Iraqi Journal of Intelligent Computing and Informatics (IJICI)
Vol. 4, 1, June 2025, pp. 33~56
ISSN: 2791-2868, DOI: 10.52940/ijici.v4i1.90

A Comprehensive Review of Advances in Tongue Image Classification Techniques for Diabetes Identification

Ghazwan H. Hussien[1], Zainab N. Nemer [2]
[1,2] Department of Computer Science, Basra University, Basra, Iraq

Article Info

Article history:

Received January 31, 2025
Revised March 5, 2025
Accepted March 12, 2025

Keywords:

Diabetes
Deep learning
Neural network
Tongue images

ABSTRACT

By grouping diabetics using tongue scans, therefore enabling non-invasive, economically priced, and efficient approaches to disease detection. Mostly focused on patient healthcare using medical diagnostics and early detection, research has evolved. This study issue has become more important since it supports early diabetes detection, helps clinicians and patients, and targets proactive treatments meant to reduce the condition. It helps doctors decide which important diabetes treatments to use because this metabolic disorder can damage many organs if it is not treated properly. Deep learning algorithms have made it possible to diagnose many diseases early, including diabetes, by processing and analyzing images of the tongue to classify diabetic patients. This makes it possible to combine feature extraction and pattern recognition. There are more people with diabetes around the world, and the number of new cases is also going up. People want accurate and reliable diagnostic tools, so they've made algorithms that look at pictures of tongues and get basic information from them. Using tongue photos, this paper presents a comprehensive review of current advancements in the classification and diagnosis of diabetes by focusing on developments in deep network designs, feature extraction techniques, assessment methods, and deep learning methodology. The approach uses tongue images to routinely analyze frequently used datasets and indicators for diabetes classification. It also covers the challenges faced and perhaps the routes of research to provide innovative ideas in this field.

Corresponding Author:

Ghazwan H. Hussien
Zainab N. Nemer
Department of Computer Science, Basra University, Basra, Iraq
Email: ghazwan.hani@uobasrah.edu.iq
Email: zainab.nemer@uobasrah.edu.iq

1. INTRODUCTION

One of the most deadly diseases worldwide, diabetes affects a growing daily count of people. Diabetes can be of two main types: Type I, sometimes known as insulin-dependent diabetes, results from the pancreas not producing the vital hormone insulin required for survival. Despite its prevalence in children and teenagers, this type primarily manifests in later life. With about 90% of all diabetes cases worldwide, the second type—called non-insulin-dependent—results from the body's inadequate response to insulin generated by the pancreas. Usually involving patients visiting a diagnostic center or doctor, the identification process takes extra time and costs resources to produce their diagnosis. Therefore, the accurate and quick diagnosis and analysis of diabetes is a topic deserving of research.[1] [2]

Diabetes mellitus (DM) has become a major public health issue for adults due to the rapid rise in population growth, aging, obesity, and sedentary lifestyles. According to an epidemiological analysis on the worldwide prevalence of diabetes, 171 million people had the condition worldwide in 2000; estimates of the number by 2030 point to 366 million. Higher blood glucose levels and prolonged diabetes run the risk of macrovascular or microvascular effects. Also, most people with diabetes had oral problems like gum disease, tooth loss, dry mouth, cavities, burning mouth syndrome, problems with taste and salivary glands, slow wound healing, lichen planus, geographic tongue, and candidiasis. People with diabetes mellitus will easily show buccal changes. [3].

With the main purposes being pulse assessment and tongue examination, diagnosis in Traditional Chinese Medicine (TCM) depends on four diagnostic techniques: observation, auscultation and olfaction, inquiry, and palpation. Studies of tongue pictures in Traditional Chinese Medicine (TCM) show a strong correlation with diabetes.[4].

The digital tongue tool TDA1 captures tongue images by establishing a stable light source environment. These images primarily concentrate on two components: the tongue body and the tongue layer. Distinguishing between these two elements is crucial for tongue diagnosis, as it facilitates hypotheses regarding color and texture analysis. The division and merging technique and the color threshold method were used to tell the difference between the tongue body and the tongue layer. This made it possible to get parameters for each part. [4].

With tongue diagnosis being a fundamental component of Traditional Chinese Medicine (TCM), which has a rich historical basis and is accepted as a common alternative medicine in Western countries. Through its many properties—color, texture, and shape—the tongue becomes a diagnostic instrument for disorders. Figure 1 shows the several tongue parts corresponding to the purposes of several body organs: The tip is connected with the lungs, heart, chest, and neck; the central portion relates to the liver, spleen, stomach, and pancreas; and the posterior section shows the abdominal organs. The small intestine and the colon (large intestine) make up the lower intestine, so the system must focus on a certain location of the tongue since this affected area indicates a malfunction in a given part of the body. [5].

Figure 1. Structure of tongue.

We extract characteristics from all tongue photos, primarily focusing on the middle region due to its connection to the pancreas. When doctors look for illnesses, they usually check the patient's health by looking at the shape, color, and size of the tongue, since changes in any of these factors can reliably point to a problem inside the body. In this proposed work, we develop a computer-based automated method to analyze alterations in the tongue, which will subsequently aid in identifying diabetes in patients.[5]

Deep learning algorithms and deep convolutional neural networks (CNNs) have come a long way in recent years. This has led to better classification accuracy and efficiency in image analysis technologies that use CNNs. Originally used widely in picture segmentation, image classification, and face recognition, convolutional neural networks (CNNs) have since become a key focus of study in the field of objective tongue diagnosis. CNN architectures can independently extract features, therefore removing the necessity for human feature selection—which is crucial for including intelligent tongue diagnosis systems into TCM clinical practice.[6]

This study looks into how machine learning and deep learning can be used to classify and find people with diabetes using pictures of their tongues. It also looks at how the system was made and highlights improvements in network architectures, feature extraction methods, and assessment metrics. The study talks about the problems this field is facing and suggests ways to do more research to make these systems more accurate and useful in real-life healthcare situations.

Previous research revealed strengths in the method of classifying diabetes based on tongue images.The newest machine learning and deep learning technologies were combined with a strong background in traditional Chinese medicine (TCM), specifically the use of images of the tongue to diagnose diseases. This is a non-invasive and relatively easy way to collect data that doesn't require complicated medical equipment or painful procedures for the patient. This combination opens up new possibilities for diagnosing diseases in non-invasive ways and allows for automation and improvements to the diagnostic process. The studies focused on linking a specific area of the tongue (the middle area) to diabetes based on the principles of traditional Chinese medicine (its connection to the pancreas), increasing the accuracy of diagnosis. This identification reduces the "noise" in the data and allows focusing on the features most relevant to the disease.Deep Convolutional Neural Networks (CNNs) have proven to be very effective in image analysis. They are able to automatically extract distinctive features from images (such as color, texture, and shape) without the need for significant human intervention. This means reducing reliance on human expertise in identifying important features (avoiding manual feature selection), making the process more objective and repeatable. This leads to a faster and less expensive way to diagnose diabetes compared to traditional methods (which rely on a doctor's visit and a laboratory).

Given all the aforementioned, this study tackles the vital demand for ongoing diabetes diagnosis improvement. This paper primarily focuses on enhancing the accuracy, dependability, and cost-effectiveness of a non-invasive diagnostic system for early and accurate diabetes diagnosis, using tongue pictures and deep learning approaches. While tongue image analysis is a useful substitute, current methods still need more work and improvement to meet present difficulties. Rather than suggesting a whole new solution, this review paper seeks to synthesize and assess current solutions, pinpoint best practices, and point up areas for future research.

2. RELATED WORKS

This section analyzes prior studies on the classification and early detection of diabetes through tongue imagery. This paragraph is to succinctly summarize many surveys on this topic to contextualize the current research within the existing literature.

2020 saw Thirunavukkarasu et al. look at the viability of using tongue thermography as a non-invasive early type II diabetes diagnosis tool. The study, which involved 140 subjects—70 healthy individuals and 70 diabetics—investigated the heat distribution in the tongue region using image processing techniques and machine learning approaches. The results showed that the tongue's surface temperature in diabetics was significantly higher than that of normal participants, showing a strong statistical link between glycemic haemoglobin (HbA1c) levels and thermal distribution in the tongue region ($r^2 = 0.5688$). Attaching the highest classification accuracy at 94.28%, the Convolutional Neural Network (CNN) approach The study found that tongue thermal imaging could be a useful, non-invasive approach for early type II diabetes diagnosis. [7].

Sagayaraj et al. (2021) investigated the use of tongue photos for the detection of diabetes and diabetic retinopathy using image processing and machine learning methods. The study sought diabetic patients using tongue features including color, texture, and shape. The Bi-Elliptical Deformable Contour (BEDT) technique helped to separate images and retrieve pertinent elements. A Support Vector Machine (SVM) splits images into two groups: healthy persons and those with diabetes. The findings revealed a decent classification accuracy—that is, an accuracy rate of 88.28% in diabetic identification. [8].

Li et al. (2022) investigated the distribution of tongue characteristics in diabetics by using unsupervised learning methods. The study's goal was to find out the rules that govern how the different parts of a diabetic's tongue are distributed. This was done so that traditional Chinese medicine (TCM) principles could be used to create a diagnostic base for personalized diabetes treatment. The TFDA-1 tongue diagnostic device was used to get images of the tongues of 598 patients, and the Tongue Diagnostic Analysis System (TDAS) was used to measure chromatic, tactile, and layer ratios. K-means and self-organizing maps (SOM) algorithms were employed to examine the distribution of tongue features. The findings indicated that the K-means algorithm categorized patients into three clusters, but the SOM method classified patients into four clusters. We identified statistically significant changes in chromatic and tactile attributes among groups, suggesting that these features could serve as precise indicators of patients' conditions. The research determined that employing unsupervised learning methods may effectively identify small alterations in tongue characteristics among diabetics, facilitating a more precise diagnosis and treatment. [9].

In 2022, Balasubramaniyan et al. developed a diagnostic model for diabetes employing panoramic tongue imaging techniques and deep learning. The researchers employed a convolutional neural network (CNN) featuring a ResNet-50 architecture to examine tongue attributes, encompassing color, texture, shape, and dental imprints. The data is categorized using the Deep Radial Basis Function Neural Network (RBFNN) method through automated learning. The model exhibited outstanding performance, with a diagnostic accuracy of 98.4%, exceeding that of rival models such as ResNet-34 and AlexNet. The study concluded that this non-invasive method could be an excellent tool for the early detection of diabetes [10].

The latest models employing machine learning and deep learning for diabetes detection have been analyzed (Wee et al., 2024). The study investigated the influence of redundant sampling techniques and feature selection on model effectiveness, highlighting the use of non-invasive and anthropometric measurements to improve diagnosis accuracy. The findings demonstrated that deep learning models, encompassing convolutional neural networks (CNN) and deep neural networks (DNN), outperformed traditional machine learning models like random forests (RF) and logistic regression (LR). The study revealed issues related to data availability and quality, emphasizing the need for data-driven methodologies to analyze medical factors for improved diagnostic reliability. [11].

This study, conducted by Usharani Thirunavukkarasu and colleagues and published in 2024, aimed to develop a non-invasive method for diagnosing type II diabetes by combining thermal and visual pictures of the tongue using discrete wavelet transform (DWT). Statistical features were extracted from embedded images using the GLCM algorithm, and the data was classified employing machine learning classifiers (SVM, LDA, K-NN) and deep learning models (VGG16, ResNet50). The results demonstrated that the mean-max waveform conversion base had enhanced performance, exceeding VGG16 with an accuracy of 94.37%. The results suggest that the combination of thermal and visual images may be an effective tool for the early detection of diabetes. [12].

Burcu Tiryaki and colleagues conducted research on tongue lesions using medical imaging using deep convolutional neural networks (DCNNs) and published it in 2024. Six thousand tongue images were arranged into five groups: natural tongue, covered tongue, geographic tongue, cracked tongue, and median rhombic glossitis. With a majority voting system to improve outcomes, the performance was evaluated using VGG19, ResNet50, ResNet101, and GoogLeNet networks. The results showed that whilst VGG19 achieved an accuracy of 83.93% in multi-class classification, ResNet101 achieved the best accuracy of 93.53% in binary classification (normal vs. pest. Following the majority vote strategy improved accuracy to 95.15% and 88.76%, respectively. The results imply that this approach is relevant in clinical environments for tongue lesion diagnosis. [13].

The associated works clearly show some restrictions and disadvantages. Many researchers suffer from a limited comparison between several approaches, usually concentrating on a small number of algorithms or structures. Furthermore, a regular difficulty is the interpretability of results, particularly for advanced methods such as deep learning. Certain techniques depend on hand-crafted feature extraction, which calls for knowledge, might not be best for catching intricate patterns, and can be arbitrary. Deep learning techniques—including CNNs—extract features automatically, therefore providing possible benefits in terms of power and flexibility. Different research, meanwhile, employs various kinds of data. While some techniques involve infrared imaging, others use plain, visible tongue images. Thirunavukkarasu et al. point out that thermal imaging can be sensitive to outside variables such as room temperature and recent meal or beverage intake. A few researchers, like Li et al., examine tongue feature distributions using unsupervised learning (K-means, SOM), which does not directly categorize patients but offers an understanding of feature patterns within a known diabetes group. Understanding the outcomes of unsupervised learning can prove difficult. Other constraints include the reliance on quite small datasets, the complexity brought about by combining several techniques (e.g., merging thermal and visual pictures and employing different classifiers), and the usage of panoramic tongue images—which may not always be available.

3.DISCUSSION

Analyzing a range of studies on tongue image classification for diabetes detection, this paper emphasizes the growing interest in combining Traditional Chinese Medicine (TCM) concepts with modern machine learning and deep learning approaches. From this research, a number of significant trends and observations emerge. First of all, deep learning models—especially convolutional neural networks (CNNs)—clearly are being embraced given their ability to automatically extract relevant features from complex visual data. Effective applications of architectures such as ResNet, VGG, DenseNet, and MobileNet illustrate the capacity of deep learning to outperform conventional techniques dependent on hand-crafted feature extraction. Second, while much research focuses simply on visual tongue photographs, some have looked at the benefits of using thermal imaging and suggest that multimodal techniques can improve diagnostic accuracy. Thirdly, as ensemble learning techniques (bagging, boosting, stacking, and voting) are being used more and more, model performance and robustness are improving.

Still, various constraints and difficulties exist. Small sample sizes and a lack of diversity in the datasets utilized in many of the examined research cause questions regarding the generalizability of the results. Furthermore, impeding repeatability and comparison among investigations is the reliance on particular databases, usually not publicly accessible. Moreover, even if CNNs show potential, their "black box" character makes it challenging to understand why a given categorization judgment is taken, which is essential for developing confidence and acceptability among medical practitioners. Methods based on, say, decision trees have less of the interpretability problem, but they usually compromise accuracy. Standardization of image acquisition presents still another difficulty. Classification model performance can be much influenced by changes in camera angles, picture resolution, and lighting conditions. Though referenced in many publications, the TDA1 instrument is a step towards standardization; wider acceptance and validation are still needed. Ultimately, a crucial factor for practical application is that few research studies specifically address the ethical and privacy issues related to gathering and using medical picture data. Larger, more varied, publicly accessible datasets of tongue pictures with standardized image capture techniques should take center stage in future studies. Explainable artificial intelligence (XAI) methods have more to be worked on to make deep learning models more interpretable and transparent. Another exciting path for enhancing diagnostic accuracy and offering a more complete evaluation of patient health is the combination of tongue image analysis with other clinical data (e.g., patient history, blood test findings). Before these approaches may be generally embraced in clinical practice, thorough clinical validation studies are finally indispensable to show their real-world efficacy and safety.

4. RECOGNITION ON CLASSIFICATION MODEL

Tongue scans of both diabetic and non-diabetic groups yielded the data, mostly related to color and textural characteristics. Preprocessing included independent variables derived from the feature parameters. The categorized variable was considered to be either diabetic or absent (dependent variable).Twenty percent of the specimens were set aside for testing, while eighty percent were used for training. Figure 2 depicts the model.[4] complete tongue image classification system, the different stages of the data starting from training data, through feature extraction (tongue color, tongue texture, etc.), feature processing (association processing, sample equalization, feature normalization, feature reduction), classifier training and optimization, and finally performance evaluation on independent test data. Each of these stages plays a crucial role in achieving high accuracy and reliability.

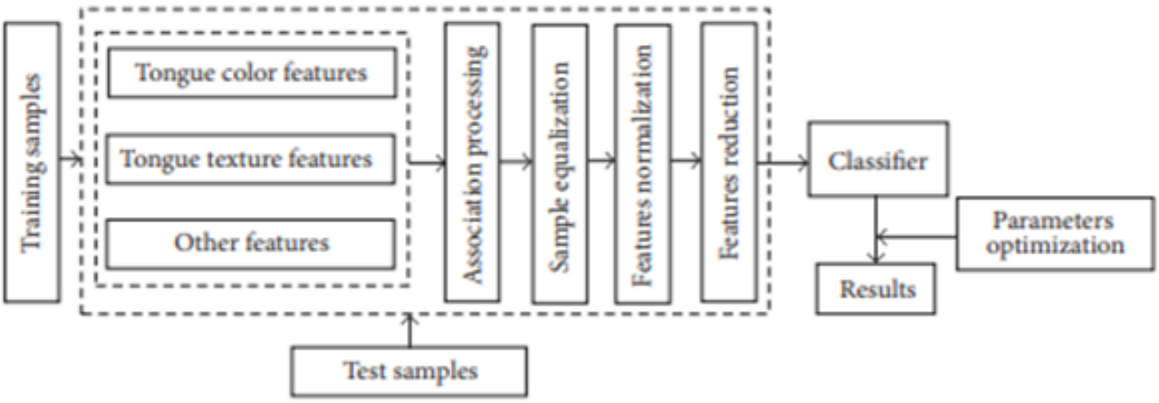

Figure 2. Map of classification model

The disparity in sample sizes affects the classification model; therefore, to equalize the samples between the diabetic and non-diabetic groups, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. The augmentation of variables would render the classification model more intricate; thus, Principal Component Analysis (PCA), a traditional method for dimensionality reduction, was employed to process the raw features obtained from tongue images, ensuring the preservation of information integrity during the dimensionality reduction procedure [4].

5. PIPELINE OF TONGUE IMAGE CLASSIFICATION

Timely detection of diabetes is crucial; however, depending just on traditional surgical methods based on blood tests is insufficient and requires the acceptance of modern non-surgical alternatives. With benefits in comfort and patient convenience, tongue image classification is a new non-invasive approach for early diabetes detection. The approaches and techniques applied in tongue image classification for several machine learning and deep learning systems are investigated in this part.

5.1. Deep Learning and Machine Learning

While deep learning is a type of machine learning that uses neural networks to emulate the human brain's learning process, in artificial intelligence (AI), machine learning refers to the ability for autonomous adaptation with minimal human intervention. These two ideas differ really significantly. Deep learning can adapt to new circumstances and correct its own mistakes even if training requires more data. On the other hand, machine learning lets one learn on smaller datasets; yet, it depends on more human engagement to learn and fix its mistakes. [14].

Based on tongue images, several well-known approaches have been suggested for the diagnosis and classification of diabetes or non-diabetic conditions, including machine learning techniques such as Gradient Boosting (GB), Support Vector Machine (SVM), AdaBoost (AB), Random Forest (RF), Decision Tree (DT), Naive Bayes, and deep learning techniques.[15]. Tongue image research Artificial intelligence (AI), machine learning (ML), and deep learning (DL) drive non-invasive diabetes screening. An area of machine learning, deep learning independently extracts and learns intricate information from medical images using several hidden layers of deep neural networks (DNNs). For the analysis of tongue pictures, where minute variations in form, color, and texture could indicate diabetes progression, these characteristics are crucial. [12].

On the other hand, developments in deep learning—particularly in relation to 2010—provided convolutional neural networks (CNNs), which greatly exceeded traditional approaches. These models independently master hierarchical properties from data. Fused thermal and visual tongue images were used to distinguish diabetic from non-diabetic patients utilizing networks including VGG16 and ResNet50. These models achieved classification accuracy of 94.37% (VGG16) and 78% (ResNet50) by combining low-level properties (e.g., edges, color) with high-level abstractions (e.g., patterns, textures), using fused images. This stands in stark contrast to the accuracy of individual modalities, such as visual images (85%) and thermal imaging (90.62%), therefore stressing the effectiveness of image fusion combined with deep learning. [12].

The application of deep learning in tongue image processing has revolutionized the diagnosis of diabetes mellitus (DM), offering a non-invasive and efficient alternative to traditional methods. Deep learning models, particularly Convolutional Neural Networks (CNNs), are particularly adept at autonomously extracting high-level features from photographs, such as texture, color, and geometry. This approach provides a more comprehensive perspective than conventional handmade methods.This method emphasizes the potential of deep learning to transform healthcare diagnostics by enabling the creation of instruments that are scalable, portable, and precise, thereby facilitating the early detection of diabetes. By integrating transfer learning, CNNs, and data augmentation, this approach creates a strong foundation for future advancements in tongue-based diagnostic systems. [15].Figure 3 shows the diabetes detection system using tongue images. The system starts by taking the patient's tongue image, then processing it and improving its quality using an adaptive median filter, then the image is segmented using a deep neural network (ResUNet++) to determine the part that contains only the tongue image, then the image is amplified by making slight changes to the image to create new images and increase the number of images, which helps in reducing the excess oversampling, then the features are extracted from the image using a convolutional neural network, and diabetes is detected using another deep network (deep residual network) used for classification.[14].

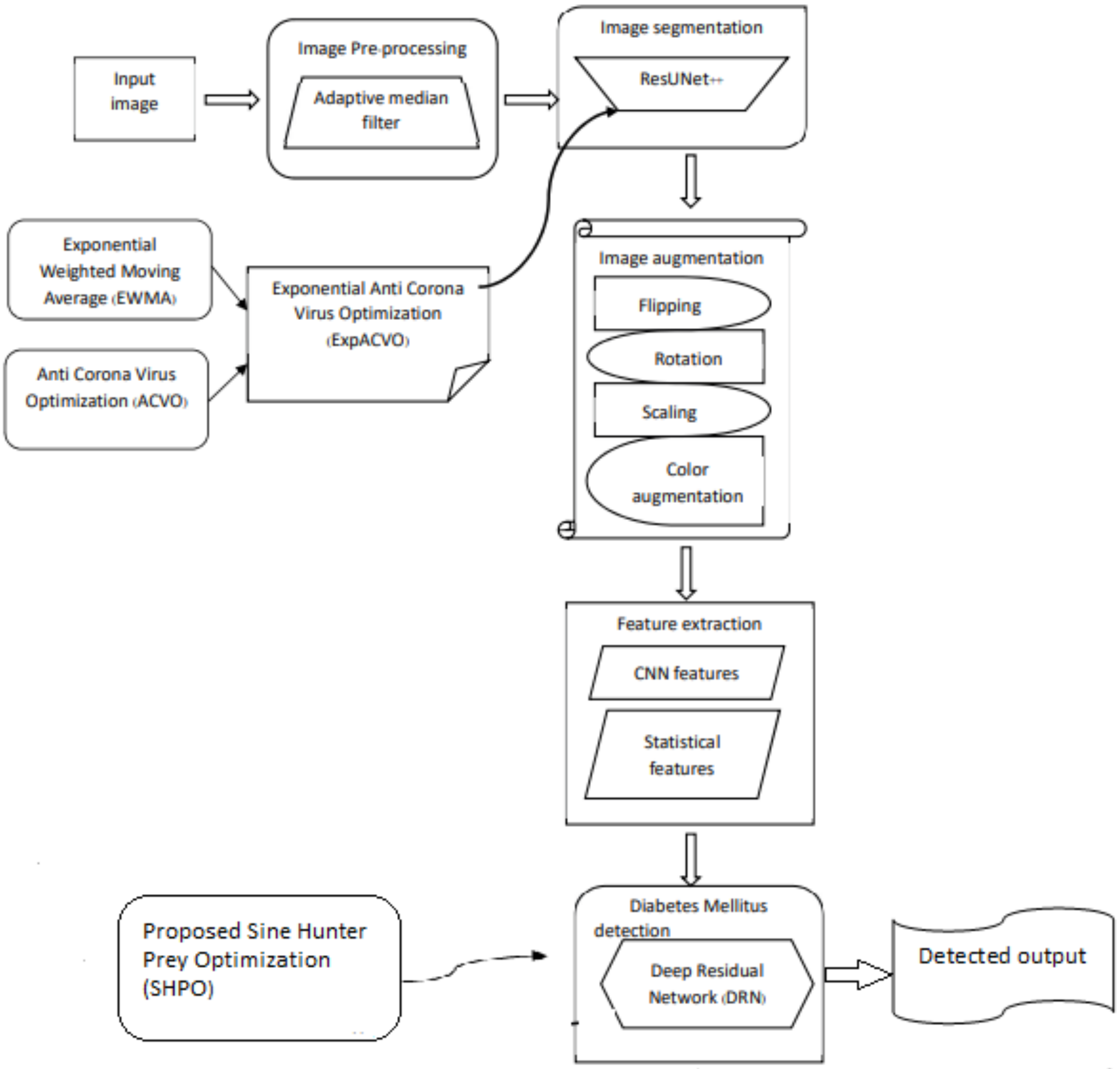

Figure 3. DM detection using tongue image

5.2. Convolutional Neural Networks

Image classification, recognition, and object detection challenges have shown remarkable performance of Convolutional Neural Networks (CNNs). Designed to allow for shifts, scales, and distortions, a standard CNN consists of input, convolutional, pooling, fully connected, and output layers. They are adept at spotting intra-class differences in texture, posture, and lighting as well as in changes in stance. Emphasizing factors including color, texture, and fissures, CNN-based models have been trained on large datasets to classify diabetes through the study of tongue images. These models fit for medical diagnosis since they show resistance to changes in tongue form and illumination. Still, their effectiveness depends much on data quality; therefore, exact results depend on high-resolution photos. CNNs are quite good at spotting complex patterns, but generalization across different datasets and model interpretability still suffer. Later studies should give top priority to improving data quality and developing explainable artificial intelligence techniques to increase their dependability in medical use.[15] [16].

Common architectures employed in tongue image classification for diabetes detection include DenseNet, GoogleNet, and MobileNet. These designs, despite utilizing relatively little datasets, effectively reduced mistakes and decreased training durations by employing efficient parameter sharing, streamlined design, and integrated graphics processing units (GPUs).

Numerous CNN-based architectures, including VGG16, VGG19, and ResNet50, are extensively employed in computer vision tasks, such as tongue image categorization for diabetes diagnosis. These models provide neural representations that efficiently extract both low- and high-level features from tongue pictures, enhancing diagnostic precision. Notwithstanding the attainment of considerable accuracy (e.g., VGG16 achieving 94.37% post-image fusion), their complexity and prolonged training duration persist as key obstacles. As tongue image analysis has advanced, the utilization of deeper neural networks such as ResNet50 has become essential. Nevertheless, augmenting the number of layers in neural networks presents challenges in training and may sometimes result in diminished accuracy due to phenomena such as overfitting or vanishing gradients. Utilizing approaches such as transfer learning and image fusion (e.g., integrating thermal and visual tongue images) alleviates these issues, enhancing performance and resilience in diabetes detection. Figure 4 shows the architecture of the modified VGG16 model. It takes a standard-sized tongue image (224 × 224 pixels) on which the model was trained and with a depth of 3 (representing the color channels: red, green, and blue—RGB), the five convolutional blocks extract features from the image with the help of filters of different sizes, each of which produces its own feature map, and the ReLU activation function helps the network learn non-linear relationships between features, then converts the outputs of the convolutional blocks (which are three-dimensional matrices) into one long vector (one-dimensional). This is necessary to prepare the data for input to the fully connected layers; then the fully connected layers learn the relationships between the features extracted in the previous stages, applying the SoftMax activation function in the last layer to convert the outputs (which are numbers) into probabilities. There is a probability for each class (in this case, "Normal" and "Diabetes"); the sum of the probabilities is equal to 1, and the class with the highest probability is the class that the network predicts as the final prediction.[12].

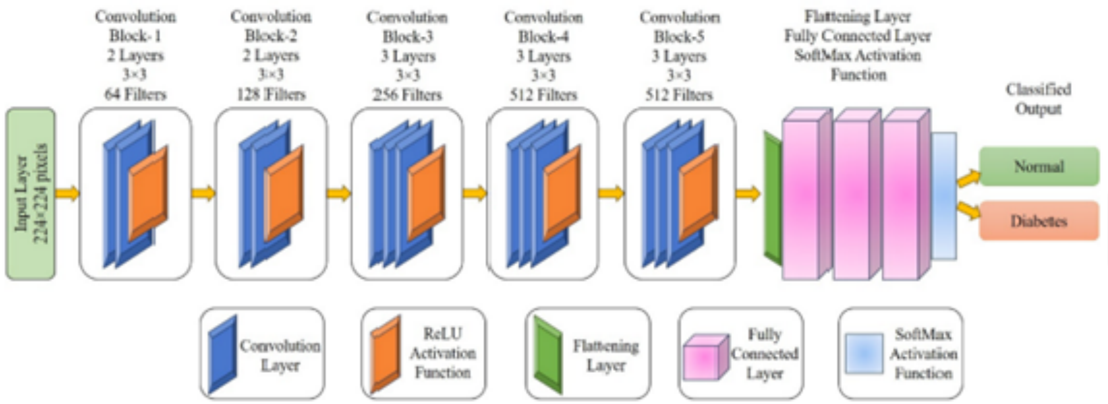

Figure 4. Pre-trained VGG-16 Net for tongue image classification.

The residual network (ResNet), developed to improve performance and accuracy through the incorporation of deeper layers, has been modified for tongue image classification to differentiate diabetic patients from healthy ones. The supplementary layers of ResNet enhance feature extraction, although they necessitate empirical assessment to confirm they do not detrimentally impact model performance. Lightweight deep neural networks, such as MobileNet, are particularly effective for tongue image processing due to their efficient computational capabilities and strong performance with optimized hyperparameters. [17]. Prominent CNN-based architectures like Inception and its variants demonstrate innovation through modular designs that enhance feature extraction in intricate images. These modules proficiently analyze texture, color, and geometric attributes vital for diagnosing diabetes. Xception, an advanced version of Inception, employs depth-wise separable convolutions, improving computational efficiency and classification accuracy for combined thermal and visual tongue data.[18]. CNN-based models' adaptability emphasizes their ability to improve non-invasive diabetes detection using tongue images.[19]. Figure 5 shows a conceptual diagram of a CNN-based system for classifying diabetic tongues by passing the tongue image through a series of convolutional layers plus an activation function (ReLU) to extract features from the image (from simple to complex), then a pooling layer to reduce the data size and make the network more resistant to minor changes, then flatten to convert matrices to a vector, then fully connected to learn the relationships between the features, then softmax to output probabilities for each class, and select the class with the highest probability.

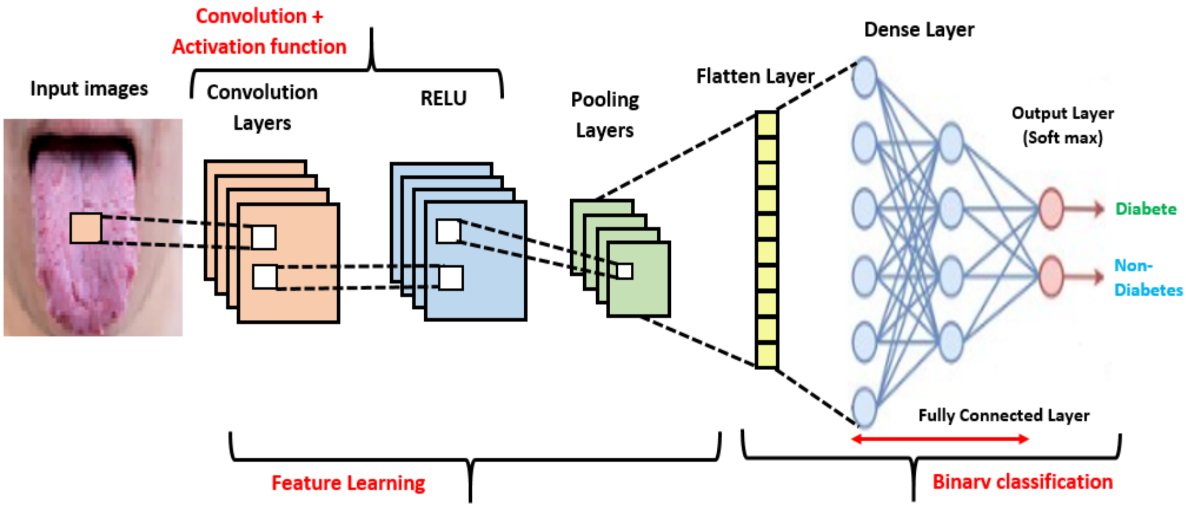

Figure 5. CNN Architecture for tongue image classification[20]

5.3. Ensemble learning

Ensemble learning is a crucial method in the field of machine learning. In recent years, ensemble learning has garnered significant attention in artificial intelligence, pattern recognition, machine learning, neural networks, and data mining. Ensemble learning has proven to be effective and practical across a broad range of problem domains and significant real-world applications. Ensemble learning generates many classifiers or a set of base learners and amalgamates their outputs to diminish overall variance. The amalgamation of many classifiers or base learners markedly enhances accuracy compared to a solitary classifier or base learner [21].

By aggregating predictions from many base models, ensemble learning techniques have achieved exceptional performance in many machine learning applications. From their inception to modern state-of-the-art algorithms, this paper offers a concise overview of ensemble learning covering the three main ensemble methods: bagging, boosting, and stacking. Previous studies have focused mostly on common ensemble techniques, including both machine learning and deep learning approaches. Machine learning, a subset of artificial intelligence (AI), has advanced thanks to the many discoveries and developments in many spheres of research, including aspects of human existence, in recent years. Machine learning can also be difficult at the same time, especially with unbalanced and high-dimensional datasets. As such, researchers regularly apply fresh and improved learning strategies, including ensemble learning. [22].

5.3.1. Ensemble learning techniques Basic

Ensemble learning is a powerful paradigm in machine learning where multiple models, often called "base learners" or "weak learners," are strategically combined to solve a given problem. The basic idea is that by combining the predictions of multiple models, the ensemble can achieve better predictive performance than any of its individual component models. This works best when the base learners are diverse (i.e., they make different types of errors). Common ensemble learning techniques include soft voting and majority voting.

5.3.1.1. Averaging (soft voting)

Using columns for the models and rows for the predictions for every data sample, soft classification arranges the computed probabilities by each ensemble model on the validation data into a matrix. For every sample, the arithmetic mean of the model predictions is calculated; the class with the highest cumulative probability is then chosen. The potential of overfitting is a main drawback. Soft voting can produce less-than-ideal performance if the underlying models are overfitted. Exhibit soft voting; Illustration 6. [23].

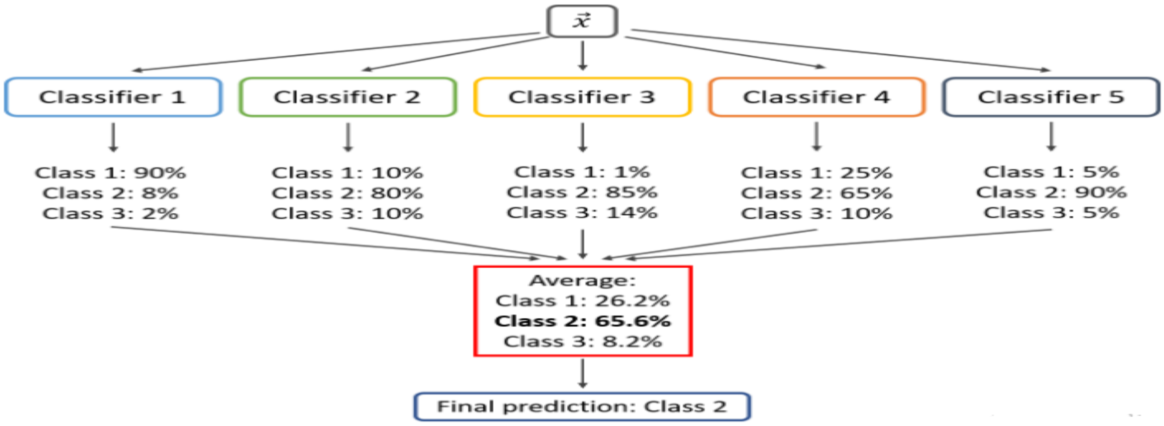

Figure 6. Show soft voting

Pseudo code (Soft Voting)

fold_predictions = []

val_predictions = predict_probabilities(model, X_val)

append(fold_predictions, val_predictions)

fold_predictions = convert_to_3D_array(fold_predictions)

average_probabilities = calculate_average(fold_predictions, across_models)

final_predictions = get_class_with_highest_average_probability(average_probabilities)

5.3.1.2. Max voting (Majority voting)

Majority voting ranks every class across all models and then determines the class with maximum prediction frequency. This yields a more realistic assessment of the model's performance. Majority voting works on a matrix whereby rows indicate models and columns represent classes or samples. Figure 7 shows overall voting.[24].

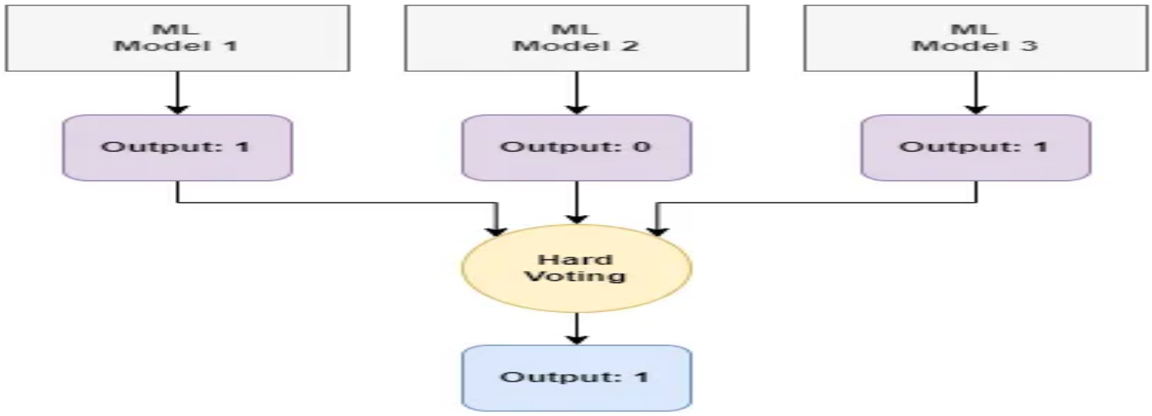

Figure 7. Show majority voting

Pseudo code (Majority Voting)

FUNCTION majority_vote(predictions):

    num_models = get_number_of_models(predictions)

    num_samples = get_number_of_samples(predictions)

    final_predictions = []

    FOR EACH sample_index FROM 0 TO num_samples - 1:

```
sample_predictions = get_predictions_for_sample(predictions, sample_index)
 most_frequent_class = find_most_frequent(sample_predictions)
  append(final_predictions, most_frequent_class)
final_predictions = convert_to_array(final_predictions)
RETURN final_predictions
```

5.3.2. Ensemble learning techniques advanced

Among advanced ensemble learning methods are bagging, boosting, and stacking. These techniques extend on the fundamental ideas of averaging and voting but provide more complex ways to generate and combine base learners.

5.3.2.1. Stacking

Stacking is the method of integrating various estimators to mitigate their biases. The forecasts from each estimator are aggregated and utilized as input for a final estimator, commonly referred to as a meta-model, which generates the ultimate prediction. The ultimate estimator is trained by cross-validation. Stacking is applicable to both regression and classification tasks. Figure 8. Illustration of the stacking technique. By training several different machine learning models (C1, C2, C3) on the training data, we use these models to make predictions (P1, P2, P3), train another model (the super-classifier) on the relationship between these predictions and the correct classification, and we use the super-classifier to make the final prediction.[25].

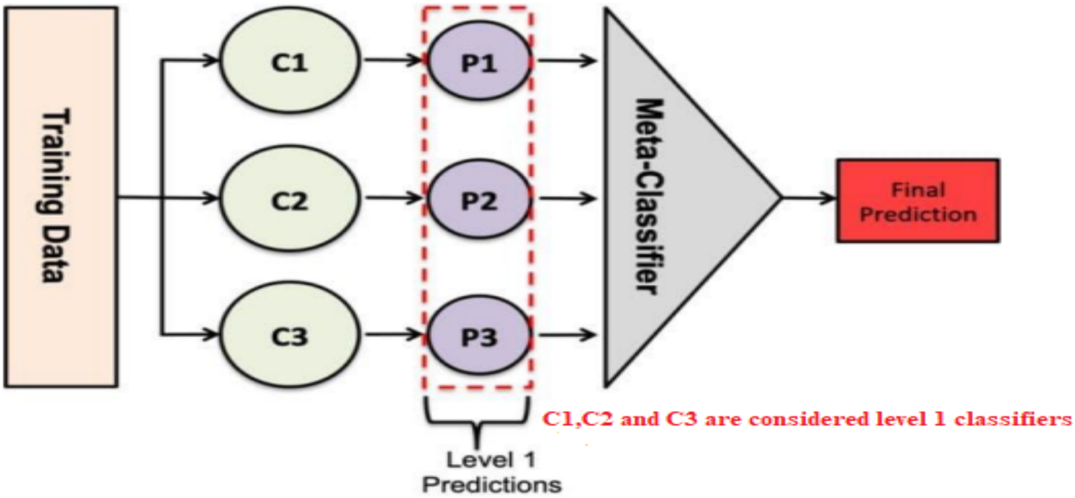

Figure 8. Show stacking technique [25].

Stacking begins with the division of data into a training set and a validation set, followed by the partitioning of the training set into K folds (for instance, 5 folds). The base model is trained on 4 folds while predictions are generated for the fifth fold, and this process is reiterated until predictions are obtained for each fold. Train the base model on the complete training set, utilize the model to generate predictions on the test set, iterate through the process from training the model on the remaining folds to predicting on the test set, and employ the test set predictions as features for a new model—the meta-model—before ultimately producing final predictions for the test set using the meta-model. In regression tasks, the inputs to the meta-model are numerical, but in classification tasks, they consist of probabilities or class labels [26].

5.3.2.2. Bagging

Obtaining random data samples, using learning algorithms, and using the mean help one to find bagging probabilities. Called aggregating bootstraps, the bagging process combines results from several samples to generate an overall output. Bagging is creating many subsets of the original dataset with replacement, building a base model for every subset, running all models concurrently, and aggregating the predictions from every model to generate the ultimate predictions.[27].

5.3.2.3. Boosting

Boosting functions by integrating a sequence of weak learners, assigning greater weight to faulty predictions in succeeding iterations and lesser weight to correct predictions. This compels the algorithm to concentrate on observations that are challenging to anticipate. The ultimate prediction derives from the majority or aggregate of the votes. Boosting can be employed to address regression and classification challenges. The "Boosting Classifier" is employed to ascertain the quantity of weak learners in the ensemble, denoted as `estimators_n. while the influence of each weak learner on the final cluster is regulated by `rate learning'. [28] [29].

5.3.2.4. Comparison between Bagging, Boosting and Stacking

Among the most important methods used in machine learning to increase model accuracy and reduce errors are ensemble learning methods. These approaches, which all rely on the idea of combining several models to improve performance relative to using a single model, include bagging, boosting, and stacking. Table 1 compares these three approaches with regard to their goals, training strategies, diversity, benefits, and drawbacks, thereby helping to clarify their features and support the choice of the most suitable way for the given work.

Table 1: Comparison between stacking, boosting and bagging [30].

Standard	Bagging	Boosting	Stacking
Objective	Reduce variance	Reduce bias	Improve overall performance
Training method	Parallel independent models	Sequential error-based models	Various models + meta model
Diversity	Depends on sample diversity	Depends on error correction	Depends on basic model diversity
Advantages	Reduces variance, easy to apply	Reduces bias, high performance	High flexibility, excellent performance
Disadvantages	Less effective at reducing bias	Prone to over fitting	More complex

5.4. Generative Adversarial Networks

Generative Adversarial Networks (GANs) may independently deduce and integrate fundamental patterns from incoming data without requiring highly annotated training datasets. A Generative Adversarial Network (GAN) comprises two neural networks: a generator and a discriminator. [31]. By means of stochastic data—more especially, random values obtained from a preset distribution—the generator generates new features. Acting as a binary classifier, the discriminator assesses the genuineness of the produced features as either real or fake.[32].

The "adversarial" feature results from the fact that in the competitive training paradigm of GANs, adversarial loss functions are maximized in a minimax game between the discriminator and generator. Two among the several medical uses of Generative Adversarial Networks (GANs) are creating lifelike medical images and improving diagnosis accuracy.[31]. Therefore, GANs can be used to improve the classification of diabetic patients by adding extra realistic photographs for training, increasing the model's ability to identify disease-related trends in tongue images.[33].

5.5. Machine learning

Machine learning classification algorithms are employed to classify a wide range of data. They facilitate the categorization of commodities into similar or distinct categories based on their characteristics. These algorithms are indispensable in the field of medical diagnostics, particularly in the context of image classification tasks. This investigation enables the classification of tongue images and evaluates their correlations with diabetes-related patterns. This exploratory study utilizes six unique conventional machine learning classification methods to accurately identify diabetes patients through tongue images.[34].

5.5.1. K-Nearest Neighbors (KNN)

The quantity of nearest neighbors to be considered is denoted by "K" in the multidimensional classification of nearest neighbors. KNN is widely employed in classification and regression problems due to its simplicity, ease of implementation, and adaptability to nonlinear data. However, it encounters challenges when dealing with large datasets, as the process of computing the distances between all data points can impede performance. KNN's accuracy may be compromised by data and high-dimensional noise. [35].

KNN is used in medicine to classify people with diseases, including diabetes and heart disease, as well as to project disease outcomes depending on symptomatology. It is quite effective in exercises on classification and regression analysis. In one specific investigation, the KNN algorithm was applied to an Alzheimer's MRI dataset for the disease diagnosis. The method showed reasonable performance with an accuracy of 45.86%, therefore highlighting its potential in medical diagnostics despite natural limitations.[36].

5.5.2. Decision Trees (DT)

A Decision Tree (DT) is a technique that employs a sequence of consecutive if-then principles to classify input into classifications or judgments. It operates by identifying the most advantageous characteristics to iteratively partition the data until categorical or predictive results are achieved. The decision tree (DT) is a widely used tool in classification and regression problems, as it has the ability to manage both numerical and categorical data and provides excellent interpretability. However, it is prone to overfitting if the tree is extremely deep, and it may demonstrate sensitivity to minor data fluctuations, which could impact the model's stability.[37].

Decision trees are flowchart-like structures that are used in medicine to make differential diagnoses based on the patient's medical history and symptoms. They are particularly skilled in identifying risk factors for conditions such as cardiovascular ailments and diabetes. The efficacy of DT models in clinical environments has been demonstrated in previous studies. The DT model was trained on clinical data in one investigation, which yielded an accuracy of 74.4%. The model's balanced accuracy was significantly improved to 91.1% as a result of the integration of medical image feature analysis. The research concluded that the integration of clinical data and medical picture elements by machine learning. [38].

5.5.3. Logistic Regression (LR)

The probability of data belonging to a specific category or a predetermined set of categories is represented by the Logistic Regression (LR) algorithm. It is particularly well-suited for binary and linear classification endeavors, as it functions primarily as a classification instrument rather than a regression model, despite its designation. LR is one of the most widely used models in practical and commercial applications due to its extraordinary efficiency and simplicity. However, its effectiveness is limited in nonlinear tasks or when the interrelations among characteristics are complex, as it struggles to capture complex patterns in these situations. [39].

In medicine, logistic regression is frequently employed to predict disease diagnoses, such as diabetes, and is valued for its ability to provide probabilities alongside classifications. This makes it highly advantageous for binary classification problems. A logistic regression model was developed and validated to differentiate between peripheral lung cancer (PLC) and solitary pulmonary tuberculosis (SP-TB) using clinical and imaging features. The model achieved a maximum AUC value of 0.878 in the internal validation cohort, demonstrating its strong effectiveness in distinguishing between the two scenarios. A further study utilized logistic regression to predict lymph node malignancy in pancreatic cancer based on ultrasound imaging features, demonstrating it to be the most effective model. Additionally, logistic regression was employed in a study to identify patients at heightened risk of kidney stones via CT-based radiography, outperforming other machine learning models and underscoring its effectiveness in medical diagnostics. [40].

5.5.4. (Support Vector Machine - SVM)

Designed for binary classification and multi-class applications, the Support Vector Machine (SVM) is a strong technique. It guarantees good generalization by finding the optimal hyperplane that divides data into discrete categories with the biggest margin. Support vector machines (SVMs) help to construct linear decision boundaries by nonlinearly translating inputs to high-dimensional feature spaces via kernel functions. This approach reduces the risk of overfitting while yet making SVMs rather effective in managing both linear and nonlinear data as well as high-dimensional datasets. Still, SVMs may show longer training times with large datasets and need careful parameter fine-tuning, including kernel type and penalty coefficient (C), to get the best performance.[33].

Support Vector Machines (SVMs) are widely utilized in the medical field for the classification of medical images to detect various conditions, including diabetes, breast cancer, lung cancer, and osteoporosis. Certain studies have employed SVMs to distinguish between the overall dimensions of the tumor and normal liver tissue in patients with hepatocellular carcinoma. In a distinct experiment, the SVM model achieved an outstanding 91.5% accuracy in classifying different types of brain tumors and distinguishing them from normal cases. Additionally, support vector machines (SVMs) were utilized to predict BRAFV600E mutations in patients with papillary thyroid cancer (PTC) using ultrasound, as part of a comparative study involving six different machine learning models, with the SVM model exhibiting superior efficacy. These applications highlight the versatility and effectiveness of SVMs in medical diagnosis and disease classification. [41].

5.5.5. Random Forest (RF)

The Random Forest (RF) algorithm is an ensemble technique that constructs several decision trees, each trained on random subsets of data and attributes. Aggregating the outcomes of these trees facilitates final classification or forecasting choices, thereby diminishing variability and improving accuracy. Random forests are user-friendly, adept at managing extensive and intricate datasets, and exhibit a reduced susceptibility to overfitting in comparison to solitary decision trees. Nevertheless, they may exhibit reduced predictive speed when processing extensive information. [31].

Random forests comprise many decision trees that deliver robust performance and provide insights on feature importance, making them highly valuable across various applications. In medical diagnostics, they are widely employed for their accuracy and ability to manage large datasets with numerous variables. In a study, six machine learning models were developed to distinguish between Epstein-Barr virus-related gastric cancer (EBVaGC) and non-EBVaGC gastric cancer using CT radiography and clinical characteristics. The Random Forest model outperformed the others, demonstrating reliable performance with high accuracy, sensitivity, and specificity in the test dataset. A distinct study analyzing plantar dynamic pressure data to detect osteoarthritis changes in the knee determined that Random Forest (RF) was the most successful model among five options, which included K-Nearest Neighbors (KNN), Support Vector Machine (SVM), AdaBoost, and XGBoost. Additionally, a study aimed at creating an AI-assisted model for diagnosing thyroid disorders revealed that the Random Forest model exhibited greater accuracy compared to eight other machine learning models, hence affirming its effectiveness in medical applications.[42].

5.5.6. Naïve Bayes (NB)

The Naive Bayes (NB) algorithm is a statistical classification technique that is based on Bayes' theorem. It facilitates the rapid and efficient processing of high-dimensional data by operating under the "naivety" assumption, which assumes that characteristics are mutually independent. This renders Naive Bayes particularly pertinent for applications such as medical classification (e.g., disease identification) and text classification (e.g., malware detection). The method is proficient in the management of missing data, is easy to implement, and is efficient in the training and prediction processes. However, its precision may be compromised if the independence assumption is violated or if the dataset is skewed.[43].

Built on Bayes' theorem, the Naive Bayes (NB) algorithm is a statistical categorization method. Operating under the "naivety" assumption—that traits are mutually independent—it allows fast and effective processing of high-dimensional data. Naive Bayes is thus particularly suitable for uses like text classification (e.g., spam detection) and medical classification (e.g., disease diagnosis). The approach is easy to apply, effective in training and prediction, and good in handling missing data. Still, its accuracy could drop if the dataset is biased or the independence condition is violated. [44].

Six unique classic machine learning classification methods are delineated, alongside additional methodologies capable of successfully identifying diabetic individuals through tongue image analysis. The precision and efficacy of these algorithms differ from one to another.

5.6. Feature extraction

Diabetic categorization algorithms using tongue images depend critically on feature extraction. Recognized as distinct traits that can point to the disease's presence are tongue pigment, scalp condition, fissures, and other signs. Still, this method is complicated by issues like tongue discoloration from food dyes or a white coating, which calls for changes to current systems to ensure exact and consistent feature representation [45].

A deep learning method used for feature extraction, convolutional networks (CNNs) have been applied to examine tongue photos and determine their correlation with either people diagnosed with diabetes or those without the ailment. These networks detect complex patterns in tongue photos by using automatic feature extraction through convolutional layers. Their training on datasets including photos of both diabetics and non-diabetics helps them to better understand subtleties such as tongue color differences or the presence of fissures. Realistic tongue images are created using generative adversarial networks (GANs), therefore improving the variety of training material. This approach generates extra images that resemble real data, therefore enhancing the model's ability to detect aberrant traits in tongue images, especially in situations with little data. [46].

One of the difficulties with feature extraction is tongue contamination since food or beverage pigments may affect its coloration, thereby complicating the extraction procedure. Furthermore, negatively affecting model performance is the lack of data resulting from small available datasets. Furthermore, the variability of features, which differs among individuals and cultural situations, needs models adept at accepting this diversity. Proposed solutions to tackle the challenges include employing deep learning methodologies such as ResNet, DenseNet, GoogleNet, and MobileNet to enhance model performance in extracting intricate features from tongue images; augmenting data diversity through techniques like GANs to generate realistic and varied tongue images; and utilizing ensemble learning strategies such as random forest, stacking, boosting, bagging, average, and majority to elevate classification accuracy by consolidating the outcomes of multiple models. Figure 9 illustrates the basic design of a CNN for feature extraction by transforming the original image into a set of feature maps that represent increasingly complex and abstract patterns.Layers of convolution pick filters to identify these trends. ReLU functions introduce nonlinearity. Combining layers lowers dimensionality and strengthens the network's resistance to small variations. Multiple consecutive layers of this process let the network learn a hierarchical representation of the image (from simple features in the first layers to sophisticated features in the deeper layers).[47].

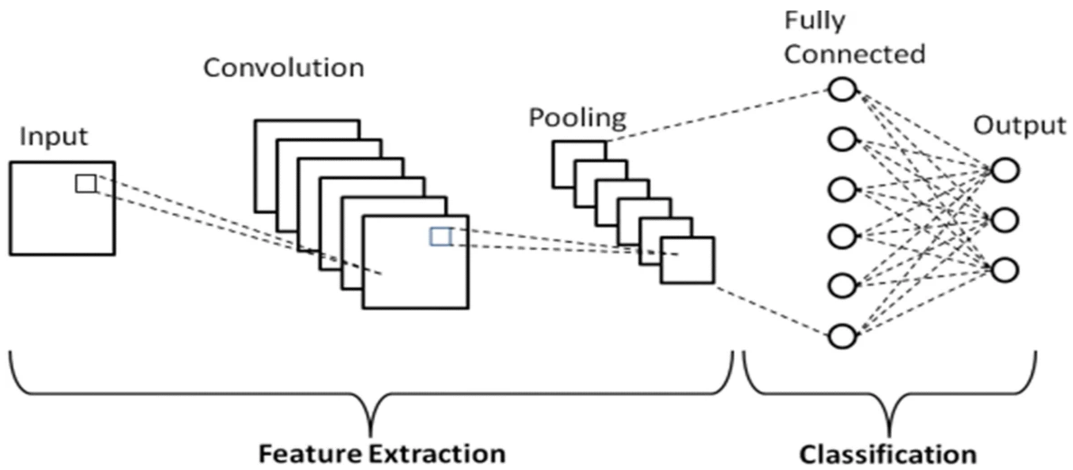

Figure 9. CNN Basic architecture for features extraction [20].

5.7. Data augmentation

One often used method to maximize the use of information gathering is data augmentation. Little changes, such as scaling, rotating, and translating, help to improve the training set. These changes generate fresh and diverse situations for training datasets, hence improving the efficacy and results of machine learning models. Deep learning and other machine learning algorithms depend on large datasets for optimum training; however, the restricted availability of datasets, particularly in emerging research areas, forces the usage of data augmentation to increase the variety of present datasets by means of modifications.[48]

The effectiveness of data augmentation is derived from fundamental modifications, including horizontal rotation, color space modification, and random cropping. These modifications encompass a multitude of variables that pose challenges for applications that rely on image identification. Sophisticated augmentation methods, in addition to fundamental approaches, include geometric modifications (such as rotation, scaling, and translation), color space transformations (including adjustments to brightness, contrast, and saturation), kernel filters (such as blurring or sharpening), and picture shuffling. Additionally, the diversity of training data is being increasingly improved through the implementation of random erasure, feature space augmentation, adversarial training, GAN-based augmentation, neural style transfer, and meta-learning strategies. [49].

Combining several augmentation techniques increases the amount of the dataset and provides several options that improve the durability and generalization of machine learning and deep learning models. This is especially crucial in disciplines such as medical imaging, where datasets are usually restricted, and the capacity to generalize across multiple circumstances is imperative for precise diagnosis and analysis. Many methods are used to improve data variety:

5.7.1. Geometric Transformations

It involves the application of elastic deformations, random inversions, and rotations to images, a method particularly effective in real-world circumstances where orientations may differ dramatically.

Understanding these technical changes lays a strong basis for the next research on data augmentation techniques; moreover, it is imperative to be able to "safely" and effectively apply these engineering improvements. Therefore, it is imperative to assess the "integrity" of the improvement to ensure that the model achieves better accuracy and efficiency in processing higher data volume. [49].

5.7.2. Cutout and Random Erasing

These techniques involve training images obscuring certain tongue areas of interest. This helps the system learn to classify tongue images even if some areas are concealed or missing by simulating variances or partial obstructions that might occur in real-world settings. Studies show that using these techniques helps tongue image categorization algorithms become more dependable and precise.[50].

5.7.3. Flipping

One finds much higher frequency in flipping the horizontal axis than in the vertical one. Among the easiest to run, this augmentation has shown success on ImageNet and CIFAR-10. This change does not maintain labels in MNIST or SVHN datasets for text recognition.[48].

5.7.4. Color space

Usually, digital picture data is expressed as a tensor with dimensions (height × width × color channels). One quite useful approach is using augmentations in the color channel space. Basic color augmentations entail separating one color channel, say Red, Green, or Blue. By separating that matrix and adding two zero matrices from the other color channels, one can quickly convert an image into its representation in one color channel. Simple matrix operations allow one to easily change the brightness of the image by means of RGB values. Advanced color augmentations are obtained from image-characterizing color histograms. Changing the intensity values in these histograms results in different lighting, much like in photo editing programs. [49].

5.7.5. Cropping

By extracting a center patch from every image, picture cropping is a good processing method for picture data with different height and width dimensions. Furthermore, arbitrary cropping might have the same effect as translations. Random cropping and translations differ in that while translations preserve image spatial dimensions, cropping reduces the input size from (256, 256) to (224, 224). The chosen lowering threshold for cropping might not be a label-preserving change.[51].

5.7.6. Rotation

Rotation augmentations rotate the image anticlockwise or clockwise around an axis between 1° and 359°.The degree of rotation parameter determines the safety of rotation augmentations in a major part. While minor rotations between 1 and 20 degrees or −1 and −20 degrees may help with digit recognition tasks like MNIST, as the degree of rotation increases, the integrity of the data label is damaged following the transformation.[52].

5.6.7. Translation

Translation of images either horizontally or vertically will help to reduce data positional bias. Should all of the images in a dataset be centered, the model also has to be tested on exactly centered images. When the original image is moved in a specific direction, the empty space could be filled with random or Gaussian noise or a constant number, say 0 or 255. This padding preserves image spatial dimensions after augmentation. [53].

5.7.8. Noise injection

The process of picture noise injection involves the incorporation of random noise, which is typically derived from a Gaussian distribution, into the images. The utilization of this technique makes it possible to replicate different illumination circumstances or to simulate inaccuracies, hence strengthening the capability of models to gain more resilient characteristics and improving their ability to effectively manage data variances. Moreno-Barea et al. conducted an evaluation of noise injection on nine datasets from the UCI collection, and their findings demonstrated that it is effective in enhancing the performance of models. The incorporation of noise improves the model's capacity to recognize patterns in surroundings that are veiled or challenging, and as a result, it serves as a crucial asset for enhancing resilience in tasks such as photo categorization. [54].

5.8. Image preprocessing

Especially in the study of tongue images, which regularly face difficulties like noise, inconsistent lighting, and resolution constraints resulting from environmental factors or equipment quality, image preprocessing is a necessary phase in improving the quality and applicability of photos. Preprocessing techniques seek to sanitize the accessible data and remove elements that can compromise performance, such as distorted, fuzzy, or incomplete images. Preprocessing is therefore very important for accurate interpretation since factors like smoke, poor lighting, or movement during image collection can compromise the quality of tongue photographs. A basic preprocessing method, grayscale equalization distributes the intensity values of an image's pixels, enhancing its contrast and brightness. The approach starts with the study of the histogram of the image, therefore showing the grayscale level distribution. While an uneven distribution may produce too bright or dark areas, therefore diminishing the visual quality of the image, a balanced histogram indicates adequate contrast and brightness. Grayscale equalization improves important characteristics in tongue image processing, including color and texture changes—qualities necessary for diagnosis. [55].

Image processing is a developing technique whereby important insights are derived from a series of operations on input images. Usually representing the three main colors—red, green, and blue (RGB), or grayscale values in monochromatic images, input images are shown as pixels. Two main approaches define image processing: analogues and digital ones. Whereas digital image processing consists of three basic phases: preprocessing, augmentation, and presentation, succeeded by knowledge extraction, analogue processing relates to tangible reproductions, such as printed photographs. Pretreatment methods in digital image processing are essential to prepare the image for the next analysis: noise reduction, grayscale normalization, and the elimination of distorted or incomplete images. Essential for medical uses, including tongue image analysis for diagnosis, these techniques improve image quality and offer more exact feature extraction and analysis. Combining these techniques helps image processing systems to effectively control changes in input data and provide consistent, high-quality results.[56].

5.9. Tongue image classification

Especially in traditional and modern medicine, tongue image categorization is regarded as one of the most interesting uses in medical image analysis. The development of more affordable and powerful computing systems has attracted increased attention in the study of tongue images for non-invasive diagnosis and health monitoring. In this discipline, digital image processing is indispensable, as it enables the extraction of critical information from tongue images to facilitate the identification and classification of illnesses. The process of tongue picture categorization involves the examination of the tongue's unique characteristics, such as its color, texture, shape, and coating, in order to identify patterns that are associated with various health conditions. The features are extracted and processed to produce a numerical representation of the tongue, which is then compared to a database of known examples for classification. The precision of tongue image analysis can be influenced by a variety of factors, such as the presence of impediments (e.g., food detritus or saliva), picture resolution, and illumination. Consequently, it is essential to implement preprocessing procedures, including noise reduction, grayscale equalization, and picture enhancement, in order to achieve the highest possible level of performance. [57].

The classification of tongue images typically involves three primary stages: the identification of the tongue region in the image, feature extraction to determine essential properties, and classification based on the identified features. In order to improve the system's resilience and accuracy, sophisticated algorithms, including deep learning models, are frequently implemented. These algorithms are capable of identifying subtle variations in tongue morphology that may indicate specific health conditions, such as infectious diseases, diabetes, digestive disorders, or systemic diseases. Tongue image categorization has a wide range of applications, such as telemedicine, which enables remote health monitoring, and personalized healthcare, which provides personalized diagnostic insights. By incorporating this technology into diagnostic instruments and electronic health records, medical evaluations can be rendered more precise and efficient. Additionally, the field will be further advanced by the implementation of standardized datasets and efficient preprocessing methods, rendering tongue image classification an essential asset in modern medicine. Figure 10 shows deep learning techniques for tongue diagnosis analysis. In Figure (a), examples of three different real tongue images (in terms of color and presence of white areas, and in terms of texture with different depths of fissures and a relatively thick white layer covering the tongue) are used as inputs for deep learning models. Figure (b) is a simple schematic of a single ResNet101 deep learning model (CNN) used to classify tongue images. This single model does all the work (feature extraction and classification). Figure (c) is a system that uses several completely different deep learning models (e.g. ResNet101, VGG16, DenseNet) together to classify tongue images. Each model classifies the image independently. The results of the three models are combined by common fusion methods (e.g. , majority voting, soft voting, and stacking) to obtain the final prediction.[58].

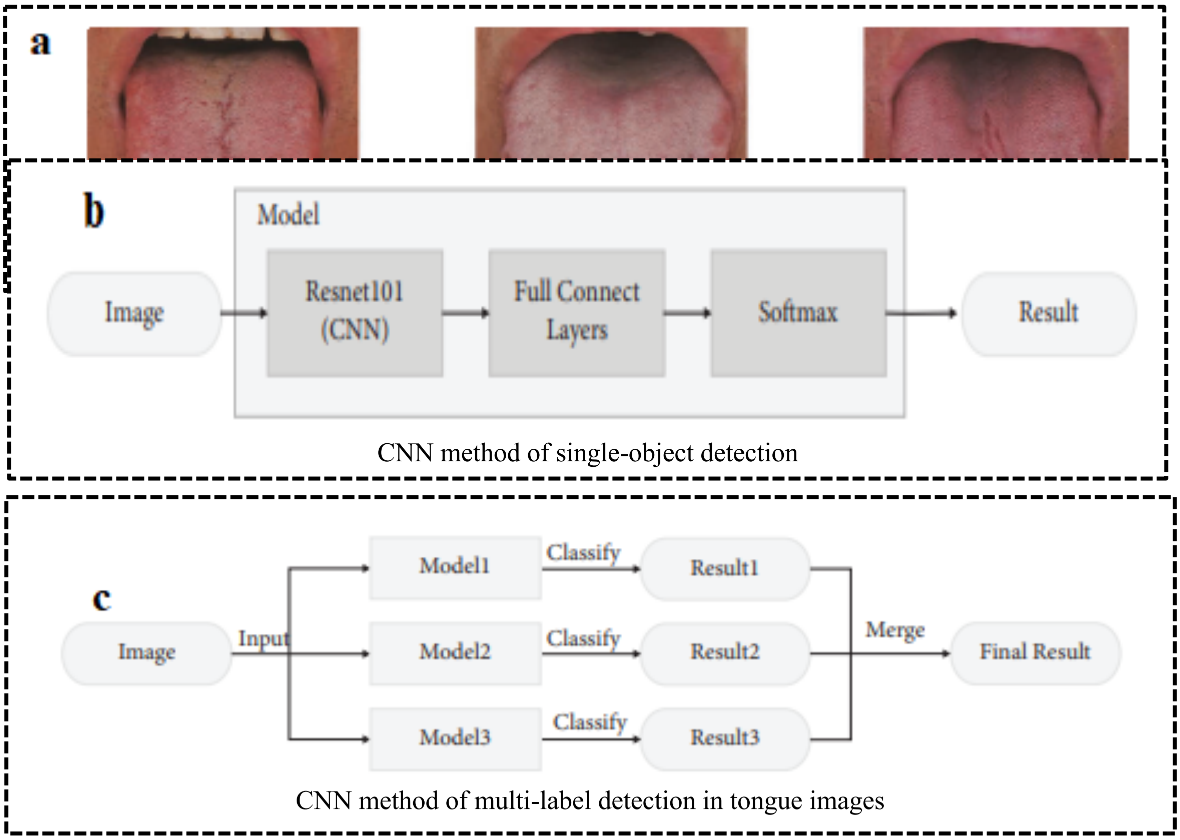

Figure 10. (a, b, c) view Deep learning methods for tongue diagnosis analyses

6. EVALUATION METRICS

Evaluating Conventional Tools for Tongue Image Classification In tongue image classification, practical application model performance measurement depends on traditional assessment instruments and metrics. Analyzing the accuracy, robustness, and efficiency of these models under several evaluation criteria and benchmarking techniques can help one to evaluate their effectiveness. The main metrics and methods used for evaluating tongue image categorization systems will be described in this part. The main evaluation gauges are: [59].

6.1. Accuracy

One of the most important evaluation measures used in many fields, including the analysis of the tongue photo ratio to the overall total count, is accuracy. High accuracy indicates, as shown in Equation (1), the model can regularly distinguish between several linguistic elements connected to different health problems.[60].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

6.2. Precision

Computes the ratio of precisely expected positive events—that is, tongue images for a certain condition—to the overall projected positive events. High accuracy means the model produces fewer false positive mistakes. As Equation (2) illustrates,[61].

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

6.3. Recall (Sensitivity)

Evaluates the model's precisely identified real positive case ratio.A high recall indicates that, as stated in Equation (3), the model ignores fewer real positive cases[62].

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \quad (3)$$

6.4. F1-Score

Considered all-around, including false positives and false negatives, the F1 score is the harmonic mean of precision and recall and is therefore a vital evaluation statistic for judging the performance of the model. Particularly helpful for imbalanced datasets and a more consistent assessment of a model's efficacy than correctness, which may ignore certain mistakes, the F1 Score provides a complete evaluation of model performance. This results from its evaluation of both mistake types. As shown in Equation (4).[63].

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

6.5. ROC (Receiver Operating Characteristic)

The ROC curve graphically shows the model's performance at certain classification thresholds. Plotting the True Positive Rate (TPR) in respect to the False Positive Rate (FPR). This image helps us to find the moment when the rates of false positives and true positives equal one another. Analyzing the ROC curve helps one assess the performance of a classification model and guide threshold choosing.[64].

6.6. AUC (Area under the Curve)

For classification problems, the AUC—also known as area under the ROC curve—offers a complete picture of a model's performance. It produces a single scalar value to evaluate the model's effectiveness. Whereas an AUC of 0.5 denotes random chance, an AUC of 1 denotes perfect classification. The AUC is widely used in performance evaluation in many different disciplines and is necessary to assess the effectiveness of categorization systems. [65].

6.7. Confusion Matrix

Comprising accurate forecasts, erroneous predictions, and false predictions, a confusion matrix summarizes the model's predictions in respect to the real labels. One can use it to identify areas of concern and determine several evaluation criteria. True Positives (TP), which are successfully expected positives; True Negatives (TN), which are correctly predicted negatives; False Positives (FP), which are incorrectly predicted positives (Type I error); and False Negatives (FN), which are incorrectly predicted negatives (Type II error).[66].

6.8. FRR (False Rejection Rate)

The False Rejection Rate (FRR) is the percentage of actual positives the model mistakenly rejects. A low false rejection rate means the model regularly ignores real positive events. As seen in Equation(5).[67].

$$\text{FRR} = \frac{FN}{FN + TP} \quad (5)$$

6.9. Specificity

Specificity, sometimes known as the true negative rate, measures the model's exact fraction of actual negatives identified. This statistic clarifies the efficiency in identifying negatives, improving the general performance and dependability of the system. High specificity is the indication that the model generates fewer false positive mistakes. As Equation (6) demonstrates.[68].

$$\text{Speci f icity} = \frac{TN}{TN + FP} \quad (6)$$

6.10. True positive rate (TPR)

The ratio of true positives (TP) accurately identified from the total of true positives (TP) and false negatives (FN) is the recall, sometimes known as the true positive rate. Equation (6) describes how one might run the computation.[68].

$$\text{TPR} = \frac{= TP}{FN + TP} \quad (7)$$

7. Challenges and Future Work

Future research has many chances to increase the present capabilities of tongue image classification algorithms to identify diseases and contribute to the future by means of numerous efforts; past research has faced many obstacles that demand addressing in future research.

7.1. Challenges

1- **Imbalanced data**: Achieving data balance is difficult since the small differences in picture amounts across various classes—e.g., healthy tongue versus tongue displaying specific diseases—are insufficient to classify the dataset as either completely balanced or significantly imbalanced. When a certain class is much underrepresented, the model may completely ignore that class, resulting in poor performance and a difficulty to consistently identify rare or uncommon diseases.

2- **Tongue image classification**:Tongue image classification datasets have to cover not only the presence of some disorders but also their degree or stage. A tongue with early-stage illness has different traits than one showing advanced disease. Misclassifying early events as advanced ones could lead to false diagnosis results, therefore reducing the effectiveness of treatment.

3- **Optimizing Model Parameters**:Hyperparameter optimization of learning rate, batch size, neural network layer, and node count directly affects the tongue image classification model's improvement. Optimal model performance is obtained by means of careful choosing of these parameters, hence enhancing the model's ability for exact and effective learning.This optimization will directly affect how well researchers understand the possibilities and limitations of the model.

4- **Detecting complex diseases:** Some diseases affecting the human body show signs on the tongue, which can be complex or vague and include subtle changes in color or texture. These changes could be difficult for systems designed mostly for the detection of common diseases. Therefore, the accuracy of the model could decrease, which would increase classification mistakes, especially in cases when the symptoms are vague.

5- **Integrating tongue classification with medical diagnostic systems**: In modern medical diagnosis systems, tongue imaging could be the first step; hence, integrating tongue categorization with other techniques, such as clinical symptom analysis or laboratory tests, can increase the accuracy and effectiveness of these systems. This method will also assist in addressing the issues related to misdiagnosis caused by symptom overlap across various diseases.

6- **Tongue image accuracy:** The precision of tongue images is contingent upon the imaging device's accuracy, illumination intensity, imaging angle, and the individual's seated posture during imaging, as well as the tongue's coloration being influenced by the dietary intake of the afflicted individual.

7.2. Future Works

1- **Expanding to discover other diseases:**Future studies could look at other diseases connected to the tongue, such as fungal infections, ulcers, and gastrointestinal and hepatic problems, which would show similar symptoms but call for different treatments.

2- **Integration of additional diagnostic methods:** Integrating other diagnostic methods, such as voice analysis to detect speech problems connected to tongue disorders or the examination of alternative medical images, helps the system's accuracy to be improved.

3- **Ethical and Privacy Considerations:**Particularly with regard to the preservation of medical data confidentiality, the ethical and privacy issues of gathering and evaluating tongue pictures must be addressed as artificial intelligence is used in medical diagnostics more and more.

4- **Early Diagnosis Alert System:** Establishing a notification system will let users or doctors be informed upon the discovery of the first signs of tongue problems, therefore promoting quick diagnosis and efficient treatment.

5- **Enhanced security and reliability:** To ensure accurate diagnosis and data protection, the system can be improved to include extra security components, such as confirming the patient's identity using alternative biometric data.

8. CONCLUSION

This paper provides a comprehensive analysis of the roles of machine learning and deep learning techniques in improving disease detection and diagnosis, specifically regarding diabetes and other tongue-related disorders. It underscores a non-invasive and effective methodology for early disease detection, highlighting improvements that address issues such as subtle symptom fluctuations, imbalanced datasets, and practical implementation in real-world contexts.Techniques for data augmentation were employed in order to improve the diversity of the data and optimize the performance of the model, which finally resulted in improved symptom detection and disease categorization. A number of cutting-edge technologies, including convolutional neural networks (CNNs), were applied in order to extract characteristics of the tongue, including its color, texture, and shape. Deep learning techniques, such as Convolutional Neural Networks (CNN) and machine learning algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression (LR), Naïve Bayes, and Random Forests, have shown efficacy in assessing tongue features like discoloration and texture, thereby improving diagnostic accuracy for diabetes and other conditions. Research indicates that non-surgical approaches can enhance the cost-effectiveness and comfort of the diagnostic process. Attaining elevated classification accuracy through the amalgamation of technologies like support vector machines (SVM) and deep neural networks. The research faced challenges such as data imbalances and discrepancies in lighting conditions and image quality, which were mitigated through pre-processing techniques like SMOTE.In the future, research should concentrate on the development of advanced deep learning and machine learning techniques, as well as the integration of tongue image analysis with traditional diagnostic tools such as blood tests.The implementation of this will make it possible to enhance the accuracy and effectiveness of the system in real time while simultaneously giving emphasis to ethical considerations and safeguarding patient data. In spite of the progress that has been made, difficulties such as symptom overlap and practical reliability still exist. It is necessary to find solutions to both specific and ethical problems in order to guarantee the safe handling of data. With the incorporation of deep learning These instruments make it possible to detect diseases, such as diabetes, at an earlier stage by the thorough collection of data and the observation of tongue symptoms.

REFERENCES

[1] M. Bharathi, D. Prasad, T. Venkatakrishnamoorthy, and M. Dharani, “Diabetes Diagnostic Method based on Tongue Image Classification Using Machine Learning Algorithms,” *J. Pharm. Negat. Results*, vol. 13, no. 4, pp. 1247–1250, 2022, doi: 10.47750/pnr.2022.13.04.174.

[2] L. T*, G. P, S. V, T. P, and S. M, “Detection of Diabetes Mellitus using Tongue images,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 4, pp. 3475–3482, 2019, doi: 10.35940/ijrte.c6409.118419.

[3] P. C. Hsu *et al.*, “The tongue features associated with type 2 diabetes mellitus,” *Med. (United States)*, vol. 98, no. 19, pp. 1–6, 2019, doi: 10.1097/MD.0000000000015567.

[4] J. Zhang *et al.*, “Diagnostic Method of Diabetes Based on Support Vector Machine and Tongue Images,” *Biomed Res. Int.*, vol. 2017, 2017, doi: 10.1155/2017/7961494.

[5] B. Gabhale, M. Shinde, A. Kamble, and M. Kulloli, “Tongue Image Analysis with Color and Gist Features for Diabetes Diagnosis.,” *Int. Res. J. Eng. Technol.*, vol. 4, no. 4, pp. 523–526, 2017, [Online]. Available: https://www.irjet.net/archives/V4/i4/IRJET-V4I4104.pdf

[6] T. Jiang *et al.*, “Tongue image quality assessment based on a deep convolutional neural network,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–14, 2021, doi: 10.1186/s12911-021-01508-8.

[7] U. Thirunavukkarasu, S. Umapathy, P. T. Krishnan, and K. Janardanan, “Human Tongue Thermography Could Be a Prognostic Tool for Prescreening the Type II Diabetes Mellitus,” *Evidence-based Complement. Altern. Med.*, vol. 2020, 2020, doi: 10.1155/2020/3186208.

[8] A. Stephen Sagayaraj, S. K. Kabilesh, A. Anand Kumar, S. Gokulnath, T. Mani, and K. Dinakaran, “Diabetes Mellitus and Diabetic Retinopathy Detection using Tongue Images,” *J. Phys. Conf. Ser.*, vol. 1831, no. 1, 2021, doi: 10.1088/1742-6596/1831/1/012028.

[9] J. Li *et al.*, “Research of the Distribution of Tongue Features of Diabetic Population Based on Unsupervised Learning Technology,” *Evidence-based Complement. Altern. Med.*, vol. 2022, 2022, doi: 10.1155/2022/7684714.

[10] S. Balasubramaniyan, V. Jeyakumar, and D. S. Nachimuthu, “Panoramic tongue imaging and deep convolutional machine learning model for diabetes diagnosis in humans,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–18, 2022, doi: 10.1038/s41598-021-03879-4.

[11] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, “Diabetes detection based on machine learning and deep learning approaches,” *Multimed. Tools Appl.*, vol. 83, no. 8, pp. 24153–24185, 2024, doi: 10.1007/s11042-023-16407-5.

[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," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-64150-0.

[13] B. Tiryaki, K. Torenek-Agirman, O. Miloglu, B. Korkmaz, İ. Y. Ozbek, and E. A. Oral, "Artificial intelligence in tongue diagnosis: classification of tongue lesions and normal tongue images using deep convolutional neural network," *BMC Med. Imaging*, vol. 24, no. 1, pp. 1–9, 2024, doi: 10.1186/s12880-024-01234-3.

[14] J. K. Mathew and S. Sathyalakshmi, "Sine hunter prey optimization enabled deep residual network for diabetes mellitus detection using tongue image," *J. Assoc. Med. Sci.*, vol. 57, no. 2, pp. 76–85, 2024, doi: 10.12982/JAMS.2024.029.

[15] L. Wu, X. Luo, and Y. Xu, "Using convolutional neural network for diabetes mellitus diagnosis based on tongue images," *J. Eng.*, vol. 2020, no. 13, pp. 635–638, 2020, doi: 10.1049/joe.2019.1151.

[16] M. M. Taye, "Understanding of Machine Learning with Deep Learning :," *Comput. MDPI*, vol. 12, no. 91, pp. 1–26, 2023.

[17] J. Li *et al.*, "Establishment of noninvasive diabetes risk prediction model based on tongue features and machine learning techniques," *Int. J. Med. Inform.*, vol. 149, no. May, 2021, doi: 10.1016/j.ijmedinf.2021.104429.

[18] J. Li *et al.*, "A multi-step approach for tongue image classification in patients with diabetes," *Comput. Biol. Med.*, vol. 149, no. October, 2022, doi: 10.1016/j.compbiomed.2022.105935.

[19] S. Fan *et al.*, "Machine learning algorithms in classifying TCM tongue features in diabetes mellitus and symptoms of gastric disease," *Eur. J. Integr. Med.*, vol. 43, 2021, doi: 10.1016/j.eujim.2021.101288.

[20] V. H. Phung and E. J. Rhee, "A High-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets," *Appl. Sci.*, vol. 9, no. 21, 2019, doi: 10.3390/app9214500.

[21] N. Thomas Rincy and R. Gupta, "Ensemble learning techniques and its efficiency in machine learning: A survey," *2nd Int. Conf. Data, Eng. Appl. IDEA 2020*, 2020, doi: 10.1109/IDEA49133.2020.9170675.

[22] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, no. September, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[23] M. Shah, H. Kantawala, K. Gandhi, R. Patel, K. A. Patel, and A. Kothari, "Theoretical Evaluation of Ensemble Machine Learning Techniques," *Proc. - 5th Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2023*, pp. 829–837, 2023, doi: 10.1109/ICSSIT55814.2023.10061139.

[24] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, "Ensemble deep learning in bioinformatics," *Nat. Mach. Intell.*, vol. 2, no. 9, pp. 500–508, 2020, doi: 10.1038/s42256-020-0217-y.

[25] A. Kumar and M. Jain, "Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases," *Ensemble Learn. AI Dev. Learn Bagging, Stacking, Boost. Methods with Use Cases*, no. January, pp. 1–136, 2020, doi: 10.1007/978-1-4842-5940-5.

[26] D. W. Opitz and R. F. MacLin, "An empirical evaluation of bagging and boosting for artificial neural networks," *IEEE Int. Conf. Neural Networks - Conf. Proc.*, vol. 3, pp. 1401–1405, 1997, doi: 10.1109/ICNN.1997.613999.

[27] P. Bühlmann, "Bagging, boosting and ensemble methods," *Handb. Comput. Stat. Concepts Methods Second Ed.*, vol. 6, no. December, pp. 985–1022, 2012, doi: 10.1007/978-3-642-21551-3__33.

[28] U. Sarmah, P. Borah, and D. K. Bhattacharyya, "Ensemble Learning Methods: An Empirical Study," *SN Comput. Sci.*, vol. 5, no. 7, 2024, doi: 10.1007/s42979-024-03252-y.

[29] T. Tong and Z. Li, "Predicting learning achievement using ensemble learning with result explanation," *PLoS One*, vol. 20, no. 1, pp. 1–25, 2025, doi: 10.1371/journal.pone.0312124.

[30] R. O. Odegua, "An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking) Rising Odegua Nossa Data An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking)," *Proc. Conf. Deep Learn*, no. January, 2019, [Online]. Available: https://www.researchgate.net/publication/338681864

[31] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3313–3332, 2023, doi: 10.1109/TKDE.2021.3130191.

[32] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An Overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[33] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent Progress on Generative Adversarial Networks (GANs): A Survey," *IEEE Access*, vol. 7, pp. 36322–36333, 2019, doi:

10.1109/ACCESS.2019.2905015.

[34] L. Malviya, S. Mal, P. Lalwani, and J. S. Chadha, “Diabetes Classification Using Machine Learning and Deep Learning Models,” *Lect. Notes Electr. Eng.*, vol. 796, pp. 487–503, 2021, doi: 10.1007/978-981-16-5078-9_40.

[35] D. V Mankar and P. S. Chaudhary, “Tongue Image Diagnosis System using Machine Learning with Hand-Crafted Features,” vol. 7588, no. 6, 2024, doi: 10.54105/ijpmh.L1097.04060924.

[36] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[37] C. Anusha, C. Anusha, and C. Engineering, “A Machine Learning Approach for Prediction of Diabetes Mellitus,” *Int. J. Emerg. Trends Eng. Res.*, vol. 11, no. 6, pp. 207–213, 2023, doi: 10.30534/ijeter/2023/031162023.

[38] K. Kangra and J. Singh, “Comparative analysis of predictive machine learning algorithms for diabetes mellitus,” *Bull. Electr. Eng. Informatics*, vol. 12, no. 3, pp. 1728–1737, 2023, doi: 10.11591/eei.v12i3.4412.

[39] M. Banday, S. Zafar, P. Agarwal, and M. Afshar, “Utilising Machine Learning Algorithms for Predictive Analysis of Diabetes,” 2021.

[40] R. D. Joshi and C. K. Dhakal, “Predicting type 2 diabetes using logistic regression and machine learning approaches,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 14, 2021, doi: 10.3390/ijerph18147346.

[41] N. A. Farooqui, . R., and A. Tyagi, “Prediction Model for Diabetes Mellitus Using Machine Learning Techniques,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 3, pp. 292–296, 2018, doi: 10.26438/ijcse/v6i3.292296.

[42] S. Jagadeesan, A. Kapoor, and S. Ghosh, “Diabetes Prediction Using Machine Learning,” *AIP Conf. Proc.*, vol. 3075, no. 1, pp. 294–305, 2024, doi: 10.1063/5.0217181.

[43] M. Chaubey, “Diabetes Mellitus Prediction using Machine Learning,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 4786–4790, 2023, doi: 10.22214/ijraset.2023.52755.

[44] P. Rani, R. Lamba, R. K. Sachdeva, P. Bathla, and A. N. Aledaily, “Diabetes Prediction Using Machine Learning Classification Algorithms,” *Int. Conf. Smart Comput. Appl. ICSCA 2023*, pp. 1664–1669, 2023, doi: 10.1109/ICSCA57840.2023.10087827.

[45] Z. Loukil, Q. K. A. Mirza, W. Sayers, and I. Awan, “A Deep Learning based Scalable and Adaptive Feature Extraction Framework for Medical Images,” *Inf. Syst. Front.*, vol. 26, no. 4, pp. 1279–1305, 2024, doi: 10.1007/s10796-023-10391-9.

[46] S. M. Vijithananda *et al.*, “Feature extraction from MRI ADC images for brain tumor classification using machine learning techniques,” *Biomed. Eng. Online*, vol. 21, no. 1, pp. 1–21, 2022, doi: 10.1186/s12938-022-01022-6.

[47] D. Rastogi *et al.*, “Deep learning-integrated MRI brain tumor analysis : feature extraction , segmentation , and Survival Prediction using Replicator and volumetric networks,” 2025.

[48] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation strategies from data,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, no. Section 3, pp. 113–123, 2019, doi: 10.1109/CVPR.2019.00020.

[49] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[50] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, “Label Refinery: Improving ImageNet Classification through Label Progression,” no. 1, 2018, [Online]. Available: http://arxiv.org/abs/1805.02641

[51] E. Randellini, L. Rigutini, and C. Saccà, “Data Augmentation Techniques and Transfer Learning Approaches Applied to Facial Expressions Recognition Systems,” *Int. J. Artif. Intell. Appl.*, vol. 13, no. 1, pp. 55–72, 2022, doi: 10.5121/ijaia.2022.13104.

[52] S. Porcu, A. Floris, and L. Atzori, “Evaluation of data augmentation techniques for facial expression recognition systems,” *Electron.*, vol. 9, no. 11, pp. 1–12, 2020, doi: 10.3390/electronics9111892.

[53] K. G. Kim and B. T. Lee, “Graph structure based data augmentation method,” *Biomed. Eng. Lett.*, 2024, doi: 10.1007/s13534-024-00446-4.

[54] J. Sheng, J. Fan, P. Ye, and J. Cao, “JNDMix: Jnd-Based Data Augmentation for No-Reference Image Quality Assessment,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2023, doi: 10.1109/ICASSP49357.2023.10096234.

[55] Y. Sun, “High-resolution image processing and entity recognition algorithm based on artificial intelligence,” *J. Intell. Syst.*, vol. 33, no. 1, 2024, doi: 10.1515/jisys-2023-0245.

[56] H. K. Gutlapalli, “Research Opportunities and Challenges in Image Processing,” *Int. Trans. Electr. Eng. Comput. Sci.*, vol. 2, no. 3, pp. 102–111, 2023, doi: 10.62760/iteecs.2.3.2023.57.

[57] Q. Liu *et al.*, “A survey of artificial intelligence in tongue image for disease diagnosis and syndrome differentiation,” *Digit. Heal.*, vol. 9, 2023, doi: 10.1177/20552076231191044.

[58] T. Jiang *et al.*, “Deep Learning Multi-label Tongue Image Analysis and Its Application in a Population Undergoing Routine Medical Checkup,” *Evidence-based Complement. Altern. Med.*, vol. 2022, 2022, doi: 10.1155/2022/3384209.

[59] N. Sulayman, “Predicting Type 2 Diabetes Mellitus using Machine Learning Algorithms Digital Image Processing View project,” vol. 44, no. 2022, pp. 89–100, [Online]. Available: https://www.researchgate.net/publication/366634353

[60] M. F. Faruque, Asaduzzaman, and I. H. Sarker, “Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus,” *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019*, 2019, doi: 10.1109/ECACE.2019.8679365.

[61] M. T. Islam, M. Raihan, F. Farzana, M. G. M. Raju, and M. B. Hossain, “An Empirical Study on Diabetes Mellitus Prediction for Typical and Non-Typical Cases using Machine Learning Approaches,” *2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019*, 2019, doi: 10.1109/ICCCNT45670.2019.8944528.

[62] A. Shitole, A. Kenchappagol, R. Jangle, Y. Shinde, and A. S. Chadha, “Enhancing Retinal Scan Classification: A Comparative Study of Transfer Learning and Ensemble Techniques,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 7s, pp. 520–528, 2023, doi: 10.17762/ijritcc.v11i7s.7031.

[63] Z. N. Abed and A. M. Al-Bakry, “Diagnose Eyes Diseases Using Various Features Extraction Approaches and Machine Learning Algorithms,” *Iraqi J. Comput. Informatics*, 2023.

[64] J. S. Akosa, “Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data,” *SAS Glob. Forum*, vol. 942, pp. 1–12, 2017, [Online]. Available: https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf

[65] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, and J. M. Moguerza, “General Performance Score for classification problems,” *Appl. Intell.*, vol. 52, no. 10, pp. 12049–12063, 2022, doi: 10.1007/s10489-021-03041-7.

[66] J. N. de Alencar, M. H. de J. Oliveira, M. C. N. Sampaio, M. F. Rego, and R. Nunes, “A Journey Through Philosophy and Medicine: From Aristotle to Evidence-Based Decisions,” *Philosophies*, vol. 9, no. 6, pp. 1–15, 2024, doi: 10.3390/philosophies9060189.

[67] M. H. Murray and J. D. Blume, “False Discovery Rate Computation: Illustrations and Modifications,” pp. 1–18, 2020, [Online]. Available: http://arxiv.org/abs/2010.04680

[68] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, E. A. Chavez-Urbiola, and J. A. Romero-Gonzalez, “Loss Functions and Metrics in Deep Learning,” 2023, [Online]. Available: http://arxiv.org/abs/2307.02694