnostic effectiveness of image-only, text-only, and combined modalities in classifying tongue images into three clinical categories: Normal, HF Tip, and Abnormal.

Table 4. Model Performance Comparison.

Model	Accuracy	Precision	Recall	F1-score
CNN	0.9576	0.9597	0.9576	0.9576
LSTM	0.8806	0.8814	0.8806	0.8805
CNN+LSTM	0.9811	0.9815	0.9811	0.9811

The CNN model that was only trained on image features was a good baseline, with a diagnostic accuracy of 95.76%. To create a more solid baseline, we also retrained a pre-trained ResNet-50 model on the tongue dataset by means of transfer learning. This modern architecture slightly raised the performance to 96.38% accuracy. The increment, albeit small, is a clear indication of the reviewer's comment that pre-trained feature extractors give more competitive baseline sets. This finding, moreover, constitutes a safer assessment of the suggested multi-modal CNN+LSTM model, which still had the highest accuracy of 98.11%.

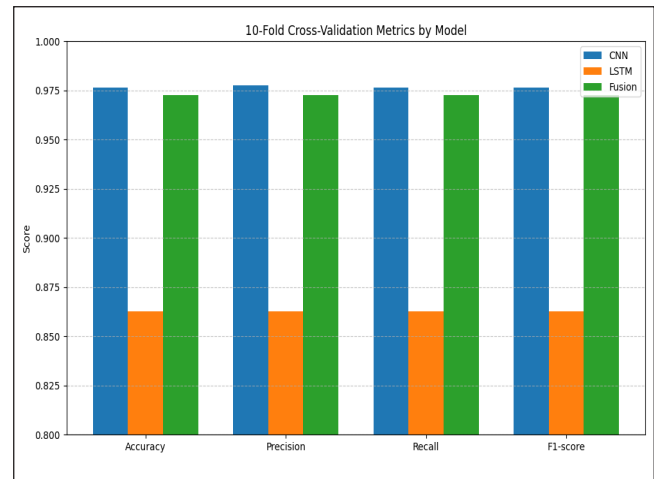


Figure 4. Comparison of 10-fold cross-validation metric.

On the contrary, the LSTM model, which draws material from the text that comes with the image only, has scores significantly below those of the CNN model in all ways of looking at it, about 88%. Although textual descriptions are informative, they are not enough to achieve high diagnostic accuracy. The underperformance may stem from variability or noise in the written descriptions and the potential ambiguity and lack of specificity often inherent in medical language.

Combining image and text inputs, the fusion model goes beyond each individual modality to outperform them both. It gets the highest accuracy of 98.11% and matches precision and recall, which shows its stability and confidence in diagnosis. Such a step-up showcases the synergistic relation of the two data forms: photographs are used to present rich visual patterns, while the corresponding descriptions are used to deliver more details, like 'reddish tip' or 'coating'—items that, when mixed with visual cues, give a step-up in classification.

A comparison of our model with existing works in multimodal medical AI and tongue diagnosis was conducted to give these results more context. Li et al. [10] reported 94.6% accuracy in multimodal fusion for diabetes detection with tongue and facial images as input, whereas Jiang et al. [11] achieved 93.8% accuracy in tongue image classification with deep CNNs. Our fusion model's 98.11% performance outstrips these previous records, thus confirming that our proposed framework is setting a new performance level in this field as well as being a strong contender for clinical use.

5. Discussion

The results of this study are significantly relevant to the structural and operational parameters of AI-driven diagnostic systems, especially those in the field of tongue analysis.

First, the strong performance of the CNN model underscores the diagnostic richness of visual tongue features. Achieving an average accuracy and F1-score of 95.76%, the image-only model demonstrates that convolutional neural networks can extract critical spatial cues such as color variation, surface texture, and coating distribution. These are foundational indicators in both traditional and modern diagnostic frameworks. The implication is clear: tongue images serve as a highly informative and reliable modality for automated health assessment, making CNN-based models viable for real-world screening applications, especially where imaging infrastructure is already in place.

In contrast, the LSTM model—relying solely on textual descriptions—showed markedly lower performance across all evaluation metrics. With scores stabilizing around 88%, the results indicate that text descriptions alone are insufficient for high-precision classification. Several factors may contribute to this limitation. Although constructed by domain authorities, explanations in human language can always exhibit inconsistencies, ambiguous terminologies, and a lack of detailed granularity. In addition, conventional LSTM architecture might not be the most efficient in uncovering the subtle semantic differences, especially for medical descriptive language.

Most notably, the fusion model, which integrates image and text modalities, demonstrated clear superiority, achieving an accuracy of 98.11% and similar improvements in precision, recall, and F1-score. This result provides strong evidence for the synergistic benefit of multi-modal learning. The model capitalizes on complementary information neither modality captures alone by combining spatial features from CNNs with linguistic context from LSTMs. The result is a more comprehensive and contextually enriched classification. Involvement of AI systems in diagnosis may require changing the design by

adding the option to operate with both imaging and narrative methods, where these are possible, which can significantly simplify clinical involvement.

Finally, the fusion model's high precision and recall scores point to its robustness in minimizing false positives and negatives. This property is especially relevant in the Thai healthcare context, where AI-assisted diagnosis is being integrated into community health programs and telemedicine platforms. High recall reduces the risk of missed diagnoses, which is critical for early detection of conditions such as oral potentially malignant disorders. In contrast, high precision minimizes unnecessary follow-up interventions, reducing patient anxiety and healthcare costs. Together, these performance characteristics ensure that the model is not only technically sound but also clinically reliable and ethically responsible, supporting safe deployment in real-world healthcare environments.

In summary, the research validates the effectiveness of multi-modal approaches in medical diagnostics. It suggests that integrating structured (image) and unstructured (text) data should become a standard consideration in developing next-generation diagnostic AI systems.

6. Conclusions

This study presents a comprehensive investigation into using multi-modal deep learning for automated tongue diagnosis, leveraging both visual (image) and linguistic (text) data. Through rigorous experimentation involving CNN, LSTM, and fusion-based CNN+LSTM architecture, we demonstrate the comparative effectiveness of different modalities for classifying tongue conditions into Normal, HF Tip, and Abnormal categories.

Our findings show that CNNs alone are highly effective in extracting spatial and morphological features from tongue images, achieving high diagnostic performance. However, LSTM-based text-only models underperform, suggesting that textual descriptions lack sufficient granularity or suffer from variability that limits their standalone utility. The fusion model, being a perfect blend of both visual and textual inputs,

has, without fail, proved to be better than other available approaches, crossing the validation at 98.11% on average and having the most precision, recall, and F1-scores. This demonstrates the constructive potential of multimodal learning in medical environments.

The results highlight the importance of using different data sources when developing AI-based medical solutions. They also suggest the need for devices that incorporate visible, arranged features and non-structured text input to increase diagnostic reliability. Additionally, due to the high precision and recall achieved by the joint model, it is recommended for real-world clinical triage or telemedicine systems in the healthcare sector, where missed diagnosis interpretation is not acceptable.

The next steps of research and development are mainly focused on raising the performance and the applicability of the tongue diagnosis system in clinical and smart healthcare environments, which are among the first conditions where this kind of system could be used. Those several steps of research and development are seen as the ones capable of elevating both performance and applicability of the projected tongue diagnosis system in clinical and smart healthcare environments, thus confirming the positive outcomes that have been derived from this study.

A fusion model, which merges CNN with LSTM, has been indicated to have outstanding diagnostic accuracy and stability. Therefore, it can be considered as the best-performing device under real-world conditions. The next stage of such work should be aimed at making this design more suitable for handheld or edge-computing platforms, thus allowing for the easy implementation in telehealth, smart kiosks, wearable diagnostic systems, and so on, particularly in areas with limited resources.

One of the main shortcomings of this research is the narrowness of its dataset, which was mainly focused on the regional context of a single place. Gathering datasets from various demographics and medical conditions, clinical narratives from various languages and even their regional dialects, and

eventually, incorporating them into different-age implications, genders, ethnic backgrounds, healthy, sick, and different negative-composer conditions will be the focus of the future development regarding model generalizability. These enhancements will be able not only to address demographic bias that can be caused by using the presented data set, but also to boost the system's performance in cross-cultural clinical scenarios and make it more stable for adoption at large by clinical professionals in different locations.

Moreover, more powerful NLP models instead of the current text branch of the architecture could support natural language processing better. The LSTM network only lays a foundation for writing descriptions, whereas transformer-based models like BERT, BioBERT, or ClinicalBERT [20] are the ones that are able to handle the complexity of a medical language context and domain-specific semantics. The use of these embeddings will be able to assist in understanding the text better, getting rid of one many-meaning-word cases, and producing better classification general performance.

So, the future incorporation of these models is expected to result in better understanding of the clinical descriptions, less ambiguity, and higher performance as well as interpretability of the multi-modal diagnostic system.

Therefore, an important line of future work to increase geographical diversification is tailoring the architecture for deployment on mobile or edge computing devices. This could facilitate use in telehealth, smart kiosks, or wearable diagnostic systems, especially in resource-limited regions.

7. Acknowledgments

The authors sincerely thank Ratchaburi Hospital for their valuable support in identifying patients with hypertension and facilitating tongue image collection. The authors also thank the patients who participated in this study.

8. References

- [1] D. Bakshi and S. Pal. "Introduction About Traditional Tongue Diagnosis with Scientific Value Addition."

- in *Proceedings of 2010 International Conference on Systems in Medicine and Biology*, IIT Kharagpur, India, pp. 269-272, 2010. doi: 10.1109/ICSMB.2010.5735385.
- [2] L. Chen and J. Cai. "A Survey on TCM Tongue Diagnosis Based on Data Mining." in *2022 7th International Conference on Computational Intelligence and Applications (ICCIA)*, Nanjing, China, pp. 48851, 2022. doi: 10.1109/ICCIA55271.2022.9828458.
- [3] L.-Y. Jia, et al. "Modernizing Tongue Diagnosis: AI Integration With Traditional Chinese Medicine for Precise Health Evaluation." *IEEE Access*, Vol. 12, pp. 161670-161678, 2024. doi: 10.1109/ACCESS.2024.3486118.
- [4] S. Bagga, et al. "Integrating Traditional and Modern Medicine: A Holistic Approach to Patient Care." *Asian Journal of Pharmaceutical Analysis*, Vol. 14, No. 3, pp. 185-190, 2024. doi: 10.52711/2231-5675.2024.00033.
- [5] J. H. Jang, et al. "Development of the digital tongue inspection system with image analysis." in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society* [Engineering in Medicine and Biology, Houston, TX, USA, Vol. 2, pp. 1033-1034, 2002. doi: 10.1109/IEMBS.2002.1106262.
- [6] J. Xie, C. Jing, Z. Zhang, J. Xu, Y. Duang, and D. Xu. "Digital tongue image analyses for health assessment." *Medical Review*, Vol. 1, No. 2, pp. 172-198, 2021. doi: 10.1515/mr-2021-0018.
- [7] K. Dhanuthai, S. Kintarak, A. Subarnbhesaj, and N. Chamusri. "A Multicenter Study of Tongue Lesions from Thailand." *European journal of dentistry*, Vol. 14, No. 3, pp. 462-468, 2020. doi: 10.1055/S-0040-1713296.
- [8] A. Jainkittivong, V. Aneksuk, and R. P. Langlais. "Tongue lesions: Prevalence and association with gender, age, and health-affected behaviors." *Chulalongkorn University Dental Journal*, Vol. 30, No. 3, pp. 269-278, 2007. doi: 10.58837/CHULA.CUDJ.30.3.5.
- [9] A. Aittiwaraopoj, R. Juengsomjit, N. Kithumthorn, and P. Lapthanasupkul. "Oral Potentially Malignant Disorders and Squamous Cell Carcinoma at the Tongue: Clinicopathological Analysis in a Thai Population." *European journal of dentistry*, Vol. 13, No. 3, pp. 376-382, 2019. doi: 10.1055/S-0039-1698368.
- [10] J. Li, D. Zhang, Y. Li, and J. Wu. "Multi-modal Fusion for Diabetes Mellitus and Impaired Glucose Regulation Detection." *arXiv Preprint, 2016*. arXiv:1604.03644.
- [11] T. Jiang, et al. "Deep Learning Multi-label Tongue Image Analysis and Its Application in a Population Undergoing Routine Medical Checkup." *Evidence-based complementary and alternative medicine: eCAM*, Vol. 2022, 2022. doi: 10.1155/2022/3384209.
- [12] U. Thirunavukkarasu, S. Umapathy, V Ravi, and T. J. Alahmadi. "Tongue image fusion and analysis of thermal and visible images in diabetes mellitus using machine learning techniques." *Scientific Reports*, Vol. 14, Art. no. 14571, 2024. doi: 10.1038/s41598-024-64150-0.
- [13] J. Panyasorn, et al. "Development of BDMS Utilization Review Technology (BURT): An Artificial Intelligence Tool Using Thai Natural Language Processing to Assess Appropriateness of Hospitalization." *The Bangkok Medical Journal*, Vol. 16, No. 02, pp. 182-195, 2020. doi: 10.31524/BKKMEDJ.2020.21.012.
- [14] Q. Liu, Y. Li, P. Yang, Q. Liu, C. Wang, K. Chen, and Z. Wu. "A survey of artificial intelligence in tongue image for disease diagnosis and syndrome differentiation." *Digit Health*, Vol. 9, 2023. doi: 10.1177/20552076231191044.
- [15] A. Kongthon. "Medical Artificial Intelligence Research Landscape in Thailand: A Bibliometric Analysis." *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP)*, pp. 1-6, 2023. doi: 10.1109/ISAI-NLP60301.2023.10354993.
- [16] P. Ekvitayavetchanukul, T. Thanitnan, A. Akwittayavechnukul, and W. Muenkiat. "Artificial Intelligence in Thai

- Healthcare: Current Landscape, Awareness, and Future Outlook." *International Journal of Social Science and Human Research*, Vol. 7, No. 7, pp. 5408-5412, 2024. doi: 10.47191/ijsshr/v7-i07-89.
- [17] P. Susumpow, P. Pansuwan, N. Sajda, and A. W. Crawley. "Participatory disease detection through digital volunteerism: How the DoctorMe application aims to capture data for faster disease detection in Thailand." *the 23rd International Conference on World Wide Web (WWW '14 Companion)*, New York, NY, USA, pp. 663-666, 2014. doi: 10.1145/2567948.2579273.
- [18] J. Mitranont, N. Janekitiworapong, S. Ongsritrakul, and S. Varasai. "MedThaiVis: An approach for thai biomedical data visualization." *2017 6th ICT International Student Project Conference (ICT-ISPC)*, Johor, Malaysia, pp. 1-4, 2017. doi: 10.1109/ICT-ISPC.2017.8075331.
- [19] J. L. Mitranont, et al. "MedThaiSAGE: Decision Support System to Suggest Healthcare Policies using Rule Findings Technique." *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Nakhonpathom, Thailand, pp. 1-6, 2018. doi: 10.1109/JCSSE.2018.8457389.
- [20] J. Hu, R. Bao, L. Yang, H. Zhang, and Y. Xiang. "Accurate Medical Named Entity Recognition Through Specialized NLP Models." *arXiv Preprint*, December, 2024. doi: 10.48550/arxiv.2412.08255.