Tongue-Based Heart Disease Classification: Data Collection, Preprocessing, and Deep Learning Implementation

By: Katarina Barbosa

1. Introduction and Research Motivation

Traditional medicine systems including Traditional Chinese Medicine and Ayurveda have long used tongue diagnosis as a window into systemic health, including cardiovascular function. This project explores whether modern computer vision and deep learning can validate and quantify these ancient diagnostic approaches by developing an automated system for cardiovascular disease detection from tongue photographs. Unlike our parallel research using periorbital eye images with hand-crafted features achieving 73.2 percent AUC, this tongue-based approach employs end-to-end deep learning to automatically discover discriminative patterns directly from raw image data.

The biological rationale underlying tongue-based cardiovascular assessment builds upon observations that tongue color, coating, moisture, texture, and shape may reflect systemic health status. The tongue's rich vascular supply and mucosal surface potentially manifest cardiovascular dysfunction through changes in perfusion, oxygenation, and microcirculation. Traditional practitioners associate pale tongues with qi deficiency potentially corresponding to reduced cardiac output, purplish discoloration with blood stasis possibly indicating poor circulation, and coating characteristics with metabolic imbalances that may accompany cardiovascular disease. This research aims to determine whether these qualitative observations translate into quantifiable patterns detectable through computational analysis.

2. Data Collection Challenges and Initial Dataset

The initial data collection presented unique challenges distinct from controlled medical imaging. Rather than standardized clinical photographs captured under consistent conditions, our dataset originated from full facial images showing individuals extending their tongues. This real-world data reflected the practical constraints of large-scale health screening where standardized clinical photography may prove impractical or unavailable. The images varied considerably in lighting conditions, camera angles, tongue extension degree, focal distance, and background environments, introducing substantial variability that both complicated analysis and tested the robustness of our computational approach.

The dataset comprised images from two groups collected with appropriate consent and ethical approval. The heart disease cohort included individuals with documented cardiovascular conditions confirmed through clinical diagnosis, while the control group consisted of healthy subjects without known cardiovascular disease based on medical history and examination. All images captured full faces with subjects instructed to extend their tongues maximally to expose the dorsal surface for analysis. This protocol aimed to reveal tongue body color, coating distribution, moisture level, and surface characteristics traditionally associated with diagnostic assessment.

3. Manual Quality Control and Image Curation

The non-standardized nature of the source images necessitated extensive manual quality control before automated analysis could proceed. This curation process represented a critical preprocessing step often overlooked in machine learning pipelines but essential for reliable model development when working with real-world medical data. Each image underwent individual visual inspection to identify and exclude problematic cases that would introduce noise or bias into the training process. This manual review proved time-intensive but irreplaceable for ensuring data quality given the variability in source material.

The primary exclusion criterion targeted images with inadequate or inconsistent lighting that obscured tongue characteristics or created misleading color representations. Overexposed images washed out subtle color variations diagnostic of cardiovascular status, while underexposed images rendered the tongue surface too dark for feature extraction. Directional lighting creating strong shadows across the tongue surface was excluded because shadows could be misinterpreted as coating or discoloration. Only images with relatively uniform, diffuse lighting sufficient to reveal surface characteristics across the entire visible tongue were retained for analysis.

A second critical exclusion criterion addressed tongue orientation. Traditional tongue diagnosis examines the dorsal surface with attention to color, coating, shape, and regional characteristics. However, some subjects inadvertently presented the ventral surface or lateral edges rather than the appropriate dorsal view. Images showing primarily the bottom of the tongue, lateral edges, or insufficient extension to reveal the posterior tongue were manually identified and removed. Only images clearly displaying the dorsal surface with adequate exposure for diagnostic assessment were included in the final dataset. This manual filtering ensured all analyzed images presented comparable anatomical views necessary for meaningful pattern recognition.

Additional quality control addressed technical issues including severe motion blur rendering surface details indistinguishable, extreme close-ups or distant shots preventing consistent scale comparison, and obstructions such as piercings or oral lesions that might confound automated analysis. Images where the tongue was partially obscured by lips, teeth, or other facial structures were similarly excluded. The manual curation process ultimately retained 404 images from healthy controls and 426 images from individuals with heart disease, establishing a reasonably balanced dataset of 830 total images suitable for supervised learning.

This extensive manual preprocessing underscores an important reality of medical machine learning: automated algorithms require clean, consistent input data to perform effectively. While the manual effort proved labor-intensive, it established a curated dataset reflecting the actual diagnostic view a clinician would assess. Future work might explore automated quality control using image quality metrics, orientation detection, and lighting assessment to scale this preprocessing step, but the current manual approach ensured high-quality training data for initial model development.

4. Image Standardization and Preprocessing Pipeline

Following manual quality control, all retained images underwent standardized preprocessing to normalize dimensions and prepare inputs for the neural network. Each image was resized to 1024 by 1024 pixel resolution using Lanczos resampling, a high-quality interpolation method that preserves fine details while changing image dimensions. This relatively high resolution was selected deliberately to maintain subtle textural and chromatic variations in tongue surface characteristics that might prove diagnostically relevant. Unlike many computer vision tasks where lower resolutions of 224 by 224 or 256 by 256 pixels suffice, tongue diagnosis potentially requires preservation of fine-grained features such as coating distribution, papillae patterns, and localized color variations that would be lost at lower resolutions.

The Lanczos resampling algorithm employs a windowed sinc function to compute interpolated pixel values, providing superior quality compared to simpler methods like bilinear or bicubic interpolation. This high-quality resampling minimized artifacts that could confound pattern recognition while standardizing all images to identical dimensions required for batch processing in the neural network. The 1024 by 1024 target resolution represented a deliberate balance: high enough to preserve diagnostic details, but manageable from a computational perspective for model training and inference.

The resized images were organized into separate directories for cases and controls, establishing clear ground truth labels for supervised learning. The heart disease cohort occupied one directory with 426 images, while the healthy control group resided in a second directory with 404 images. This organizational structure facilitated straightforward data loading with automatic label assignment based on directory membership. The slight imbalance between groups (51 percent cases versus 49 percent controls) was addressed during dataset preparation by limiting both classes to 400 images through random sampling, creating a balanced training corpus less susceptible to class imbalance bias.

5. Dataset Partitioning and Augmentation Strategy

The curated and balanced dataset of 800 images underwent stratified partitioning into training and test sets using an 80-20 split ratio commonly employed in machine learning research. Stratification ensured both subsets maintained the 50-50 balance between cases and controls, preventing one class from dominating either partition. This yielded 640 training images (320 per class) for model fitting and 160 test images (80 per class) for unbiased performance evaluation. The test set remained completely withheld during all training and development activities, serving as a held-out evaluation dataset to estimate real-world generalization performance.

Before training, images underwent additional preprocessing including resizing from the standardized 1024 by 1024 dimensions down to 224 by 224 pixels to reduce computational requirements while maintaining sufficient resolution for feature learning. This dimension reduction balanced detail preservation against training efficiency, with the network architecture designed to operate effectively at this more modest resolution. Each image was converted from unsigned integer pixel values to floating-point tensors and normalized using ImageNet statistics with mean values of 0.485, 0.456, 0.406 and standard deviations of 0.229, 0.224, 0.225 for red, green, and blue channels respectively.

Although these normalization constants derived from the ImageNet natural image database rather than tongue photographs specifically, this approach facilitated potential transfer learning in future work while providing reasonable standardization for current training. Using pre-computed normalization constants simplified the data pipeline and aligned with standard practices in computer vision research. The normalized images were organized into mini-batches of 32 examples during training, striking a balance between gradient estimate quality and GPU memory constraints.

6. Convolutional Neural Network Architecture

The classification system employed a custom convolutional neural network designed to balance learning capacity against overfitting risk given the modest dataset size of 640 training images. The architecture comprised three convolutional blocks, each containing a convolutional layer with 3 by 3 kernels, rectified linear unit activation introducing nonlinearity, and max pooling with 2 by 2 windows for spatial downsampling. This sequential application of convolution, activation, and pooling represents a fundamental building block of successful image recognition architectures, progressively extracting higher-level feature representations through multiple processing stages.

The first convolutional layer contained 16 filters learning to detect low-level visual primitives such as edges, color gradients, and simple textures across the input tongue images. The same padding preserved spatial dimensions allowing each filter to analyze all image locations. The subsequent max pooling reduced dimensions by half to 112 by 112 pixels while retaining the most prominent features, providing translation invariance and computational efficiency. The second convolutional layer doubled capacity to 32 filters, learning combinations of first-layer features to represent more complex patterns such as coating textures, color regions, and papillae configurations.

The third convolutional layer further increased to 64 filters capturing high-level semantic features potentially corresponding to diagnostic tongue characteristics. After three convolution and pooling rounds, the spatial dimensions reduced from 224 by 224 input to 28 by 28 feature maps, with 64 feature channels. This three-dimensional activation volume was flattened into a vector of 50,176 activations feeding into a fully connected layer with 128 hidden units. This bottleneck compressed the feature representation into a compact embedding capturing the most discriminative information for classification.

To combat overfitting on the relatively small training set, a dropout layer with 50 percent drop probability was inserted before the final classification layer. During training, dropout randomly zeroed half the activations at each iteration, forcing the network to learn redundant representations and preventing feature co-adaptation. This regularization technique proved particularly important given only 640 training examples. The final fully connected layer with two output units and softmax activation produced probability estimates for control and heart disease classes, with the higher probability determining predicted classification.

7. Training Procedure and Optimization

Model training employed the Adam optimization algorithm with learning rate 0.001, combining adaptive learning rates and momentum for accelerated convergence. Cross-entropy loss served as the training objective, quantifying discrepancy between predicted probability distributions and true labels. This loss function naturally handles probabilistic softmax outputs while providing strong gradients for effective learning. Training proceeded for 10 epochs, with each epoch comprising complete passes through all 640 training images organized into 20 batches of 32 images.

The learning trajectory revealed characteristic dynamics with progressive improvement followed by apparent convergence. Initial performance at epoch one showed training loss 0.7660 and accuracy 56.41 percent, marginally above the 50 percent chance baseline for balanced binary classification. This baseline reflected random weight initialization and inherent task difficulty. Subsequent epochs demonstrated steady improvement, with loss decreasing to 0.6531 and accuracy increasing to 62.50 percent by epoch two, indicating the network began discovering meaningful patterns in the training data.

Training continued through epochs three through six with gradual improvements punctuated by occasional fluctuations characteristic of stochastic gradient descent. By epoch six, training accuracy reached 64.06 percent with loss 0.6281. Epochs seven through nine showed more substantial gains, with epoch nine achieving 71.72 percent training accuracy and loss 0.5670. This performance suggested the network discovered genuinely discriminative patterns, though the gap between training and ultimate test performance indicated overfitting. The final epoch reached 68.44 percent training accuracy with loss 0.5486, a modest decrease potentially reflecting overfitting onset or stochastic variation.

8. Test Performance and Evaluation

Evaluation on the held-out test set of 160 images revealed modest classification performance with accuracy 53.75 percent, only marginally exceeding chance. This indicated the model correctly classified approximately 86 of 160 test images, barely distinguishing between cases and controls. The confusion matrix detailed performance with 47 true negatives, 33 false positives, 41 false negatives, and 39 true positives. The model correctly identified 47 of 80 controls and 39 of 80 heart disease patients, showing similar performance across both classes without strong bias.

Precision measured 54.17 percent, meaning among subjects classified as having heart disease, only 54 percent actually had the condition while 46 percent were false positives. This modest precision indicated substantial false alarms requiring confirmatory testing in clinical deployment. Recall of 48.75 percent demonstrated the model detected fewer than half of actual heart disease cases, missing 51.25 percent of affected individuals. This low recall posed significant concerns for screening applications where failure to identify diseased patients carries serious clinical consequences.

The F1-score of 51.32 percent represented the harmonic mean of precision and recall, providing a single metric balancing both measures. This below-average F1-score confirmed the model struggled on both dimensions. The area under the receiver operating characteristic curve measured 0.6055, exceeding the 0.5 random guessing threshold but falling well short of the 0.7 minimum typically considered adequate for clinical applications. This AUC indicated that for randomly selected case-control pairs, the model assigned higher disease probability to actual cases in only 60.55 percent of instances. The misclassification rate of 46.25 percent quantified that nearly half of predictions were erroneous.

9. Comparison with Eye-Based Classification

The tongue-based classification performance of AUC 0.6055 contrasted markedly with the parallel eye-based system achieving AUC 0.732 on a comparable dataset. This 12 percentage point performance gap suggested fundamental differences in either the discriminative information content of the two modalities or the effectiveness of the analytical approaches employed. The eye-based system used traditional machine learning with 67 hand-crafted features targeting known physiological mechanisms linking periorbital appearance to cardiovascular function, including red channel dominance, vessel tortuosity, circulation asymmetry, and regional blood flow patterns.

These expertly designed features directly encoded medical knowledge accumulated over decades of clinical observation, essentially embedding domain expertise into the feature representation. In contrast, the tongue-based approach relied on end-to-end deep learning to automatically discover discriminative patterns from raw images without explicit incorporation of medical knowledge about tongue characteristics associated with heart disease. While this automatic feature learning offers theoretical advantages by potentially capturing subtle patterns that escape manual engineering, it typically requires substantially larger training sets than the 640 images available here.

The performance difference may also reflect genuine biological differences in the diagnostic signal strength between the two modalities. The periorbital region contains rich vascular networks with established connections to cardiovascular function through microcirculation patterns, vessel organization, and blood flow symmetry. These manifestations prove relatively consistent and well-characterized in the medical literature. Tongue characteristics associated with cardiovascular disease may prove more subtle, variable across individuals, or less directly linked to cardiac function, requiring larger datasets or more sophisticated analytical approaches to capture reliably.

10. Multi-Modal Integration: Future Directions

The complementary nature of eye and tongue-based cardiovascular assessment suggests significant potential for multi-modal integration to improve diagnostic accuracy beyond either modality alone. Each imaging approach captures distinct aspects of systemic cardiovascular function through different physiological mechanisms. Periorbital images reveal microvascular circulation, vessel organization, and blood flow patterns through highly vascularized tissue surrounding the eyes. Tongue images potentially reflect tissue oxygenation, coating characteristics possibly related to circulation, and color variations that might indicate cardiovascular status through mechanisms recognized in traditional medicine.

A critical next step in this research program involves matching corresponding eye and tongue images from the same individuals to enable paired analysis. Currently, the eye and tongue datasets exist as separate collections with no linking information to identify which images originated from the same subjects. Establishing these correspondences would enable development of multi-modal classification systems that integrate information from both sources. Such integration could proceed through several architectural approaches including early fusion concatenating raw images, intermediate fusion combining learned feature representations, or late fusion ensembling separate predictions from each modality.

Early fusion would concatenate preprocessed eye and tongue images as multi-channel inputs to a unified deep network learning joint representations. This approach allows the model to discover complementary patterns and inter-modal relationships but requires careful normalization to prevent one modality from dominating. Intermediate fusion would process each modality through separate convolutional pathways before combining learned feature vectors for final classification. This architecture preserves modality-specific processing while enabling integration of complementary information at a semantic level.

Late fusion would train independent classifiers for each modality and combine their probabilistic predictions through weighted averaging, majority voting, or meta-learning. This ensemble approach proves robust to modality-specific failures and allows incorporation of the high-performing hand-crafted feature system for eye images alongside deep learning for tongue images. Optimal fusion weights could be learned on validation data to emphasize the more reliable modality. Research on other multi-modal medical imaging tasks suggests intermediate or late fusion often outperform early fusion by allowing specialized processing per modality before integration.

Beyond architectural considerations, multi-modal integration requires addressing practical challenges. The image correspondence problem demands careful record-keeping to ensure each subject contributes matched eye and tongue photographs captured under comparable conditions. Missing modalities for some subjects necessitate handling strategies such as single-modality prediction for incomplete cases or imputation of missing views. The relative contribution of each modality to final predictions should be interpretable to clinicians, suggesting attention mechanisms or learned fusion weights that reveal which source drove each classification decision.

The potential benefits of successful multi-modal integration prove substantial. If eye and tongue characteristics reflect partially independent aspects of cardiovascular function, their combination could substantially improve diagnostic accuracy through complementary information. Even if one modality proves superior overall, the secondary modality might excel for specific patient subgroups, disease subtypes, or image quality scenarios. Multi-modal approaches also offer redundancy improving robustness: if one imaging view proves technically inadequate, the system could fall back to the alternative modality rather than failing completely.

11. Limitations and Challenges

Several limitations constrain interpretation and application of these findings. The modest dataset size of 830 images likely proved insufficient for deep learning to discover robust discriminative patterns, as convolutional neural networks typically require thousands or tens of thousands of training examples for strong performance on complex visual recognition tasks. The network architecture, while appropriate for the available data, represents a relatively simple design that might benefit from increased capacity, deeper architectures, or incorporation of modern design principles such as residual connections, batch normalization, or attention mechanisms.

The lack of rigorous image standardization in the source data, while reflecting real-world screening conditions, introduced substantial variability that complicated pattern recognition. Professional medical imaging employs strict protocols for lighting, positioning, focal distance, and exposure to ensure consistent, reproducible images suitable for diagnostic interpretation. Our more variable data tested robustness but likely degraded performance relative to standardized clinical photography. Future data collection should emphasize consistent image capture protocols while maintaining practical feasibility for large-scale screening.

The extensive manual quality control, while necessary given source image variability, proves difficult to scale for larger datasets. Developing automated quality assessment tools using image quality metrics, lighting analysis, and orientation detection would enable efficient processing of substantially larger image collections. The current approach also introduced potential selection bias: manual curation might systematically exclude certain image types that actually contain diagnostic information, or include others based on subjective quality judgments that inadvertently correlate with disease status.

The confidential nature of medical images prevents public dataset release that would enable independent validation and methodological comparison. While protecting patient privacy remains paramount, this limitation hinders reproducibility and community engagement that accelerate scientific progress. Federated learning approaches enabling model training across distributed datasets without sharing raw images, or generation of realistic synthetic images preserving statistical properties while protecting individual privacy, might help address this challenge in future research.

12. Future Work and Research Directions

The most immediate priority involves establishing correspondences between eye and tongue images from the same individuals to enable multi-modal classification development. Once correspondences are established, the research can proceed to comparative analysis of single versus multi-modal performance, fusion architecture optimization, and investigation of whether the modalities provide complementary or redundant information.

Dataset expansion through additional image collection would substantially improve deep learning performance. A target of 2,000 to 5,000 images per class would better support convolutional neural network training while enabling reserved validation sets for hyperparameter tuning separate from final test evaluation. Collecting multiple images per subject under varying conditions would improve robustness and enable subject-level rather than image-level classification. Longitudinal data tracking individuals over time could reveal dynamic patterns and validate whether tongue changes correlate with cardiovascular disease progression.

Architectural improvements including deeper networks, modern designs incorporating residual connections and attention mechanisms, and transfer learning from models pretrained on large medical image datasets could improve feature learning from limited data. Hybrid approaches combining automatic feature learning with explicit encoding of medical knowledge about diagnostic tongue characteristics might capture the benefits of both deep learning and expert feature engineering. Ensemble methods combining multiple architectures or training runs could improve prediction reliability.

Explainability techniques such as gradient-weighted class activation mapping could visualize which tongue regions drive classification decisions, enabling clinical validation of whether the model attends to diagnostically relevant areas. This interpretability proves essential for clinical acceptance and regulatory approval, allowing domain experts to assess whether automated predictions rely on medically plausible patterns rather than spurious correlations. Uncertainty quantification providing confidence estimates for each prediction would enable intelligent triage, flagging uncertain cases for human review.

13. Conclusions

This project successfully established a complete pipeline for tongue-based cardiovascular disease classification from non-standardized source images through extensive manual quality control, systematic preprocessing, and deep learning model development. While the achieved performance of 60.55 percent AUC falls short of clinical utility thresholds and lags behind the parallel eye-based system's 73.2 percent AUC, the work demonstrated feasibility of computational tongue analysis and identified key challenges for future improvement.

The extensive manual preprocessing required to curate usable images from variable source data underscores the importance of data quality in medical machine learning. The effort invested in removing images with inadequate lighting, incorrect orientation, and technical defects proved essential for reliable model development but highlighted the need for automated quality control approaches to scale this work. The experience gained through manual review provides valuable insights for developing such automation, having identified the specific image characteristics requiring assessment.

The most exciting future direction involves multi-modal integration of eye and tongue images from matched individuals. This approach could substantially improve diagnostic accuracy by combining complementary information from multiple physiological indicators of cardiovascular health. The successful eye-based classification system demonstrates that computational analysis of non-invasive images can achieve clinically meaningful performance, while the tongue-based approach, despite modest current results, captures different aspects of systemic health that might prove valuable in combination.

Ultimately, this research contributes to a broader vision of affordable, accessible cardiovascular screening using simple smartphone photography rather than expensive medical equipment. While current performance does not support clinical deployment, the methodology established here provides a foundation for continued development. With larger datasets, refined architectures, and multi-modal integration, automated analysis of tongue and eye photographs might eventually provide valuable preliminary screening to identify at-risk individuals for follow-up with traditional diagnostic testing, particularly in resource-limited settings where access to cardiac imaging and specialist care remains constrained.

14. Citations

1. Bhosekar, S., Singh, P., Garg, D., Ravi, V., & Diwakar, M. (2023). Simulated Annealing with Deep Learning Based Tongue Image Analysis for Heart Disease Diagnosis. *Intelligent Automation & Soft Computing*, 37(1). https://doi.org/10.32604/iasc.2023.037476
2. Duan, M., Mao, B., Li, Z., Wang, C., Hu, Z., Guan, J., & Li, F. (2024). Feasibility of tongue image detection for coronary artery disease: based on deep learning. *Frontiers in Cardiovascular Medicine*, 11, 1384977. https://doi.org/10.3389/fcvm.2024.1384977
3. Duan, M., Zhang, Y., Liu, Y., Mao, B., Li, G., Han, D., & Zhang, X. (2024). Machine learning aided non-invasive diagnosis of coronary heart disease based on tongue features fusion. *Technology and Health Care*, 32(1), 441-457. https://doi.org/10.3233/THC-230590
4. Li, Y., et al. (2025). AI-driven multimodal fusion of tongue images and clinical indicators for identifying MAFLD patients at risk of coronary artery disease: An exploratory study. *ScienceDirect*. (AUC 0.858-0.933)
5. Wang, X., et al. (2025). A Machine Learning Model Integrating Tongue Image Features and Myocardial Injury Markers Predicts Major Adverse Cardiovascular Events in Patients with Coronary Heart Disease. *PMC*, PMC12242533.
6. Zhou, T., Ruan, S., & Canu, S. (2024). A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 176, 108552. https://doi.org/10.1016/j.compbiomed.2024.108552