applied sciences

MDPI

Article

Convolution Neural Network Based Multi-Label Disease Detection Using Smartphone Captured Tongue Images

Vibha Bhatnagar * and Prashant P. Bansod

Department of Biomedical Engineering, Shri G. S. Institute of Technology & Science, Indore 452003, India; pbansod@sgsits.ac.in

* Correspondence: vbhatnagar@sgsits.ac.in

Abstract: Purpose: Tongue image analysis for disease diagnosis is an ancient, traditional, non-invasive diagnostic technique widely used by traditional medicine practitioners. Deep learning-based multi-label disease detection models have tremendous potential for clinical decision support systems because they facilitate preliminary diagnosis. Methods: In this work, we propose a multi-label disease detection pipeline where observation and analysis of tongue images captured and received via smartphones assist in predicting the health status of an individual. Subjects, who consult collaborating physicians, voluntarily provide all images. Images thus acquired are first and foremost classified either into a diseased or a normal category by a 5-fold cross-validation algorithm using a convolutional neural network (MobileNetV2) model for binary classification. Once it predicts the diseased label, the disease prediction algorithm based on DenseNet-121 uses the image to diagnose single or multiple disease labels. Results: The MobileNetV2 architecture-based disease detection model achieved an average accuracy of 93% in distinguishing between diseased and normal, healthy tongues, whereas the multilabel disease classification model produced more than 90% accurate results for the disease class labels considered, strongly indicating a successful outcome with the smartphone-captured image dataset. Conclusion: AI-based image analysis shows promising results, and an extensive dataset could provide further improvements to this approach. Experimenting with smartphone images opens a great opportunity to provide preliminary health status to individuals at remote locations as well, prior to further treatment and diagnosis, using the concept of telemedicine.

Keywords: tongue feature extraction; disease diagnosis; segmentation; MobileNetV2 architecture; DenseNet-121 architecture; convolutional neural network

Citation: Bhatnagar, V.; Bansod, P.P. Convolution Neural Network Based Multi-Label Disease Detection Using Smartphone Captured Tongue Images. *Appl. Sci.* **2024**, *14*, 4208. https://doi.org/10.3390/app14104208

Academic Editors: István Vassányi, István Kósa and László Balkányi

Received: 4 April 2024
Revised: 9 May 2024
Accepted: 12 May 2024
Published: 15 May 2024

1. Introduction

A technological breakthrough has made clinical investigations and tests for disease diagnosis feasible to a great extent; high-end diagnostic tools are readily available. Ayurveda uses certain human indices like pulse (Nadi), eyes, nails, and tongue to diagnose common ailments. Ancient traditional medicine used these parameters as a means of diagnosis, with no advanced technological tools available, to predict many abnormalities related to the health status of internal body organs. These traditional, non-invasive diagnostic practices can be adapted to give a preliminary prognosis by quantifying and automating the entire analysis process to remove the associated subjectivity.

The paper presents a novel method for multilabel disease classification from tongue analysis. Observation of the tongue is an important part of traditional Chinese medicine as well as Ayurveda. It was established that different parts of the tongue correspond to different internal organs. The tip of the tongue reflects the heart and lungs; the middle part reflects the spleen, pancreas, and stomach; its root represents the kidneys and intestines, and the right and left sides represent the liver and gall bladder.

Over the past few decades, the use of decision-support systems in clinical practice has increased. Deep learning models allow reliable classification and object detection of medical images, showing remarkable accuracy comparable to that of physicians. An automated tongue diagnosis system is a way to bridge the gap between traditional diagnostic methods and modern western medicine. Quantitative analysis can overcome the subjective aspect, which stems from a skilled traditional practitioner's knowledge base and meticulous practice and establish its accountability and acceptability. In this paper, we have developed a multi-disease classification algorithm that is able to predict multiple disease classes with appreciable accuracy. According to the collected data samples, eight diseased states are considered. This paper contributes to the following fields:

- The paper presents a MobileNetV2 deep learning model to detect the diseased tongue from a normal, healthy tongue. The model employs a five-fold cross-validation and transfer learning technique to achieve this binary classification. The main achievement is the sanguine results obtained with images taken by smartphone cameras rather than standard equipment.
- A multilabel classification model for eight common categories of ailments is developed. The DenseNet-121 architecture for disease classification achieves satisfactory results with the small dataset. There are eight disease labels: diabetes, hypertension, acidic peptic disease, pyrexia, hepatitis, cold cough, gastritis, and others.

The organisation of this paper is as follows: Section 2 summarises the techniques used for tongue feature extraction. Section 3 provides an explanation of the related work, dataset, and training details, along with the experimentation and evaluation metrics used. Finally, Section 4 reports the performance analysis. Section 5 concludes the research.

2. Related Work

Quantifying the tongue's diagnostic attributes is one of the main challenges in automating tongue analysis for disease diagnosis. Ref. [1] presents a study of tongue conditions based on Ayurveda, which pertain to an individual's health status. Ref. [2] presented a summary of various tongue attributes such as colour coating, texture, and geometric shape to predict specific diseased conditions, followed in oriental medicine. According to TCM (traditional Chinese practice), the best illumination for tongue inspection is sunshine in an open area at 9 am. Artificially, this can be generated with a source with a colour rendering index greater than ninety and a colour temperature around 5000 K. The automation of tongue analysis systems essentially requires an image-capturing device with high-resolution images for accurate extraction of tongue features for disease predictions, in agreement with traditional medicine practitioners. Researchers explore various imaging setups using high-end CCD cameras [3–9], hyperspectral cameras [10–16], and smartphone cameras [17–24]. Before feature extraction and classification, it is necessary to segment the tongue region from the captured images, which include teeth, lips, and skin areas. Over time, researchers have explored conventional approaches, artificial neural networks, and deep learning algorithms for tongue area segmentation and feature extraction. Probably due to the lack of a digital dataset covering all possible features for various diseases, most of the work in this area focused on a specific disease and its associated tongue features. Some common diseases, such as diabetes, appendicitis, and gastritis, were targeted, and relevant features and classification were done using statistical techniques [25–28]. A hybrid model with statistical methods for feature extraction and machine learning algorithms for classification was also developed by [29–33]. Table 1 summarises some feature and disease classification methods published.

Table 1. Summary of a literature review of ML-based models for specific tongue features.

Tongue Features/Disease Targeted	Dataset	Method Employed	Reference
5 tongue body colours, 6 tongue coating colours	1080 images acquired using the DSO1 state-of-the art acquisition system	k-means clustering algorithm	[34]
Diabetes - Texture and Coating features	The TFDA-1 captured 732 subject images.	The auto-encoder algorithm extracts tongue features, and then the k-means algorithm fuses the two sets of features for classification.	[35]
The study focuses on tongue area detection, calibration, and constitution classification.	50 subjects	Tongue detection using faster RCNN, feature extraction models ResNet-50, VGG-16, and Inception-V3, alongside LBP for texture features and Colour-Moment for colour features The model was evaluated using the classifiers SVM and Decision Tree.	[36]
Colour andTexture features	702 images,	The Grey Level Co-occurrence Matrix (GLCM), in conjunction with the LEAD (Multilabel Learning Algorithm) and a threshold-determining algorithm, yields superior results compared to other existing techniques.	[37]
Multifeature extraction	268 images,	GLA (Generalised Lloyd Algorithm) to extract colour and texture features from the tongue surface.	[38]
Seven categories: fissured tongue, tooth-marked, statis, spotted, greasy, peeled, and rotten coating	8676 images	The faster R-CNN, a region-based network, achieved an accuracy of 90.67%.	[39]
considered 11 features on the tongue surface.	482 images	The ResNet-34 architecture has achieved 86% accuracy for the 11 features identified.	[40]
Tooth-marked tongue related to spleen deficiency	1548 images	ResNet-34 architecture, 90% accuracy.	[41]
Gastritis	263 gastritis patients, 48 healthy	Features related to gastritis were extracted using a constrained, high-dispersal neural network.	[42]
		Ada Boost, SVM (support vector machine), and MLP (multilayer perceptron classifier) are some examples.	[43]
11 disease categories, plus healthy tongue images.	There are 936 images, with 78 images for each of the 12 disease categories, including healthy ones.	The VGG-19 network's extracted tongue features, supported by a Random Forest classifier, achieved 93.7% accuracy	[44]
12 disease categories, including healthy	936 images For each of the 12 disease categories, there are 78 images.	Designed IoT base automated synergic deep learning tongue colour image analysis model providing 98.3% accuracy for disease diagnosis and classification.	[45]
Iron deficiency	95 images from the Harvard dataset	Explored the possibility of monitoring health status by tongue images using the CNN algorithm, which could be deployed on an Android mobile app.	[46]

Ref. [47] reviewed current trends in tongue diagnosis. They also trained a classification model using Random Forest and Support Vector Machine on a tongue dataset, with the tongue region divided into five parts as per the layout of the internal organs and extracted seven colour spaces from the five extracted parts. Further, they trained VGG and ResNet pretrained models on tongue constitution classification. ResNet 50 showed the best performance, with 64.52% accuracy.

Over the past two decades, enormous efforts have been made to effectively utilise the full potential of the tongue diagnosis system. Enhancement of tools and techniques for achieving the holistic approach and overcoming the subjective nature of diagnosis is an ongoing process, and there is a need to set some standard protocols so that they have uniform characteristics and are readily acceptable by end users. Most of the research has focused on specific diseases and sought to extract tongue features. Refs. [44,45] achieved notable and accurate results by classifying 12 disease categories, including a healthy tongue, and assigning a specific single disease label to each tongue sample; they did not focus on multi-label classification. None of the references mentioned aimed to classify multiple disease labels.

The aim of this paper is to take a step towards developing a decision support system using deep architecture to predict a holistic tongue diagnosis, not just any particular tongue features or diseases. In the future, it is proposed to further develop it as a mobile app that allows users to consult an expert medical professional and receive a preliminary treatment for immediate relief before conducting detailed investigations.

3. Methodology

The primary goal of this study is to determine whether smartphone images can yield accurate results for tongue analysis, enabling the provision of a preliminary diagnosis whenever and wherever needed. Tongue images captured by an individual and symptoms mentioned by him or her can aid the medical professional in giving a preliminary line of treatment before directing further detailed investigations. We collaborated with two consulting physicians to collect data over a two-month period using smart phone devices with a camera resolution greater than or equal to 8 megapixels. Analysis and identification of the images to be labelled with single and/or multiple disease labels for each case are performed.

We classified seven common ailments, each with a significant number of samples, and grouped the remaining diseases diagnosed, with few samples, under a single label, 'others'. The doctor's diagnosis serves as the ground truth for the classification task. The approach used for multilabel classification starts with the initial segmentation of the tongue area from the complete image. Thereafter, a binary classification architecture distinguishes an unhealthy tongue from a healthy one, followed by final disease label prediction using a multi-label classifier model. The workflow for implementing a multi-label classification model is illustrated in Figure 1. Subsequent sub-sections follow, providing brief details of all the steps involved. The strategy for the deep learning classification models used is as follows:

- Stratified 5-fold cross-validation for disease risk classification.
- Ensemble learning strategy: bagging to reduce variance in data.
- Upsampling of the dataset to set some minimum sample size in each class.
- Extensive real-time data augmentation is needed for training models.
- Class-weighted focal loss to tackle class imbalance.
- Individual training for multi disease classification and disease risk detection.

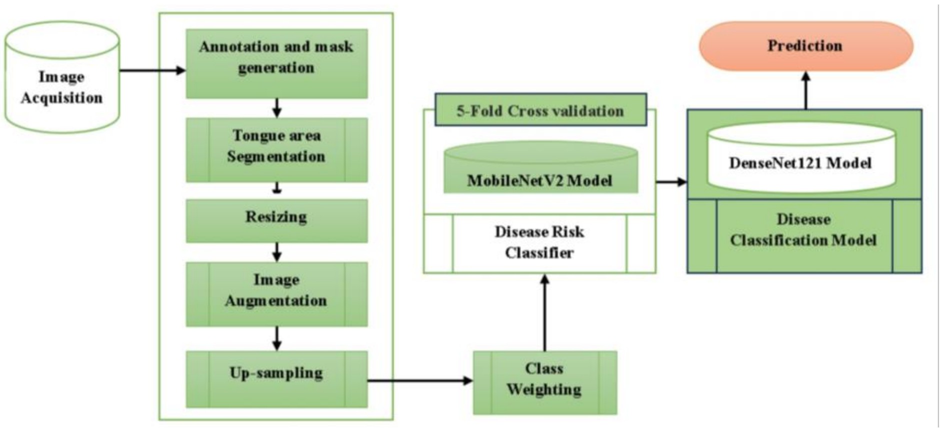


Figure 1. Proposed pipeline for multi-label disease detection using tongue images.

3.1. Tongue Analysis Dataset

A total of 1095 images of subjects suffering from one or more than one ailment are acquired using smartphone cameras with an image resolution greater than or equal to 8 mega pixels (Samsung A50, iPhone, one plus). A total of 822 images of healthy individuals were also collected from willing individuals. All images were collected with the necessary consent of individuals eager to be part of our study. Raw images come in a variety of resolutions and sizes. For the proposed model, images are annotated with eight conditions other than normal and disease risk categories, as listed in Table 2. Other label classes include some uncommon ones like CAD (coronary artery disease), CKD (chronic kidney disease), COPD (Chronic Obstructive Pulmonary Disease), epilepsy, and vertigo.

Table 2. Annotation frequency for each class in the dataset.

Disease	Samples	Disease	Samples
Diabetes (DM)	112	Hepatitis	183
Blood Pressure (BP)	138	Cold Cough	150
Acid Peptic Disease (APD)	156	Gastritis	189
Pyrexia	98	Others	429

3.2. Preprocessing and Image Augmentation

The acquired images are of various sizes and formats, and we perform primary processing by converting all images to a 256 × 256-pixel jpeg format. These basic preprocessing steps are used ensure a uniformity in all images in the dataset. The next step involves segmentation of the tongue area of interest by Double U-Net architecture [48]. In order to increase data variability, further preprocessing methods are used, such as image augmentation for upsampling to balance class distribution and real-time augmentation during training to obtain unique and novel images in each epoch, thus improving the model's performance. Upsampling is done to ensure each label occurs at least 150 times in the dataset, which increased the total number of diseased tongue images to 2729. Rotation, flipping, and altering brightness, saturation, and hue are used for real-time augmentation.

3.3. Deep Learning Models

Our pipeline combines two different types of image classification methods: a binary classification for normal or diseased tongues and a disease label classifier for multilabel annotated images. The AUCMEDI [49] platform is employed to develop our classification model; in both cases, we are using pretrained models to reduce the time and cost of training a fresh model. We apply transfer learning with frozen layers, with the exception of the classification head. After 10 epochs of training, the freezing is undone so that the weights can be adapted to the new task.

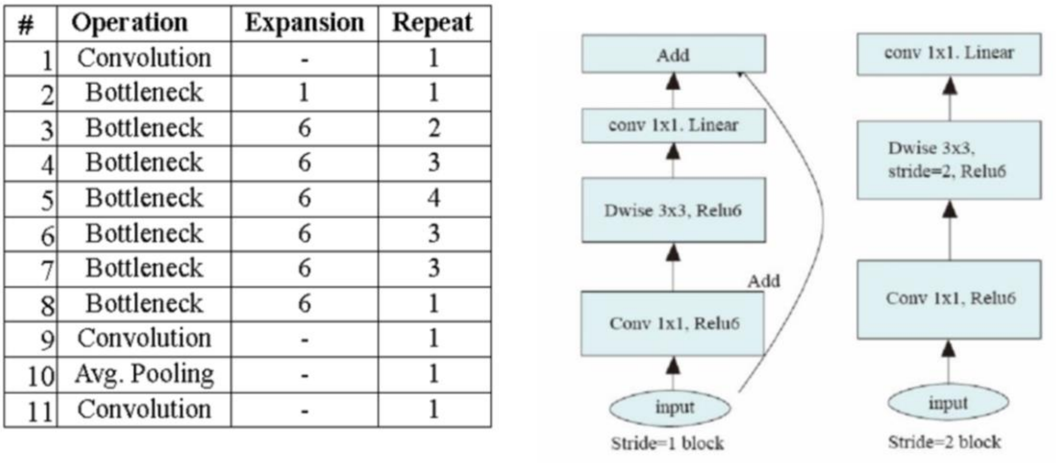
We chose the two architectures to be compatible with low resource requirements and feasible for deployment on the Android platform. A brief introduction to the two architectures is provided in the following paragraphs, along with the parameter settings for each model.

3.3.1. Diseased Tongue Detector

MobileNetV2 is used for binary classification to distinguish between normal and unhealthy tongue images. MobileNet [50] is the first computer vision model open-sourced by Google and designed for training classifiers, detectors, and segmentation. The model's highlights are its small, low-latency, and low-power models, which are designed to effectively maximise accuracy while considering resource constraints for on-device or embedded applications. The use of depth-wise separable convolutions significantly reduces the number of parameters compared to other regular convolution networks. MobileNetV2, as defined in [50], basically contains 53 convolution layers and 1 average pooling layer, as shown in Figure 2a. It incorporates two new architectural elements within its layout: a linear bottleneck across the layers and a shortcut between the bottlenecks. The initial layer is a fully convolutional layer with 32 filters, followed by 19 residual bottleneck layers. The structure consists of two crucial components:

1. Residual Block inverted;
2. Residual Bottleneck Block.

Each block has three different layers, consisting of point-wise 1 × 1 convolution and 3 × 3 depth-wise convolution. There are Stride 1 blocks and Stride 2 blocks, as shown in Figure 2b. Bottleneck is either an inverted residual block, a bottleneck residual block, or a Stride 1 or Stride 2 block. Bottleneck layers enhance models' ability to transition from lower-level pixels to higher-level variables identifying image categories.



#	Operation	Expansion	Repeat
1	Convolution	-	1
2	Bottleneck	1	1
3	Bottleneck	6	2
4	Bottleneck	6	3
5	Bottleneck	6	4
6	Bottleneck	6	3
7	Bottleneck	6	3
8	Bottleneck	6	1
9	Convolution	-	1
10	Avg. Pooling	-	1
11	Convolution	-	1

(a) (b)

Figure 2. Basic Blocks of MobileNetV2 (**a**) Layers in MobileNet V2. (**b**) Stride 1, stride 2 blocks [50].

The use of 5-fold cross-validation with the bagging method for ensembles with Random Forest and Soft Majority Voting is done. The dataset is split into three parts: a train, a test model, and a test ensemble set. The five-fold cross-validation further divides the train set into a train and a validation set, dividing the training set into five parts and performing training on each part. Each time, one of the five parts is used for validation, and the other four parts are used for training. Five models are trained according to a 5-fold train dataset split; individual models may not give the best results by themselves. The best model for each fold is saved. The predictions of the individual models are combined into an ensemble in order to produce a single, stable prediction. Two techniques—Random Forest (RF) and Soft Majority Voting (SMV)—are used, and results are compared.

Random Forest

The Random Forest algorithm itself is an ensemble method. The implementation steps begin with a bootstrapped dataset, from which we draw random samples. We can draw one sample from the dataset multiple times, resulting in a new set of data that is the same size as the original, but does not necessarily contain all of the original dataset. The next step is to build a decision tree with random variables based on the number of folds. The bootstrapping step is repeated, thereby generating multiple decision trees. The final step is tree bagging for the prediction of new data; the most common decision becomes the final result of the decision tree. For the bagging method, the Random Forest classifier uses an ensemble dataset.

Soft Majority Vote

For SMV, the summation of each model's prediction is done to create a new prediction matrix. The final prediction for new data then becomes the category having the maximum value in the matrix.

Parameter settings for the said model are as follows; due to the imbalance in the number of images in the two categories, class weights are computed, and categorical focal loss is considered the loss function. Testing of model performance is done on the test data, and evaluation metrics for training each fold are computed and saved. Ensemble by 5-fold bagging is done by Random Forest and Soft Majority Vote methods. Each model is trained for 50 epochs, and the dynamic learning rate is set to a maximum decrease of 1×10^{-7} with a factor of 0.1 in case of no improvement. Validation loss is monitored after 5 epochs. Early stopping and the model checkpoint method are also used, with a patience level set at 12 epochs with no improvements in validation loss. Finally, we compare all the individual fold F1 scores with the ensemble models to determine which model performs best.

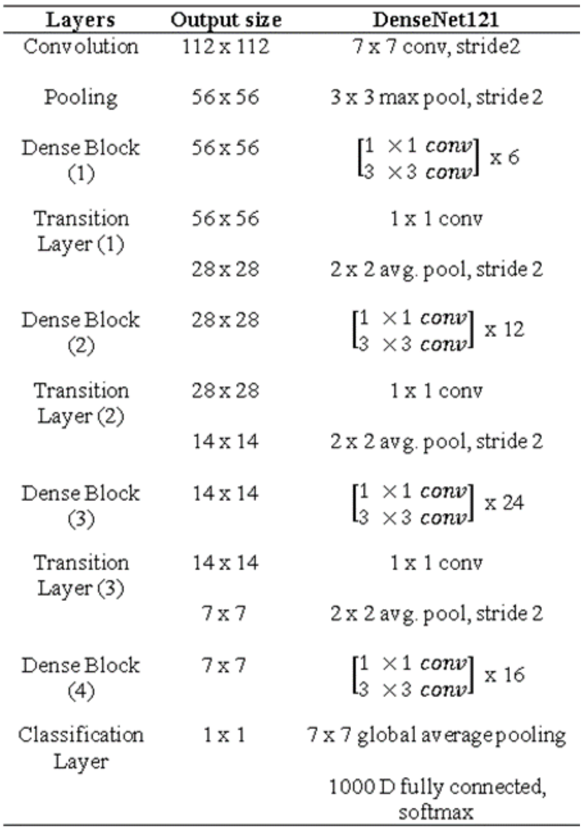
3.3.2. Disease Label Classifier

This is a multilabel classifier using the DenseNet-121 [51] architecture for the eight disease classes. DenseNet, as its name implies, is a densely connected network that connects each layer to every other layer. The input of a layer is the concatenation of feature maps from previous layers. Concatenation requires feature maps of similar size; this issue is resolved by dividing the network into multiple densely connected networks, facilitating both down sampling and feature concatenation. This keeps the feature map size similar within a block.

$$x_l = H_l([x_0, x_1, x_2, \dots x_{l-1}]) \tag{1}$$

where x_0 represents the input, H_l represents the non-linear transformation occurring in each block, and x_l represents the output of the Lth layer. A transition block, consisting of a convolution and pooling layer, follows each dense block. Lastly, the dense block is terminated by a classification layer that accepts the feature map of all previous layers of the network to perform the classification task. Bottleneck layers are used within each block, where 1 × 1 convolution reduces the number of channels in the input, followed by 3 × 3 convolution for feature extraction. This helps to relax large computational requirements while also improving efficiency. DenseNet-121 consists of four dense blocks with [6,12,24,16] number of layers in each block as shown in the summary chart in Figure 3.

This network has the following imperative advantages: it mitigates the vanishing gradient problem, strengthens feature propagation, encourages feature reuse, and has a smaller number of parameters.



Layers	Output size	DenseNet121
Convolution	112 x 112	7 x 7 conv, stride2
Pooling	56 x 56	3 x 3 max pool, stride 2
Dense Block (1)	56 x 56	$\begin{bmatrix}1 & \times 1\ conv\\ 3 & \times 3\ conv\end{bmatrix}$ x 6
Transition Layer (1)	56 x 56	1 x 1 conv
	28 x 28	2 x 2 avg. pool, stride 2
Dense Block (2)	28 x 28	$\begin{bmatrix}1 & \times 1\ conv\\ 3 & \times 3\ conv\end{bmatrix}$ x 12
Transition Layer (2)	28 x 28	1 x 1 conv
	14 x 14	2 x 2 avg. pool, stride 2
Dense Block (3)	14 x 14	$\begin{bmatrix}1 & \times 1\ conv\\ 3 & \times 3\ conv\end{bmatrix}$ x 24
Transition Layer (3)	14 x 14	1 x 1 conv
	7 x 7	2 x 2 avg. pool, stride 2
Dense Block (4)	7 x 7	$\begin{bmatrix}1 & \times 1\ conv\\ 3 & \times 3\ conv\end{bmatrix}$ x 16
Classification Layer	1 x 1	7 x 7 global average pooling
		1000 D fully connected, softmax

Figure 3. Summary of DenseNet-121 Architecture.

For this study, multilabel focal loss is used as the loss function along with categorical accuracy. The model is trained with hyperparameter settings as mentioned: 100 epochs with a dynamic learning rate set to a maximum decrease of 1×10^{-7} with a factor of 0.1 in case of no improvement. It is also monitored for validation loss with a patience of 5 epochs. Early stopping and the model checkpoint method are also used to stop after 12 epochs with no improvements monitored.

The experiments were performed with the available resources, consisting of a HP Pavilion laptop with a 1.60 GHz Intel i5 8th generation processor and 8 GB of RAM. Training of the model was done on Google Colab Python 3. Backend GPU for Google Compute Engine.

3.4. Evaluation Metrics

The metrics considered for quantitative evaluation of the model on test data are precision, recall, and F1-score.

Precision is the number of true positives divided by the number of total positive predictions.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

Recall measures the model's ability to predict the positives; it is the true positive divided by the true positive and false negative.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive\ (TP) + False\ Negative(FN)}$$

*F*1 *score* is the harmonic mean of precision and recall given by

$$F1Score = 2 \times \frac{Precision \ \times Recall}{Precision + Recall}$$

Categorical Accuracy is the percentage of predicted values that match the actual truth values. The calculation for the same is done as follows:

- First, identify the index at which the maximum value occurs using argmax.
- If it is the same for both predicted and true value, it is considered accurate.

Here, since the maximum value index is observed, the predicted value can be a logit or probability function.

$$Categorical\ accuracy = \frac{accurately\ predicted\ records}{total\ number\ of\ records}$$

- *Reciever Operating Characteristics Curve*

The receiver operating characteristic (ROC) curve is determined by calculating the average ROC for each of the five folds. Practically, for each fold, different false- and true-positive-rates exist, and the mean cannot be calculated; hence, the first aggregation of all False-Positive-Rates (FPR) into one vector is done, serving as the x axis of the average ROC curve. The True-Positive Rate (TPR) needs to be interpolated. The TPR represents the corresponding y-points to the previously collected x-points.

4. Results and Discussions

A disease risk detection model is trained on the dataset with healthy and unhealthy tongue images. Upsampling of the unhealthy image data is done to ensure at least 150 images in each of the 8 categories. Real time Image augmentation is used to increase the image set artificially by adding small transformations to the original images, such as rotations or changes in contrast or saturation. Online image augmentation applies the transformations to each image upon loading with the data generator, eliminating the need for disk storage. For multilabel disease classification evaluation, one hot-encoded file serves as an interface defining true labels for the model.

4.1. Performance Analysis of Disease Risk Detector

When trained on a shared GPU from COLAB, the sequential training of the disease risk detector with 5-fold cross validation with the MobileNetV2 architecture took approximately 4.5 h with 45 epochs for each fold on average. The disease risk model's performance results with respect to metrics considered on the test model dataset are shown in Table 3. The test dataset consists of 485 images in total, with 361 unhealthy tongue images and 124 normal, healthy tongue images. Concise ROC Curve Figure 4, for the 5-folds and ensemble techniques with a mean value, indicates the model's appreciable performance for binary classification.

Table 3. Performance metrices for the 5 folds of cross-validation of the disease risk model.

		Precision	Recall	F1-Score	Accuracy
Fold-1	Diseased	0.99	0.91	0.95	0.92
	Normal	0.78	0.97	0.86	
Fold-2	Diseased	0.97	0.94	0.96	0.93
	Normal	0.85	0.90	0.88	
Fold-3	Diseased	0.98	0.98	0.98	0.96
	Normal	0.93	0.94	0.93	
Fold-4	Diseased	0.97	0.94	0.95	0.93
	Normal	0.84	0.92	0.88	
Fold-5	Diseased	0.98	0.91	0.95	0.92
	Normal	0.79	0.95	0.86	

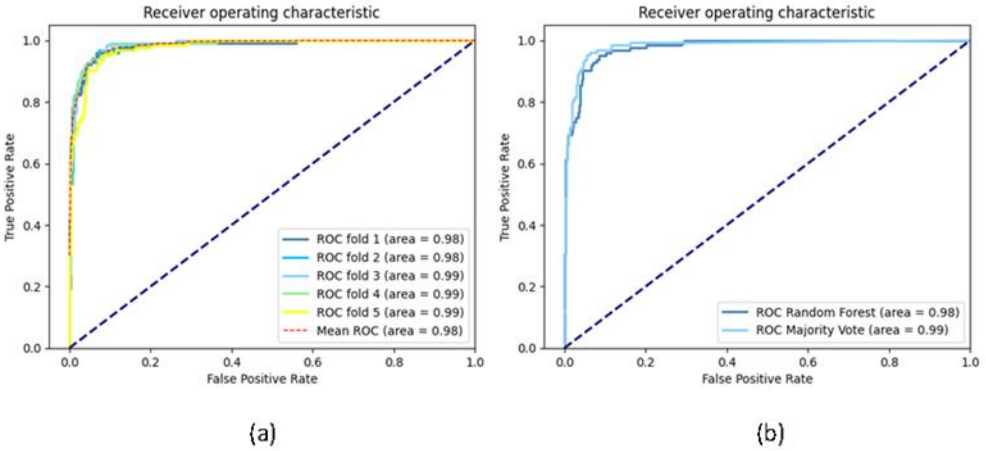


Figure 4. Receiver operating characteristics curve (**a**) 5-fold cross-validation models. (**b**) Ensemble techniques.

The best model is saved for each fold of cross-validation. Predictions of each fold on a test ensemble dataset with 485 images are combined to form an ensemble. Implementation of the Random Forest ensemble is done by picking random samples and features to build many decision trees on a bootstrapped dataset through a repeated process. Prediction of new data is accomplished by bagging operations, wherein the final decision is the most common decision amongst all decision trees.

The Soft Majority Voting Ensemble aggregates the generated predictions from the five folds to create a new prediction matrix. The category with the maximum value is considered the final prediction for each sample. In this particular case, the ensemble of 5-fold predictions with two classes does not show enhanced performance; F1 scores for Random Forest and Soft Majority Voting are comparable to the individual fold F1 score as depicted in Figure 5.

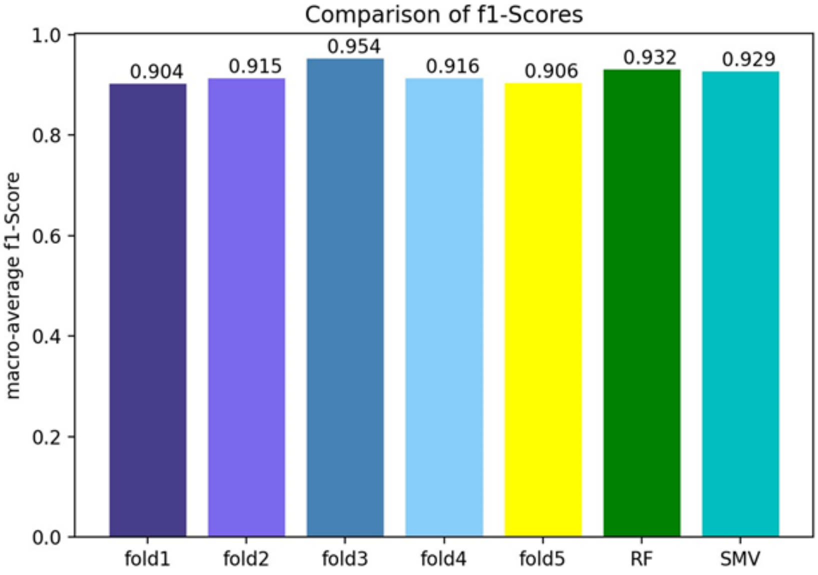


Figure 5. Comparison bar graph for 5-folds and ensemble techniques.

4.2. Performance Analysis of the Multi-Label Disease Classification Model

The DenseNet-121 model is utilised for the classification of diseases. The dataset comprised seven distinct disease labels and an additional one, which consisted of all diseases with fewer than 10 subjects in each category. Model training stopped at 41 epochs. The disease classification model with DenseNet-121 architecture took over 18 h to train. Table 4 shows the performance parameters for the eight labels considered in the test data of 474 images. The ROC curve in Figure 6 gives an idea of true positive predictions for each category.

Table 4. Performance Parameters for the Disease Classification Model.

Disease	Precision	F1-Score	Accuracy
DM	0.9722	0.8203	0.9148
BP	0.9803	0.8658	0.9425
APD	0.9130	0.8038	0.9240
Pyrexia	0.9473	0.9183	0.9703
Hepatitis	0.9885	0.8958	0.9629
Cold Cough	0.9878	0.8901	0.9629
Gastritis	0.9798	0.9652	0.9870
Others	0.9034	0.7553	0.8093

It is observed that each of the diseased classes achieved an average accuracy of over 90%. Some sample results presented in Table 5, where bold values indicate the true predictions with respect to ground truth, the success of the DenseNet-121 model for disease classification.

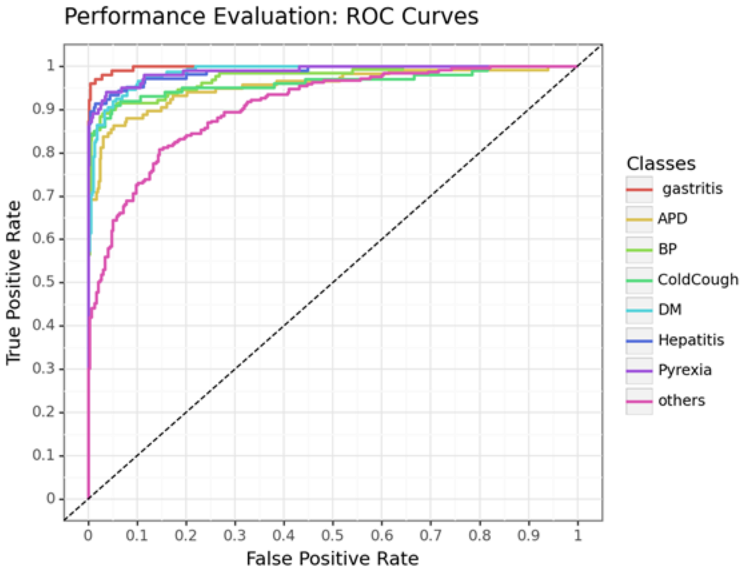
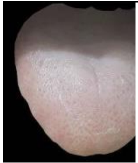
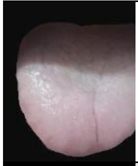
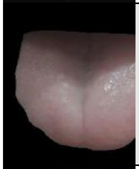
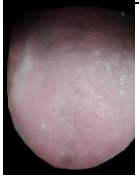
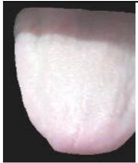
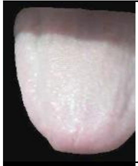
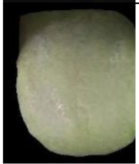
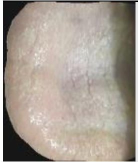
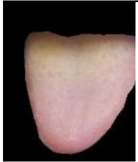
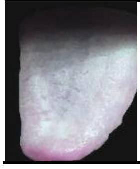


Figure 6. ROC curve for DenseNet-121 model on Test dataset.

Highlighted rows in the table indicate inaccurate classification for two sample images. The results for the class “others” were also observed to be slightly less accurate. This is also evident from the ROC curve and performance indices.

Table 5. Sample test images with prediction probability and truth data.

Test Image Samples	Prediction Probability								Truth
	DM	**BP**	**APD**	**PYR**	**HEP**	**CC**	**GAS**	**Others**	
	0.0456	0.0190	0.0400	0.0131	**0.7392**	0.0207	0.0256	**0.7650**	Hepatitis and others
	**0.9177**	0.0405	0.0055	0.0001	0.0033	**0.9872**	0.0071	**0.9968**	Diabetes, cold cough, others
	0.0129	0.0198	0.0118	**0.4670**	0.0729	0.1006	0.0051	0.3405	Pyrexia
	**0.4555**	**0.6903**	0.0591	0.0014	**0.8166**	0.0654	0.0259	0.3004	Diabetes, hypertension, hepatitis
	0.0059	0.0235	**0.7697**	0.0071	0.0074	0.0315	**0.9693**	**0.8268**	APD, gastritis, others
	0.0004	0.0103	**0.9436**	0.0413	0.0029	0.0142	**0.9705**	**0.7960**	APD, gastritis, others
	0.0335	0.0188	0.0108	0.0987	**0.4677**	0.0124	**0.7906**	0.0795	Hepatitis, gastritis
	0.1365	0.0628	0.2372	0.0261	0.0131	0.2798	0.0423	**0.5847**	Pyrexia
	0.0264	0.0036	0.1321	0.1261	0.0083	**0.4880**	0.1867	0.3476	Cold cough
	**0.9768**	**0.8526**	0.0079	0.0013	0.0009	**0.9293**	0.0010	**0.9519**	DM, hypertension, cold cough, others

The proposed model for multi-disease classification shows satisfactory results. The most important outcome of this work is that images captured in different environments and on various mobile phones demonstrated appreciable results. Further improvement is possible with an ensemble of different deep learning models to achieve high accuracy for all disease labels. As it is not practically possible for any individual model to deliver uniform accuracy for all classification labels under consideration, multiple models will ultimately enhance the overall prediction accuracy. In future work, the aim is to increase the number of disease labels to encompass all potential cases that tongue analysis can diagnose, in collaboration with a medical expert to compile ground truth data for classification.

Even with the limitation of a small dataset with a smaller variety of disease classes, model performance is appreciable. Since only a single dense model is considered, its performance accuracy for all class labels could not be uniform. As previously mentioned, we can enhance this by incorporating the ensemble learning method.

5. Conclusions and Future Work

The main challenging aspect of automating tongue analysis for disease diagnosis is quantifying the diagnostic features of the tongue image on par with the expert practitioner's observations. Automating this non-invasive method of diagnosis can facilitate the consultation of a sick individual with medical personnel using a telemedicine network, even from a remote location. On the Indian continent, with its diverse and large population, automatic tongue analysis systems can aid medical personnel to some extent in giving a preliminary diagnosis and initiating immediate treatment without direct physical interaction with the subject.

In this study, we introduced a powerful multi-disease detection pipeline for tongue image analysis that uses ensemble learning to combine the predictions of five individually trained models to improve the performance of identifying a diseased tongue. In this particular case, the ensemble method's performance is comparable to that of individual models, showing no notable improvement. A single image of the tongue can be used to predict more than one disease by employing techniques like transfer learning, class weighting for imbalance in different classes, large amounts of real-time data enhancement, and focal loss utilisation in the deep learning architecture. An average accuracy of 90% is achieved with the multi-label classification model for eight disease classes. Two different deep learning models are specifically used, primarily due to their lower resource requirements compared to their counterparts and their potential for Android platform deployment.

This study's most significant accomplishment is the successful use of mobile phone images for tongue image analysis, which represents a step towards reducing the cost and expertise of a high-end, sophisticated image-capturing device and strengthens the possibility of developing an Android-based app for easy and quick prognosis. Automation will also be helpful in training aspiring clinicians to use this non-invasive traditional practice and enhance their skills with experience. This opens the opportunity to explore more enhanced models with a larger dataset for performance improvement.

In our future work, we aim to build a more robust model with a judicious selection of a group of deep learning architectures for an ensemble model capable of classifying all possible disease labels achievable using tongue analysis from smartphones. This inherently includes the need to compile an exhaustive tongue image dataset for research.

Author Contributions: V.B. has compiled the work after conducting a conducting a literature survey and developing an algorithm that is then trained on the dataset and rigorously analysed by P.P.B., my guide and mentor. P.P.B. critically revised and corrected the first author's rough draft for intellectual content, ultimately approving the submitted version. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This is an observational study, and hence no ethical approval is required. The individual consent form was obtained from all the subjects and recorded alongside the clinician’s diagnosis. The only use of the collected data is for research purposes, never disclosing the identity of any subject.

Informed Consent Statement: All of the individual participants in the study gave their informed consent. The authors made it clear to them that they would only use the provided images for research purposes and would not disclose any personal information on any platform.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors acknowledge the support of medical professionals who contributed to the collection of datasets and the annotation and labelling of the same. We are also thankful to student volunteers of the department of Biomedical Engineering, SGSITS, Indore.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Patil, M.K. Anatomical Study of Jhiva W.R.T Sam and Niram Prakriti Pariksha. *Int. Ayurvedic Med. J.* **2017**, *1*, 151–159.
2. Vocaturo, E.; Zumpano, E.; Veltri, P. On discovering relevant features for tongue colored image analysis. In Proceedings of the 23rd International Database Applications & Engineering Symposium (IDEAS ‘19), Athens, Greece, 10–12 June 2019; Association for Computing Machinery: New York, NY, USA; pp. 1–8. https://doi.org/10.1145/3331076.3331124.
3. Chiu, C.C. A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue. *Comput. Methods Programs Biomed.* **2000**, *61*, 77–89.
4. Wang, Y.; Zhou, Y.; Yang, J.; Xu, Q. An Image Analysis System for Tongue Diagnosis in Traditional Chinese Medicine. In Proceedings of the International Computational & Information Science Conference, Shanghai, China, 16–18 December 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp.1181–1186.
5. Zhang, H.; Wang, K.; Zhang, D.; Pang, B.; Huang, B. Computer aided tongue diagnosis system. In Proceedings of the 27th Annual Conference on Engineering in Medicine & Biology Society, Sydney, Australia, 17–18 January 2005; IEEE: Piscataway, NJ, USA, 2006; pp. 6754–6757.
6. Jiang, L.; Xu, W.; Chen, J. Digital imaging system for physiological analysis by tongue color inspection. In Proceedings of the 3rd Innovative Engineering & Applications Conference, Washington, DC, USA, 3–5 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp.1833–1836.
7. Xu, J.; Tu, L.; Ren, H.; Zhang, Z. A Diagnostic Method Based on Tongue Imaging Morphology. In Proceedings of the 2nd International Conference on Bioinformatics & Biomedical Engineering, Shanghai, China, 16–18 May 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 2613–2616.
8. Wang, X.; Zhang, B.; Yang, Z.; Wang, H.; Zhang, D. Statistical Analysis of Tongue Images for Feature Extraction & Diagnostics. *IEEE Trans. Image Process.* **2013**, *22*, 5336–5347.
9. Wang, X.; Zhang, D. A High-Quality Color Imaging System for Computerized tongue Image Analysis. *J. Expert Syst. Appl.* **2013**, *40*, 5854–5866.
10. Liu, Z.; Zhang, D.; Yan, J.Q.; Li, Q.L.; Tang, Q.L. Classification of hyperspectral medical tongue images for tongue diagnosis. *Comput. Med. Imaging Graph.* **2007**, *31*, 672–678.
11. Li, Q.; Liu, J.; Xiao, G.; Xue, Y. Hyperspectral tongue imaging system used in tongue diagnosis. In Proceedings of the 2nd International Conference on Bioinformatics & Biomedical Engineering, Shanghai, China, 16–18 May 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 2579–2581.
12. Li, Q.; Wang, Y.; Liu, H.; Sun, Z.; Liu, Z. Tongue fissure extraction and classification using hyperspectral imaging technology. *Appl. Opt.* **2010**, *49*, 2006–2013.
13. Li, Q.; Lui, Z. Tongue color analysis and discrimination based on hyper spectral images. *Comput. Med. Imaging Graph.* **2009**, *33*, 217–221.
14. Yamamoto, S.; Tsumura, N.; Nakaguchi, T.; Namiki, T.; Kasahara, Y.; Terasawa, K.; Miyake, Y. Early Detection of Disease Oriented State from Hyperspectral Tongue Images with Principal Component Analysis and Vector Rotation. In Proceedings of the Annual International Conference on Engineering in Medicine & Biology Society, Buenos Aires, Argentina, 31 August–4 September 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 3025–3028.
15. Li, Q.; Wang, Y.; Liu, H.; Sun, Z. AOTF based Hyperspectral Tongue Imaging System and Its Applications in Computer-aided Tongue Disease Diagnosis. In Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics; Yantai, China, 16–18 October 2010, IEEE: Piscataway, NJ, USA, 2010; pp. 1424–1427.
16. Liu, Z.; Wang, H.J.; Li, Q. Tongue Tumour Detection in Medical Hyperspectral Images. *Sensors* **2012**, *12*, 162–174.
17. Ryu, I.; Itiro, S. A tongue diagnosis system for personal healthcare on smartphone. In Proceedings of the 5th Augmented Human International Conference, Kobe, Japan, 7–9 March 2014; Waseda University Press: Tokyo, Japan, 2014.
18. Duan, Y.; Xu, D. *I Tongue: An iPhone App for Personal Health Monitoring Based on Tongue Image*; Final Report; Interdisciplinary Innovations Fund (IIF) 2012/2013 Awards (MU); University of Missouri: Columbia, MO, USA, 2014.

19. Hu, M.C.; Cheng, M.H.; Lan, K.C. Color Correction Parameter Estimation on the Smartphone and its Application to Automatic Tongue Diagnosis. *J. Med. Syst.* **2016**, *40*, 18.
20. Tania, M.H.; Lwin, K.T.; Hossain, M.A. Computational Complexity of Image Processing Algorithms for an Intelligent Mobile Enabled Tongue Diagnosis Scheme. In Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications, Denver, CO, USA, 15–16 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 29–36.
21. Kanawong, R.; Obafemi-Ajayi, T.; Liu, D.; Zhang, M.; Xu, D.; Duan, Y. Tongue Image Analysis and Its Mobile App Development for Health Diagnosis. *Adv. Exp. Med. Biol*. **2017**, *1005*, 99–121.
22. Hu, M.C.; Lan, K.C.; Fang, W.C.; Huang, Y.C.; Ho, T.J.; Lin, C.P.; Yeh, M.H.; Raknim, P.; Lin, Y.H.; Cheng, M.H.; et al. Automated Tongue Diagnosis on the Smartphone and its Applications. *Comput. Methods Programs Biomed.* **2019**, *174*, 51–64.
23. Zhou, Z.; Peng, D.; Gao, F.; Leng, L. Medical Diagnosis Algorithm Based on Tongue Image on Mobile Device. *J. Multimed. Inf. Syst.* **2019**, *6*, 99–106.
24. Smith, Z.J.; Chu, K.; Espenson, A.R.; Rahimzadeh, M.; Gryshuk, A.; Molinaro, M.; Dwyre, D.M.; Lane, S.M.; Matthews, D.; Wachsmann-Hogiu, S. Cell-Phone-Based Platform for Biomedical Device Development and Education Applications. *PLoS ONE* **2011**, *6*, e17150.
25. Haralick, R.M.; Shanmugan, K.; Dinstein, I. Textural features for Image Classification. *J. IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621.
26. Pang, B.; Zhang, D. Computerized tongue diagnosis based on Bayesian networks. *IEEE Trans. Biomed. Eng*. **2004**, *51*, 1803–1810.
27. Pang, B.; Zhang, D.; Wang, K. Tongue Image analysis for appendicitis diagnosis. *Inf. Sci.* **2005**, *175*, 160–176.
28. Kim, J.; Son, J.; Jang, S.; Nam, D.H.; Han, G.; Yeo, I.; Ko, S.J.; Park, J.W.; Ryu, B.; Kim, J. Availability of Tongue diagnosis system for Assessing Tongue Coating thickness in Patients with Functional Dyspepsia. *Evid Based Complement. Altern. Med.* **2013**, *2013*, 348272.
29. Jayanti, S.K.; Shanmugapriyanga, B. Detecting Diabetes Mellitus Gradient Vector Flow Snake Segmented Technique. *Int. Res. J. Eng. Technol.* **2017**, *4*, 1238–1244.
30. Zhang, D.; Zhang, H.; Zhang, B. Detecting Diabetes Mellitus and No proliferative Diabetic Retinopathy Using CTD. In *Tongue Image Analysis*; Springer: Singapore, 2017; pp. 303–325.
31. Sandhya, N.; Rajasekar, M. Tongue Image Analysis for Hepatitis Detection Using GA-SVM. *Indian J. Comput. Sci. Eng.* **2017**, *8*, 526–534.
32. Mrilaya, D.; Pervetaneni, P.; Aleperi, G. An Approach for Tongue Diagnosing with Sequential Image Processing Method. *Int. J. Comput. Theory Eng.* **2012**, 4, 322–328.
33. Dhanalakshmi, M.; Pervetaneni, P.; Aleperi, G. Applying Wavelet Transforms and Statistical Feature Analysis for Digital Tongue Image. *IOSR J. Comput. Eng.* **2014**, *16*, 95–102.
34. Kawanabe, T.; Kamarudin, N.D.; Ooi, C.Y.; Kobayashi, F.; Xiaoyu, M.; Sekine; M.; Wakasugi, A.; Odaguchi, H.; Hanawa, T. Quantification of tongue colour using machine learning in Kampo medicine. *Eur. J. Integr. Med.* **2016**, *8*, 932–941. https://doi.org/10.1016/j.eujim.2016.04.002.
35. Li, J.; Hu, X.; Tu, L.; Cui, L.; Jiang, T.; Cui, J.; Ma, X.; Yao, X.; Shi, Y.; Wang, S.; et al. Diabetes Tongue Image Classification Using Machine Learning and Deep Learning. In Proceedings of the 2008 3rd IEEE Conference on Industrial Electronics and Applications, Singapore, 3–5 June 2008. https://doi.org/10.2139/ssrn.3944579.
36. Ma, J.; Wen, G.; Hu, Y.; Chang, T.; Zeng, H.; Jiang, L.; Qin, J. Tongue image constitution recognition based on Complexity Perception method. *arXiv* **2018**, arXiv:1803.00219. https://doi.org/10.48550/arXiv.1803.00219.
37. Zhang, X.F.; Zhang, J.; Hu, G.Q.; Wang, Y.Z. *Preliminary Study of Tongue Image Classification Based on Multi-Label Learning*; ICIC 2015, Part III LNAI9227; Springer: Berlin/Heidelberg, Germany, 2015. https://doi.org/10.1007/978-3-319-22053-6_23:2080220.
38. Chen, L.; Wang, B.; Zhang, Z.; Lin, F.; Ma, Y. Research on Techniques of Multifeatures Extraction for Tongue Image and Its Application in Retrieval. *Comput. Math. Methods Med.* **2017**, *2017*, 8064743. https://doi.org/10.1155/2017/8064743.
39. Jiang, T.; Lu, Z.; Hu, X.; Zeng, L.; Ma, X.; Huang, J.; Cui, J.; Liping, T.; Zhou, C.; Yao, X.; et al. Deep Learning Multi-label Tongue Image Analysis and Its Application in a Population Undergoing Routine Medical Checkup. *Evid.-Based Complement. Altern. Med.* **2022**, *2022*, 3384209. https://doi.org/10.1155/2022/3384209.
40. Li, J.; Zhang, Z.; Zhu, X.; Zhao, Y.; Ma, Y.; Zang, J.; Li, B.; Cao, X.; Xue, C. Automatic Classification Framework of Tongue Feature Based on Convolutional Neural Networks. *Micromachines* **2022**, *13*, 501. https://doi.org/ 10.3390/mi13040501. PMID: 35457806; PMCID: PMC9025353.
41. Wang, X.; Liu, J.; Wu, C.; Liu, J.; Li, Q.; Chen, Y.; Wang, X.; Chen, X.; Pang, X.; Chang, B.; et al. Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 973–980. https://doi.org/10.1016/j.csbj.2020.04.002.
42. Meng, D.; Cao, G.; Duan, Y.; Zhu, M.; Tu, L.; Xu, D.; Xu, J. Tongue Images Classification Based on Constrained High Dispersal Network. *Evid.-Based Complement. Altern. Med.* **2017**, *2017*, 7452427. https://doi.org/10.1155/2017/7452427. PMID: 28465706; PMCID: PMC5390589.
43. Kanawong, R.; Ajayi, T.O.; Ma, T.; Xu, D.; Li, S.; Duan, Y. *Automated Tongue Feature Extraction for Zheng Classification in Traditional Chinese Medicine*; Hindawi Publications Corporation: London, UK, 2012. https://doi.org/10.1155/2012/912852.
44. Rajakumaran, S.; Sashikala, J. An Automated Tongue Color Image Analysis for Disease Diagnosis and Classification Using Deep Learning Techniques. *Eur. J. Mol. Clin. Med.* **2021**, *7*, 4779–4796.

45. Mansour, R.F.; Althobaiti, M.M.; Ashour, A.A. Internet of Things and Synergic Deep Learning Based Biomedical Tongue Color Image Analysis for Disease Diagnosis and Classification. *IEEE Access* **2021**, *9*, 94769–94779. https://doi.org/10.1109/ACCESS.2021.3094226.
46. Soma, P.; Saradha, K.R.; Jothika, S.; Dharshini, S. Tongue Diagnosis using CNN for Disease Detection. *Int. J. Electr. Electron. Res.* **2022**, *10*, 817–821.
47. Xie, J.; Jing, C.; Zhang, Z.; Xu, J.; Duan, Y.; Xu, D. Digital tongue image analyses for health assessment. *Med. Rev.* **2022**, *1*, 172–198. https://doi.org/10.1515/mr-2021-0018. PMID: 37724302; PMCID: PMC10388765.
48. Bhatnagar, V.; Bansod, P.P. Double U-Net a Deep Convolution Neural Network for Tongue Body Segmentation for Diseases Diagnosis. In Proceedings of the International Conference on Communication and Computational Technologies, Jaipur, India, 26–27 February 2022; Algorithms for Intelligent Systems; Springer: Singapore; pp. 293–303. https://doi.org/10.1007/978-981-19-3951-8_23.
49. Mayer, S.; Müller, D.; Kramer, F. Standardized Medical Image Classification across Medical Disciplines. *arXiv* **2022**, arXiv:2210.11091.
50. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Computer Vision and Pattern Recognition (cs.CV). *arXiv* **2019**, arXiv:1801.04381.
51. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2018**, arXiv:1608.06993.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.