Skin Diseases Detection and Diagnosis Support System Using YOLOv12s and Late Fusion Technique

Sahil Muneer, Aleena Azam, Yasir Arfat Malkani, Noor e Hira
Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan
***Correspondence**: sahilmuneersariyo@gmail.com, aleenaazamchandio12@gmail.com, yasir.malkani@usindh.edu.pk, noorehira.noorani@gmail.com

Citation | Muneer. S, Azam. A, Malkani. Y. A, Hira. N. E, "Skin Diseases Detection and Diagnosis Support System Using YOLOv12s and Late Fusion Technique", IJIST, Vol. 7 Issue. 10 pp 55-67, November 2025

Received | October 31, 2025 **Revised |** November 22, 2025 **Accepted |** November 25, 2025 **Published |** November 28 2025.

Millions of people worldwide suffer from dermatological conditions, but in areas with limited resources, such as Pakistan, access to prompt diagnosis is still restricted. In order to increase diagnostic reliability, this study suggests a multimodal framework for the detection of skin diseases that integrates patient-reported symptoms with visual lesion analysis. Using a late fusion approach that combines a Logistic Regression classifier trained on structured symptom features with a YOLOv12s-based object detection model for lesion localization, the system targets 16 common skin conditions. While the symptom model encodes clinical indicators such as redness, itching, and pustules, the visual model captures discriminative lesion patterns. To handle visually ambiguous situations that are challenging for Image-only approaches, outputs from both modalities were combined at the decision level. Experiments on a multimodal dataset show that the proposed fusion framework outperforms unimodal baselines in terms of accuracy and F1-score. The robustness of the proposed method under realistic dataset conditions is demonstrated through a comparison with previous research. The findings indicate that multimodal late fusion improves the performance of skin disease screening, making it suitable for tele dermatology and initial clinical decision support applications in low-resource environments.

Keywords: Clinical Decision Support, YOLOv12s, Multimodal Learning, Late Fusion, Skin Disease Detection.

Introduction:

An estimated 1.9 billion people worldwide are affected by skin-related illnesses [1]. These conditions constitute the largest group of diseases diagnosed by dermatologists [1][2]. The ability to make an immediate and accurate diagnosis is critical for the effective management of these types of conditions. Patients with these conditions may experience complications due to disease progression, increased treatment costs, and emotional trauma [2][3] if not diagnosed accurately and promptly. For this reason, it is extremely important to be able to communicate accurate and timely diagnoses of dermatological conditions in regions with limited resources.

Dermatologists provide their expertise through clinical evaluations to diagnose skin conditions. Due to the shortage of dermatologists, large numbers of patients present at healthcare facilities. Additionally, because visual assessment is subjective, the potential for error in clinical examinations increases.

The dermatologist-to-patient ratio in Pakistan is 1:460,000, compared to 1:25,000 in developed countries [4]. Patients may have to wait one to twelve weeks for an appointment at a public hospital, during which time treatable skin conditions can worsen significantly. Recent advancements in computer vision and AI technologies have significantly improved the speed of medical image analysis [5]. As illustrated in Table 1, this significant disparity results in waiting times of 1 to 12 weeks for clinical appointments in Pakistan.

Table 1. Comparison of Dermatology Healthcare Accessibility: Pakistan vs. Developed Nations

Metric	Pakistan	Developed Nations (UK/US)	Gap / Impact
Dermatologist to patient ratio [4]	1: 460,000	1: 25,000	**18× shortage**
Registered Dermatologists [4]	~1000	~12,000+ (US)	**Severe undersupply**
Concentration of specialist (Urban) [4]	80% in cities	Equally shared	**Rural areas underserved**
Population with Skin Conditions [1]	60% (~138M)	15–20%	**High disease burden**
Wait Time (Public Hospitals) [Estimation based on Market Survey]	1–12 weeks	1–2 weeks	**Longer delays to diagnosis**
Misdiagnosis Rate (GPs) [6][7]	40%	10–15%	**Higher errors**
Private Consultation Cost [Market Survey]	Rs. 1,500 to Rs. 10,000	Mostly insured	**Unaffordable for the majority**

Note: Bold values indicate major disparities between Pakistan and developed nations.

Deep learning approaches can save time and assist specialists by accelerating patient diagnosis [6][7]. However, current automated systems typically rely on a single modality (image analysis), and therefore have limitations in analyzing visually ambiguous conditions or incorporating patient-reported symptoms, which contain critical diagnostic information.

This research focuses on developing an AI-based system for skin condition detection and diagnosis, combining image-based deep learning with symptom-based machine learning. For visually observable skin conditions, the system employs YOLOv12s (You Only Look Once version 12 small), an experimental attention-based architecture adapted for real-time object detection [8], while Logistic Regression is applied to analyze structured patient symptom data.. Predictions from both modalities are later combined to improve diagnostic certainty by integrating visual and symptom-based evidence.

The system uses three different dashboards to provide administrators with oversight, doctors with tools to treat patients effectively, and Patients with the ability to monitor their progress, and its cost is roughly 1% of a traditional consultation. This paper provides a comprehensive overview of a combined image- and symptom-based detection system for various dermatological conditions, including wrinkles, acne, hyperpigmentation, rosacea, and skin cancer, evaluated using precision, recall, F1-score, and mean average precision (mAP).

The dermatological analysis of images has been one of the most popular research areas in AI medicine. Most of the initial studies utilized traditional techniques, such as image processing and machine learning algorithms (Support Vector Machine, K-Nearest Neighbor, and Random Forests), to classify skin lesions using manually extracted feature sets produced by experts, i.e., color, shape, and texture. Although these early studies achieved moderate accuracy for classifying standardized lesions, reliance on manually extracted features and researcher expertise limited their applicability across diverse skin tones and lesion types.

The introduction of deep learning has transformed the practice of dermatological diagnosis. The most widely used Convolutional Neural Network (CNN) models in dermatology include AlexNet, VGGNet, and ResNet, all of which have demonstrated high precision in identifying and classifying skin lesions. For instance, a recent study by [6] demonstrated that human-computer collaboration significantly improves diagnostic accuracy compared to clinicians working alone, demonstrating a high level of confidence in detecting skin cancers in advanced dermatological diagnosis systems. Similarly, [7] developed a deep learning system for the differential diagnosis of 26 common skin diseases with dermatologist-level accuracy using their deep neural network. Based on the findings of these studies, CNNs have now become the foundation of automated dermatology systems.

However, existing automated systems predominantly rely on single-modality approaches using image analysis alone, which struggle with visually ambiguous conditions and cannot incorporate patient-reported symptoms that provide crucial diagnostic information, [9][10]. Compared in Table 2, these image-only models cannot integrate non-visual attributes, which limits diagnostic accuracy, particularly for visually similar conditions such as acne vs. rosacea or benign vs. malignant pigmentation. The inability of Image-only models to integrate non-visual attributes limits diagnostic accuracy, particularly for visually similar conditions such as acne vs. Rosacea or benign vs. malignant pigmentation. Additionally, deep learning generally requires large labeled datasets and high computational resources, making it challenging to scale in resource-constrained healthcare systems such as Pakistan.

To address these challenges, researchers have recently explored multimodal learning and fusion-based frameworks. Several researchers have merged image features with clinical details—such as age, sex, and the site of the lesion to improve sorting accuracy [10]. Other works have applied early, intermediate, or late fusion strategies to merge heterogeneous data sources, achieving measurable gains in accuracy and consistency [11][12]. Nevertheless, only a limited number of studies have combined symptom data with complex image patterns, which together provide complementary diagnostic information. Additionally, most of the tools developed up until now do not allow for instantaneous usage and custom view capacities for physicians, making it impractical to deploy such a solution into daily practice.

Table 2. Comparative Analysis of AI-Based Dermatological Systems

Approach & Modality	Dataset & Classes	Performance & Limitations
CNN-based Image Classification [5]	Large web-sourced dermatology images	Varies (80-90%) Acc; noisy labels
CNN-Human Collaboration (Image-only) [6]	129,450 images, 2 classes (benign vs malignant)	91% Acc; limited to skin cancer, binary classification
Differential Diagnosis DLS (Image-only) [7]	50k+ images, 2 classes	89% Acc; single disease
Deep CNNs (ResNet)	10,015 demoscopic images,	88-93% Acc; /AUC > 0.90; rare

(Image-only) [7]	7 classes	class misclassification
Fine-tuned YOLOv3 (Image-only) [12]	ISIC 2018 dermoscopic images, multi-class lesion localization	96% mAP; evaluated for lesion localization, Comprehensive survey of YOLO models from v1 to v8.
YOLOv5 + ResNet50 (Image-only) [13]	10k+ images (HAM10000, multiple lesion types)	98.3% mAP@0.5; strong performance, but focused only on melanoma
This System (YOLOv12s + Logistic Reg.) (Image + Symptoms) [8][14]	2,880 images, 16 classes	79.17% mAP@0.5; class imbalance, limited dataset size

Note: It should be noted that many of these benchmarks are based on existing research found in the literature. However, individual cases will vary not only due to dataset size, but also due to how balanced the datasets are between class types.

This research extends existing methodologies by implementing a multi-input approach for skin classification, integrating YOLOv12s-based lesion detection with symptom-based analysis using Logistic Regression (late fusion). This new framework is intended to enhance diagnostic accuracy and decrease overlapping classification of similar appearing skin lesions while also providing more affordable and user-friendly access to skin classification services in low-resource areas of medical care.

Materials and Methods:

This section covers the data sources, the beginning processing methods, the machine learning models, the fusion techniques, and the framework behind intelligent dermatological recognition diagnosis presented by this project. This study consists of two complementary components: a visual deep learning framework in conjunction with symptoms-based classification and multimodal Fusion, which will provide more accurate diagnoses of skin pathologies.

Dataset Collection and Preparation:

To develop and improve the model, we put together a large dataset of images. The datasets were taken from different sources, which are publicly available dermatology datasets that can be easily found on Kaggle [15] and DermNet [16]. They were enlarged using Adobe Stock and Pinterest as additional sources of high-resolution clinical images. Each source of images used strict criteria to acquire its images. In total, there are 16 unique dermatological conditions and diseases in the dataset, including acne, eczema, psoriasis, rosacea, and skin cancer, which also include a large variety of clinical morphological characteristics to train machine-learning algorithms to allow for generalization in diagnosis across many different dermatological conditions.

A total of 2,880 images were collected. The collection of images was split into three subsets: training (70%), validation (20%), and testing (10%). The images were processed to all have the same size (640 x 640 pixels). As required by the YOLOv12s architecture. The images were annotated manually using the LabelImg tool [17], with bounding boxes drawn around each lesion and label information exported in YOLO format with normalized coordinates. Various augmentation techniques were implemented on the training subset to increase generalization and reduce the likelihood of the model becoming too fitted to the training set (overfitting), including random horizontal flipping, jittering (i.e., adjusting the brightness, saturation, or hue of an image) and mosaic augmentation (i.e., creating one composite image from four separate training images to facilitate the modeling of more complex visual environments, thus increasing the size of the training dataset).

A dataset of structured symptoms was developed simultaneously with clinical symptoms to build a dataset of clinical indicators using structured data collection methods.

The resulting dataset consists of 248 individual records that contain binary and categorical clinical indicators reported by patients regarding the presence of symptoms, including redness, itching, pustules, scaling, pain, and symptom duration. The curated, de-identified dataset also served as a key link between visual signs/symptoms and patient-reported experiences, helping to create a multimodal learning model to enhance the accuracy of diagnostic predictions.

Image-Based Detection Model: YOLOv12s:

The R-LEAN module is the underlying architectural foundation of YOLOv12s and has provided an increase in gradient flow stability through the whole network, providing improved representation of input features. YOLOv12s also employs the use of 7 x 7 separable convolutions, which provide efficient spatial pattern extraction compared to other convolutional algorithms. The function of the neck within this network was built to pull together features from all input/output to allow for synthesis of features across several levels of hierarchy, in conjunction with a detailed area-based attention mechanism (FlashAttention), to allow for robust analysis of areas within the utilized space for possible clinical diagnostic purposes. The detection head of YOLOv12s is functionally decoupled, which enables separate localization/classification operations to increase the predictive performance of the overall detection systems.

Through pre-training and standard development of a visual pattern using the MS COCO [18] dataset, we used Stochastics Gradient Descent (SGD) to establish the model's performance in the dermatology database by way of SGD training over 25 epochs with momentum of 0.937, an initial learning rate of 0.01, and scheduled cosine annealing/learning rates to optimize model generalization across the datasets. During testing, the model converged around the 25th epoch; all subsequent evaluations had very high overfitting (more than generalizing) based on the number of samples available in this dataset to learn the data from; for all stages of training, we employed random horizontal flipping, mosaic composition, and color shifting to increase the robustness of the model and improve its ability to generalize on unseen and other datasets.

After the training phase, the model was exported into the Open Neural Network Exchange (ONNX) format [19], which allowed fast and efficient deployment across multiple platforms. Using the ONNX Runtime allows for inference with very low delays, that is good for applications that need to run in real-time. The last processing step specified was Non-Maximum Suppression, which used an IoU score value of 0.45 to prune duplicate bounding boxes and leave only the most confident prediction in any group of identical predictions.

Symptom-Based Prediction Model: Logistic Regression:

Logistic Regression is the algorithm that was used by the diagnostic portion of the software to classify conditions through symptoms that were presented by a user (i.e., doctor or other caregivers). Logistic Regression was a commonly used method of statistical classification that is widely utilized for multiple class prediction problems and thus is suitable for use in the clinical environment. Structured and tabular type symptom data is suited to Logistic Regression, therefore offering a transparent method of making clinical decision while providing clinicians with probability-based outputs.

To implement classifiers for the study, a classifier scikit-learn library [20] was used; L2-Ridge-regularization (ridge regression) is a common default used in scikit-learn to reduce overfitting and therefore encourage generalization. Categorical variables were converted to on-hot-encoded format, and continuous features were normalized. The disease labels were converted from text labels to numeric values using a LabelEncoder. An evaluation dataset was then created by splitting the resulting prepared dataset into 168 samples for training and 80 samples for testing.

Multinomial classification was performed using the Broyden-fletcher-Goldfarb-Shanno known as L-BFGS [21]. This is summarized in Table 5 – experimental setup specifications used during the process. Training was conducted for up to 1,000 iterations to ensure convergence, with a fixed random state of 42 to guarantee reproducibility.

The logistic Regression model created a probability distribution of all disease classes associated with the input set of symptoms. These probability scores indicated the model's confidence in diagnosing each disease and served as an additional source of evidence that supplements image-based analysis, thus improving the overall decision support system.

Late Fusion Strategy:

A late fusion architecture was used to aggregate independent predictions obtained from two models: one producing result based on images, the second producing results based on symptoms. Results from each model were compiled and merged once all analyses had been completed, thereby maintaining both models' respective strengths in terms of visual and clinical information.

An integrated sequential pipeline was used for this approach. The YOLOv12s model outputs bounding boxes and confidence scores for detected skin abnormalities, which are converted into normalized probabilities. Concurrently, the Logistic Regression classifier processes patient symptom data to produce a separate probability distribution over all possible diagnoses. These probabilities are then combined using a weighted average, with a relative weight of 6.0 for visual input and 0.4 for symptom data, determined via grid search on a validation dataset. The result of this fusion is a fused probability distribution that provides both an initial diagnosis and a confidence score for that diagnosis.

This methodology for the weighted fusion of clinical diagnostic data coincides with the sequence of working through clinical diagnoses in standard medical practice, in which visual evidence takes precedence over other forms of evidence, such as the use of symptomology and other clinical characteristics to differentiate morphologically similar conditions. The adjustable framework allows for an increase in the weight of symptoms for some conditions, therefore improving the overall accuracy and utility of a clinical diagnostic classification system.

System Architecture and Implementation:

The underlying architecture of the web-based system for an end-user diagnostic pipeline includes the Django REST API [22] as the backend and PostgreSQL [23] as the data storage system. A core component of the system is the end-user diagnostic pipeline, which supports real-time dermatological assessments. A case begins when a patient submits clinical images of their skin condition. After an image is submitted, the YOLOv12s object detection system will detect objects within the submitted image and produce a baseline visual classification by identifying the specific outline of each lesion.

Following an initial diagnosis, the user interface will request a "Targeted" symptoms questionnaire to narrow down possible conditions and include the patient's past medical information. The survey contains a series of structured follow-up questions that address the current symptoms, at least some of which can be answered in binary (e.g., "do you have itching, yes/no"), and /or give categorical answers (e.g., symptom duration). The results will be processed through the Logistic Regression classifier to determine the final classification.

Upon retrieving probability results from both the image model and the symptom model, the final fusion module synthesizes these scores into a final diagnostic consensus. The platform produces a results dashboard that includes a confidence score for the final diagnosis, the original image overlaid with bounding boxes, and a personalized list of management recommendations, including referral instructions for a qualified healthcare provider.

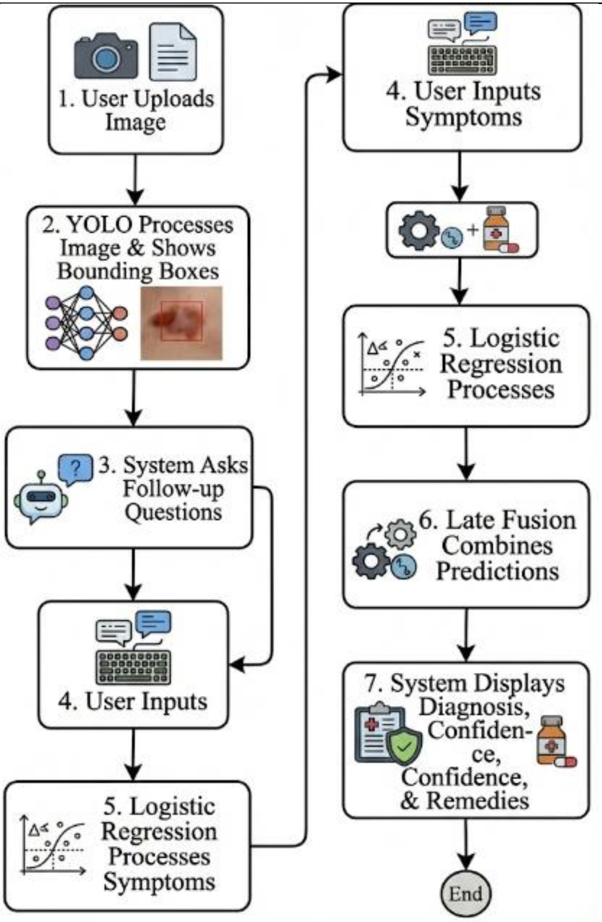
1. User Uploads Image
2. YOLO Processes Image & Shows Bounding Boxes
3. System Asks Follow-up Questions
4. User Inputs
5. Logistic Regression Processes Symptoms
4. User Inputs Symptoms
5. Logistic Regression Processes
6. Late Fusion Combines Predictions
7. System Displays Diagnosis, Confiden-ce, Confidence, & Remedies
End

Figure 1. Workflow of Multimodal Diagnostic System

Evaluation Metrics:

The proposed method was evaluated based on the results of three different analyses of performance: the evaluation of object detections (YOLOv12s), the evaluation of single modality classification, and the evaluation of multimodal predictions.

YOLOv12s Image Detection Performance:

To evaluate the accuracy of detection from YOLOv12s, evaluating how accurately it detected object using mAP@0.5 (at an Intersection over Union (IoU) threshold of 0.5) was the primary measurement for the YOLOv12s model's accuracy on all dermatology classes, and for measuring accuracy by determining the trade-offs between Precision (proportion of correctly identified positives) and Recall (correctly identifying all relevant ground truth lesions) from YOLOv12s. The F1-Score was the harmonic mean of Precision and Recall. These scores provide an overall evaluation of model performance.

Detection results, along with computational profiles, are provided by YOLOv12s Table 3 on the held-out test set. The mAP@0.5 for the model was calculated at 0.638, Precision was 0.715, Recall was 0.597, resulting in an optimal F1 of 0.648. The structure was highly effective, with each inference processing 10.4 milliseconds and a total of 9,237,072 parameters; it can therefore be used efficiently in resource-constrained settings for deployment.

Table 3. Performance Metrics of YOLOv12s on the Test Set

Metric	Value
mAP@0.5	0.638
Precision	0.715

Multimodal System Performance:

To evaluate the performance of the Logistic Regression symptom classifier, researchers assessed overall accuracy and individual class Precision, Recall, and F1-Score to determine how well it discriminates between different skin conditions. The entire system, specifically the late fusion of image and symptom predictions, was evaluated by final matching of fused diagnoses with labels assigned by expert physicians. A confusion matrix was created to identify the patterns in any errors made by the model.

Table 4. Comparison of Single-Modality and Multimodal System Performance

Model	Modality	Accuracy (%)	F1-Score (%)
Symptom-based LR	Symptoms only	100	100
Image-based YOLOv12s	Image only	67	64
Proposed Late Fusion	Image + Symptoms	79.17	77.78

Table 4 presented a comparative summary of unimodal and multimodal performance. The symptoms-only classifier achieved 100% accuracy and F1-score, which was directly influenced by the small size and nature of the symptom set. The YOLOv12s images model at its regular status had an accuracy of approximately 67% and an F1-score of approximately 64%. both classifiers—the symptom classifier and YOLOv12s images model—are combined into the integrated late fusion system, there is approximately 79.71% accuracy and an F1-score of approximately 77.78%. Therefore, the use of visual markers combined with clinical symptoms greatly increases diagnostic accuracy.

Experimental Setup:

We have implemented the architecture of the model using PyTorch as the deep learning engine (2.0 version) and the Scikit-learn library (1.2 version) for the classical machine-learning portion in the workflow, which was created using a programming environment that runs Python 3.9 on an experimental basis. For training purposes, Google Colab Pro was used with an NVIDIA Tesla V100 GPU (16 GB of video memory). To simulate real-world scenarios, inference speed and quality of the model were evaluated using standard CPUs during inference. The source code was managed under version control via Git and saved at various intervals to ensure that the models' states were reproducible during each phase of training.

Table 5. Experimental Setup Specifications

Component	Specification
Environment	Python 3.9
Deep Learning Framework	PyTorch 2.0 [24]
Machine Learning Library	Scikit-learn 1.2
Training Hardware	NVIDIA Tesla V100 GPU (16 GB VRAM)
Inference Hardware	Standard CPU Configuration
Version Control	Git
Check pointing	Periodic save for reproducibility

Results and Discussion:

The evaluation of the multimodal system was conducted using standard metrics designed for the detection and classification of objects, as well as measurements for evaluating the diagnostic performance across all disease types. First, a wide range of diagnostic performance was assessed on the YOLOv12s image object detection model, followed by a broader analysis of results for all disease types.

YOLOv12s Image Detection Performance:

To assess the YOLOv12s model's ability to accurately detect all types of lesions, we considered the average precision (AP) of the model at a specific Intersection over Union (IoU) threshold of 0.5. As shown in Table 3, detection results provided by YOLOv12s reached an mAP@0.5 of 0.638. The test dataset included 16 skin condition types. The average AP achieved by YOLOv12s across all classes was 0.638 at an IoU threshold of 0.5, demonstrating its capability in real-time lesion detection. The Precision-Recall curve shows the aggregate average accuracy aligns with the per-class average accuracy for each of the types of skin condition.

As determined by class- specific Average Precision scores (AP), the precision of detected objects to their real-world equivalents varies by object type. The highest average precision was reported to be 0.881 for detection of Eye Bags and 0.831 for detection of Skin Cancer. The lowest detection AP scores were found in classes such as psoriasis (0.479 AP) and sunspots (0.497 AP). This disparity is likely due to visual similarities between these classes, class imbalance in the dataset, and fewer training examples for these conditions. Evaluation of the F1-confidence curves revealed that a detection confidence threshold of 0.310 produced the highest overall F1-score of 0.620 across all diagnostic categories.

[10][7] reported benchmarks that had much better accuracy metrics on their test (88-91%) than we did on the image-only YOLOv12s, our metric only reflects accuracy on those images but out performance would appear to be poorer than theirs due to the complexity of our task with 16 disease types or classifications, having fewer images for training, testing and random image sources being used to create our dataset. Our 79.17% performance and predictive performance are improved by using patient-symptoms-based late fusion as part of the final prediction regarding the potential benefit of combining visual and clinical information in underserved clinical settings.

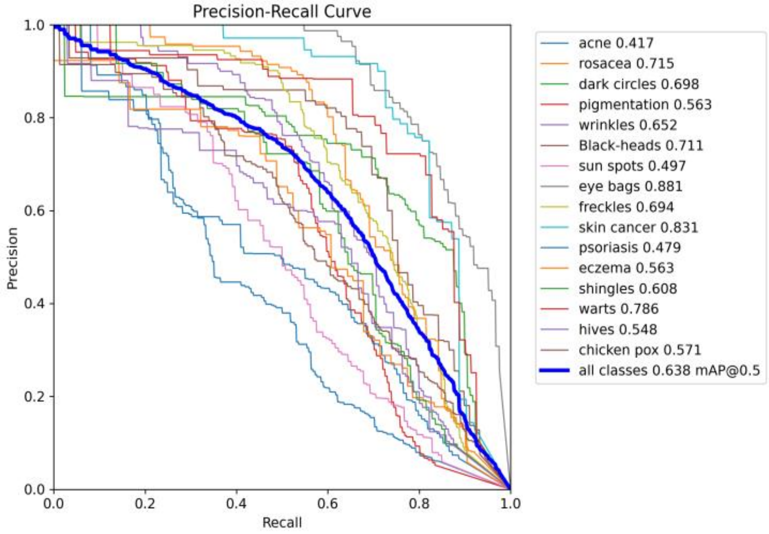

Figure 2. Precision-Recall Curve for YOLOv12s Detection

The plot of Average Precision (AP) for each skin condition shows the average mAP@0.5 of 0.638.

Model Training and Convergence:

During model training, the monitoring of loss curves and evaluation metrics occurred at all stages of training. The three main loss functions—box, classification, and focal loss—consistently decreased over the 25 training epochs. The downward trend on the loss curves of the training data indicates that the model can accurately localize lesions despite performing a complex task with multiple classes, as shown in Figure 3. Additionally, the downward trend demonstrates that the model achieved stable convergence over time. Validation losses mirrored the training losses, suggesting minimal risk of overfitting (i.e., no substantial overfitting). Effective regularization and data augmentation contributed to the absence of significant overfitting.

As a reduction in loss, we saw a consistent upward trend across all evaluation metrics, including Precision and Recall, from the start to the end of training. The important metric of mean Average Precision (mAP) on a mean Average Precision (mAP) at 0.5 to 0.95 has continued to increase consistently through training cycles. Improvements in map@50 were 0.63, and in mAP@50 to mAP@95 were 0.35 at the end of our training cycle confirmed that localization and detection accuracy have continued to improve as training has continued.

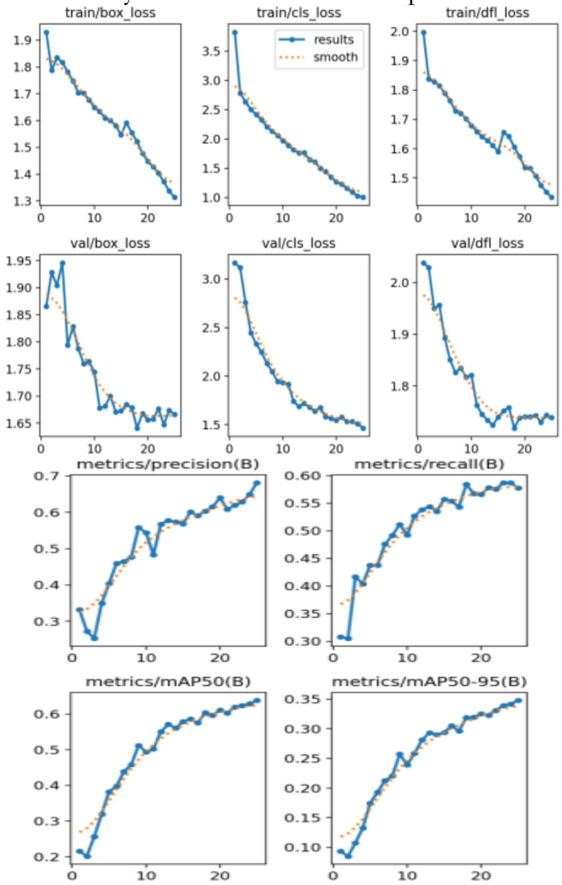

Figure 3. YOLOv12s Training and Validation Metrics

Figure 3: The training and validation metrics for each epoch are displayed in this plot. The x-axis shows the training epochs, and the y-axis shows the corresponding loss or performance metric for each epoch. There are five metrics demonstrated on this plot, which contain bounding box loss, classification loss, distribution focal loss, precision, recall, mAP@0.5, and mAP@0.5:0.95. There are two colored curves on this plot, with blue curves showing the actual performance metrics for each epoch and the orange curve showing a smoothed version of the curves for ease of visualization in terms of their respective trends regarding convergence.

Classification and Misclassification Analysis:

An accurate view of a performance from a classification model can be shown through the use of a normalized confusion matrix. A confusion matrix contains counts indicating correct and incorrect classifications, that will show an indication of the performance of the model's ability to correctly classify a condition. High numbers on the diagonal of confusion metrics indicate good detection of conditions and are known as True Positive rates or Recall, i.e., the recall or number of recalled conditions for a particular condition. The conditions with the highest recalls are Skin Cancer (0.73), Dark Circles (0.73), and Blackheads (0.71), with Eye Bags (0.71) and Shingles (0.60) closely behind.

Examining off-diagonal entries reveals patterns of misclassification among diagnoses. Misclassification of Rosacea occurred at a rate of 67%, or approximately two-thirds of all true

Rosacea cases were misdiagnosed as dark Circles. Acne was also misclassified more than half the time; over 57% of all true acne images were classified as “Background”. In contrast, the “Background” category was mistakenly classified for other diagnoses, including Acne (0.57), Rosacea (0.28), and Chicken Pox (0.42). Differentiating images showing minor lesions from normal skin is very difficult in these areas. Additionally, the majority of the predicted images in the Wart category had previously been classified as Chicken Pox; nearly 22% of the predicted Wart images were identified as Chicken Pox.

The high level of inaccuracy in visually overlapping conditions (such as Rosacea and Dark Circles) demonstrates that a more reliable diagnostic approach is to include adjacent or visually similar conditions in conjunction with other relevant data, creating a clinically more relevant diagnosis. Additionally, the data collected to support the development of the system validates the design by indicating that using a late fusion strategy with patient-reported symptom data improves overall diagnostic accuracy.

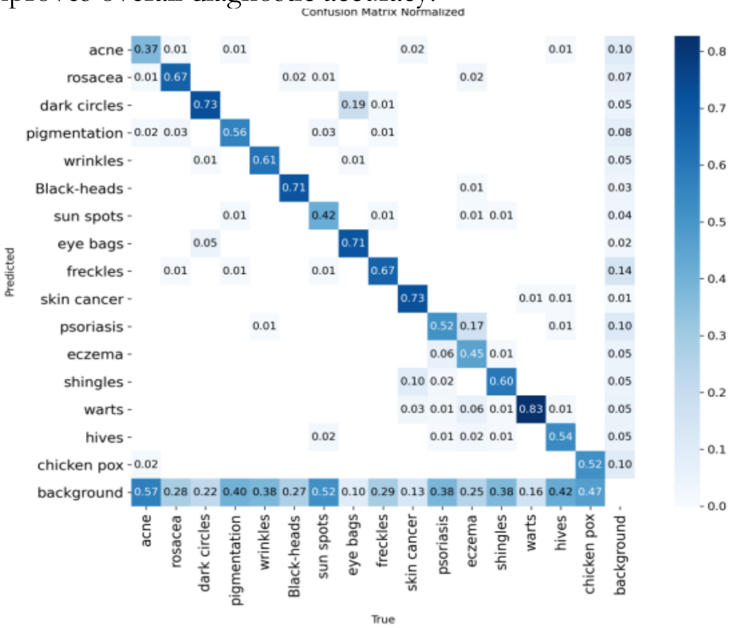

Figure 4. Normalized Confusion Matrix

Integrated Multimodal System Performance:

The diagnostic platform created utilizes both the confidence scores of the YOLOv12s image model and the classifications of symptoms identified by the Logistic Regression methods for a more comprehensive understanding of a patient’s health. The two approaches were combined using a methodology known as “weighted late fusion,” whereby experimental weighting was used for developing handcrafted weights, and the final resulting weights were 0.6 for images and 0.4 for symptoms. The multimodal diagnostic platform provides a significant improvement in accuracy and F1-scores relative to either a single modality approach, achieving 79.19% accuracy and 77.78% F1-Score.

While benchmark studies [6][7] suggests that image alone system have higher accuracies (88-91%) than those which rely on a range of other input modalities; because they are generally focused on only a limited subset of binary classifications – little more than two options – the current study will analyze an entirely different level of complexity: by examining 16 different dermatological conditions from many different perspectives using real-world clinical images versus a much smaller training database than world sally be required give how often you clinician might see and evaluated patients before being able to provide treatment. Comparing the accuracy of the image only YOLOV12s program (67% accuracy, 64% F1), shows expanding the use of patient-sourced symptoms documents through late fusion adds significant face-value clinical validity towards improving treatment reliability; thus, demonstrating the promise of multimodal systems with clinical practice where access to dermatologists is limited across our country.

Conclusion:

The YOLOv12s visual detection model was optimized to work with a second classifier based on Logistic Regression to analyze symptomatic conditions from patients by employing a late fusion method in which weighted factors were applied to each of the modalities (visual and symptomatic) to develop a multimodal supportive method of identifying and diagnosing skin disease. By combining a variety of modalities for evaluating symptoms, we were able to provide reliable and interpretable diagnostic support relative to systems that only utilized one modality of input. Longitudinal measure supported the effectiveness of the multimodal support system and demonstrated its ability to provide practical and timely diagnostic support to the end-user by providing accurate and efficient identification and diagnosis of skin lesions. Ultimately, this system demonstrated its sufficient performance to achieve a final diagnostic accuracy of 79.17%, while also providing a real-time capability for detecting lesions during the application of the YOLOv12s module. The findings support an essential idea that interconnectedness between imaging evidence and patient symptoms improves diagnostic accuracy, especially in instances of similar visual disease. Designed for affordability, this system is based on a web platform to provide separate user interfaces for the user, clinician, and administrator. This simplifies interactions between users and the system. Because of the low-cost operation compared to the expenses of visiting a doctor-in-person, the use of this system makes it a viable option for limited resource locations that cannot access specialized medical services. Throughout the entire process of image acquisition to individual recommendations, there exists evidence for the implementation of accessible tele dermatology. The intent of future investigation will include developing more complex neural networks for the modelling of symptoms, conducting additional rigorous clinical validation research, and expanding the training database in order to provide greater consistency of the model across a broad range of patient demographics. Thus, this method may improve diagnostic access and outcomes for underserved populations worldwide by applying an established AI-assisted clinical care approach in real-life settings.

References:

[1] H. Chen et al., (2022) published "The global burden of skin and subcutaneous diseases from 1990 to 2019: a systematic analysis of the Global Burden of Disease Study 2019," in the British Journal of Dermatology. The paper is available in volume 187, no 3, and pages 344–353.

[2] Henry W. Lim, Scott A.B. Collins, "The burden of skin disease in the United States," *J Am Acad Dermatol*, vol. 76, no. 5, pp. 958–972, 2017, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28259441/

[3] L. Tomas-Aragones and S. E. Marron, (2023) published "The Psychosocial Burden of Skin Diseases," Actas Dermo-Sifiliográficas (English Edition), This paper is available in volume 114, no 5 and pages 403–412, [Online]. Available: https://doi.org/10.1016/j.ad.2023.02.004

[4] Awan, M. O., Khan, M. N., Ejaz, A, "The dermatology workforce crisis in Pakistan: Challenges and opportunities," *J. Pakistan Assoc. Dermatologists*, vol. 30, no. 2, pp. 222–228, 2020.

[5] K. M. Hosny, M. A. Kassem, and M. M. Fouad, (2023) published "Deep learning in dermatology: A systematic review," Medical Image Analysis, vol. 85, p. 102744. [Online]. Available: https://doi.org/10.1016/j.media.2023.102744

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "Human–computer collaboration for skin cancer recognition," Nature Medicine, vol. 26, pp. 1229–1234, 2020.

[7] Y. Liu *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nat. Med. 2020 266*, vol. 26, no. 6, pp. 900–908, May 2020, doi: 10.1038/s41591-020-0842-3.

[8] Yunjie Tian, Qixiang Ye, David Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," *arXiv:2502.12524*, 2025, [Online]. Available: https://arxiv.org/abs/2502.12524

[9] Farhat Afza, Muhammad Sharif, "Multiclass Skin Lesion Classification Using Hybrid Deep Features Selection and Extreme Learning Machine," *Sensors*, vol. 22, no. 3, p. 799, 2022, [Online]. Available: https://www.mdpi.com/1424-8220/22/3/799

[10] Chubin Ou, Sitong Zhou, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Front. Surg.*, vol. 9, 2022, [Online]. Available: https://www.frontiersin.org/journals/surgery/articles/10.3389/fsurg.2022.1029991/full

[11] Sören Richard Stahlschmidt , Benjamin Ulfenborg, Jane Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Brief. Bioinform.*, vol. 23, no. 2, 2022, doi: https://doi.org/10.1093/bib/bbab569.

[12] J. Terven, D. M. Córdova-Esparza, and J. A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1680–1716, Dec. 2023, doi: 10.3390/make5040083.

[13] Manar Elshahawy, Ahmed Elnemr, "Early Melanoma Detection Based on a Hybrid YOLOv5 and ResNet Technique," *Diagnostics*, vol. 13, no. 7, p. 2804, 2023, [Online]. Available: https://www.mdpi.com/2075-4418/13/17/2804

[14] Karen Drukker, Weijie Chen, "Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations," *Eur. J. Radiol. Artif. Intell.*, vol. 3, p. 100030, 2025, doi: https://doi.org/10.1016/j.ejrai.2025.100030.

[15] "Dermatology image dataset." Accessed: Feb. 07, 2026. [Online]. Available: https://dermnetnz.org/dermatology-image-dataset

[16] "DermNet." Accessed: Feb. 07, 2026. [Online]. Available: https://dermnetnz.org/

[17] "(PDF) A Survey in Deep Learning Model for Image Annotation." Accessed: Mar. 09, 2026. [Online]. Available: https://www.researchgate.net/publication/332109738_A_Survey_in_Deep_Learning_Model_for_Image_Annotation

[18] Zhengxia Zou, Keyan Chen, "Object Detection in 20 Years: A Survey," *Proc. IEEE*, 2019, [Online]. Available: https://arxiv.org/abs/1905.05055

[19] "ONNX | Home." Accessed: Feb. 07, 2026. [Online]. Available: https://onnx.ai/

[20] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Jun. 16, 2024. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[21] R. Y. Sun, "Optimization for Deep Learning: An Overview," *J. Oper. Res. Soc. China 2020 82*, vol. 8, no. 2, pp. 249–294, Jun. 2020, doi: 10.1007/s40305-020-00309-6.

[22] " Django: The web framework for perfectionists with deadlines." Accessed: Feb. 07, 2026. [Online]. Available: https://jashangill3592.pythonanywhere.com/article/6

[23] "PostgreSQL: The world's most advanced open source database." Accessed: Feb. 07, 2026. [Online]. Available: https://www.postgresql.org/

[24] Zakariya Ba Alawi, "A Comparative Survey of PyTorch vs TensorFlow for Deep Learning: Usability, Performance, and Deployment Trade-offs," *arXiv:2508.04035v1*, vol. 8, 2025, [Online]. Available: https://arxiv.org/html/2508.04035v1