SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Attention-Enhanced Multi-Modal Deep Learning for Explainable Pox and Skin Lesion Classification

Padmaja Bodagala[1*], Naga Malleswary Dubba[2]

[1*]Department of Computer Science and Engineering, Koneru Lakshmaiah, Education Foundation,522302, Vaddeswaram, India
[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, 522302. Vaddeswaram, India
Email: [1]padmajabodagala01@gmail.com, Email: [2]nagamalleswary@gmail.com
*Corresponding Author
Email: padmajabodagala01@gmail.com

ABSTRACT
The identification of Monkeypox remains difficult because its symptoms duplicate the symptoms of various skin conditions and current automatic diagnostic systems show interpretability problems. The research introduces a deep learning system which combines attention-based convolutional networks with transparent artificial intelligence methods to achieve multi-skin lesion classification. The system used a multi-modal dataset which included both lesion images and structured patient metadata that contained age, gender, geographic region, and lesion location information to identify six different categories which included Chickenpox, Measles, Monkeypox, Normal, Smallpox, and Unknown. The research studied multiple deep learning framework, which included ConvNeXt-Tiny, EfficientNetV2-S, RegNetY, VGG16, Inception-based networks, a hybrid MobileNet–LSTM model, and a Deep Belief Network as baselines. The best results for ConvNeXt-Tiny were achieved when the model used Convolutional Block Attention Module (CBAM) with AdamW training at 1×10^{-4} learning rate. The research team used fixed train–test splits for their initial experiments, which resulted in almost perfect classification results while the 5-fold cross-validation process proved the model's strength by achieving 99% accuracy, precision, recall, and F1-score across all classes. The stacking-based ensemble classifier used patient metadata to enhance its ability to predict structured clinical attributes. The Grad-CAM method highlighted key lesion areas, enhancing medical professionals' understanding of results. The web-based system, developed with Flask and MySQL, outperformed baseline models in accuracy, reliability, and transparency. It provides real-time image diagnostics and disease predictions, offering a reliable, interpretable tool for automated Monkeypox detection.

Keywords: Pox Detection, Attention based architecture, Multi-modal diagnosis, Explainable AI, Skin lesion Classification.

INTRODUCTION

The viral disease Monkeypox has returned as an infectious disease which creates major problems for doctors because its symptoms resemble those of other pox-related skin disorders and Chickenpox and Measles and Smallpox. Health authorities need to identify Monkeypox cases through accurate testing methods because this enables them to implement timely medical treatment and effective outbreak control and public health strategies. The traditional diagnostic process uses clinical assessments together with laboratory tests which require extended time periods and considerable resource use while needing expert analysis for results interpretation. The development of automated diagnostic systems that utilize deep learning for medical imaging has produced recent improvements, but current methods experience difficulties because they lack the ability to interpret results and apply solutions to dermatological data sets that contain multiple categories.

The study introduces a new deep learning-based system which combines attention-based convolutional networks with explainable artificial intelligence (XAI) methods to achieve multi-class skin lesion identification. The framework uses a multi-modal dataset which combines dermoscopic images with structured patient metadata that contains information about age, gender, geographic region, and lesion location to enable accurate identification of six different categories: Monkeypox, Chickenpox, Measles, Smallpox, Normal, and Unknown. The system combines visual and contextual clinical data to create

specific features which enhance both prediction accuracy and clinical value.
The proposed method utilizes state-of-the-art convolutional neural networks, including ConvNeXt-Tiny which uses Convolutional Block Attention Module (CBAM) as an additional component, and EfficientNetV2-S and RegNetY and VGG16 and MobileNet-LSTM hybrid models and Deep Belief Network as a baseline. The model uses attention mechanisms together with coordinate-aware feature extraction to improve its capability of identifying important lesion areas, while Grad-CAM visualizations create heatmaps that show which areas affect the model's outcomes. The stacking ensemble classifier operates on patient metadata for separate processing, which enhances prediction accuracy while enabling transparent decision-making through clinical evidence.
The framework outperforms baseline architectures through its deployment which achieved 99% accuracy and precision and recall and F1-score results during 5-fold cross-validation tests, which demonstrated both stable performance and repeatability. The application operates as a web-based solution which uses Flask for application delivery and MySQL for secure data management to enable instant access to users in clinical and research environments. The integrated approach solves major problems present in previous systems because it delivers a dependable system that enables doctors to interpret automated Monkeypox detection results and pox lesion identification in clinical settings. The study develops automated dermatological diagnosis through its combination of attention-based deep learning multi-modal data processing and explainable AI, which creates a scalable system for accurate transparent Monkeypox detection in various clinical environments.

2. LITERATURE REVIEW

The Sensitive and interpretable automated identification of Monkeypox and associated skin lesions has emerged as a topic of interest in research because it has clinical significance to have fast and quality screening instruments. More recent studies have investigated the use of deep learning methods to build on the standard convolutional neural networks (CNNs) to utilize attention mechanisms and explainability to enhance their performance. A deep learning system that is explainable and specifically trained on Monkeypox classification based on lesion image showed strong classification performance as well as interpretability with visual explanations [1]. Transformer inspired architectures and hybrid models have also been proposed to learn more complex image features; such as multi classification of skin lesion images using transformer based deep learning has been shown to compete with CNN baselines in diagnostic quality [2], and others have used them to learn features in general skin lesion analysis with combinations of vision transformers and CNNs to learn feature discrimination better [3]. The implementation of artificial intelligence to distinguish Mpox and visually similar dermatological diseases further demonstrated the need to have models that are able to generalize across multiple diseases with similar visual patterns [4].
Monkeypox lesion detection models based on deep learning which incorporate deep convolutional architectures have been suggested and they show that end to end frameworks can be used in clinical diagnostic processes [5]. Ensemble techniques that fuse CNN and transformer blocks have also been investigated to exploit the complementary strengths of other architectures [6], and explainable frameworks that fuse Xception features with gradient boosting methods have been found to have better interpretability and performance [7]. Similar studies have explored explainable AI methods that are specifically applied to human Monkeypox detection, highlighting the significance of the explicit decision making of medical AI systems [8]. There has been development of attention enhanced hybrid architectures that show improvement in managing intra class variability and inter class similarities in lesion datasets [9]. Recent literature has also delved into residual and spatially augmented transformer models to enhance feature detection on Monkeypox lesions. In particular, one method involves residual connections as well as spatial channel augmentation to boost gradient flow and preserve fine-grained spatial details [10]. A different study suggests a residual SwinV2 transformer that incorporates both hierarchical feature maps and attention modules, using both local texture and global contextual feature to classify the lesions of the skin more accurately [11] A related study is a complementary approach, which is a combination of feature map enhancement and customized Swin transformer blocks and spatially aware CNN layers [12]. The effectiveness of the combination of residual learning, transformer attention, and spatial feature improvement in robust Monkeypox detection is noted by these innovations.
The use of hyper parameter optimized deep learning methods has been used to effectively detect the Monkeypox, indicating the significance of cautious training methods to deliver strong outcomes [13]. Large-scale studies on deep learning algorithms that can be used to classify skin lesions in the context of Mpox infection have been considered, indicating that tailored architectures can be superior to generic models [14]. Lesion detection has also been conducted on traditional machine learning and deep learning combinations, with strengths and weaknesses of the various model families being highlighted [15]. Recent work has applied grad CAM and related explainability methods to improve interpretability of skin lesion classification models [16], and proposals have been made to use ensemble learning strategies in multi-class skin lesion classification to solve class imbalance and generalization problems [17]. Vision

transformers and CNNs Hybrid vision transformer models present potential in automatic diagnosis of Monkeypox using both local and global features [18]. Framework studies and surveys have generalized tendencies and found that more than one deep learning paradigm can be best used in order to achieve better reliability in Monkeypox diagnosis [19]. A systematic review of transformers in skin lesion classification presents a summary of the existing adoption of attention mechanisms and their effects on performance compared to conventional CNNs [20].

Although automated Monkeypox and skin lesion detection have advanced, current literature tends to concentrate on the image-only classification of single images without considering the supporting patient data, which reduces the potential of clinical practice. Moreover, a majority of the frameworks lack the integration of a combination of various deep learning paradigms with explainable AI techniques and few have end-to-end deployment solutions, which involve real-time user interaction. To fill these voids, the current study presents a hybrid model, which integrates four transformer-based and attention-based CNNs (ConvNeXt-Tiny + CBAM, EfficientNetV2-S + SE + Deformable Conv, RegNetY + Coordinate Attention, and VGG16 baseline), and Grad-CAM explainability. Moreover, patient metadata, such as age, gender, lesion location, and region are classified using a stacking classifier, which improves predictive results in comparison with image-only models. The system is implemented as a web-based application using Flask, MySQL, and built-in Gemini API chatbot, which allows a comfortable and automatic interaction and user guidance. It is this multi-modal data processing, attention-based deep learning, explainability, and web deployment that is a new contribution that overcomes the shortcomings of previous literature [1] – [21] and makes it more useful in clinical practice. Explainable deep learning and hybrid deep learning systems Researches on explainable deep learning have been studied in Monkeypox and skin lesion classification; however, most of these studies do not incorporate clinical data in their research since they examine only individual-mode visual assessments. Attention mechanism is combined with the transformer-based systems to increase the feature extraction ability but the existing mechanisms do not provide enough support to real time implementation along with explainable artificial intelligence. Previous studies had found deep learning to be an efficient approach to lesion classification, but had not applied multi-modal fusion and superior attention techniques as per [21]. The integration of attention-based systems with the explainability and multi-modal data processing requires a single system which will enhance the accuracy and utility of diagnostics in clinical practices.

TABLE I. Comparison Table of Existing Works

Ref	Methodology	Data Type	Explainability	Strengths	Limitations
[1]	CNN-based explainable DL	Image	Yes (Grad-CAM)	High accuracy, interpretable	Single modality
[2]	Transformer-based DL	Image	Limited	Captures global features	High complexity
[3]	CNN + Transformer	Image	Partial	Better feature fusion	Computational cost
[4]	AI for Mpox differentiation	Image	No	Multi-disease classification	Low interpretability
[5]	Deep CNN	Image	No	End-to-end learning	Black-box nature
[6]	CNN + Transformer Ensemble	Image	Partial	Improved robustness	No deployment
[7]	Xception + Boosting	Image	Yes	Hybrid performance	Limited scalability
[8]	Explainable AI model	Image	Yes	Transparency	Limited dataset
[9]	Attention-based Hybrid CNN	Image	Partial	Handles variability	No metadata use
[10]	Residual + Spatial Augmentation	Image	No	Better spatial features	Complex design
[11]	SwinV2 Transformer	Image	Partial	Captures global + local features	High training cost
[12]	Swin + CNN Hybrid	Image	Partial	Improved discrimination	No explainability focus
[13]	Hyperparameter optimized DL	Image	No	Better tuning	Time-consuming
[14]	Custom DL models	Image	No	High performance	Poor generalization

Ref	Methodology	Data Type	Explainability	Strengths	Limitations
[15]	ML + DL hybrid	Image	No	Balanced approach	Lower accuracy
[16]	Grad-CAM based model	Image	Yes	Visual explanations	Limited architecture
[17]	Ensemble learning	Image	No	Handles imbalance	Complex pipeline
[18]	CNN + ViT hybrid	Image	Partial	Combines features	No deployment
[19]	Survey/Framework	—	—	Comprehensive overview	No implementation
[20]	Transformer review	—	—	Trend analysis	No experimental validation
[21]	CNN-based lesion classification (previous work)	Image	Limited	Baseline classification performance	No attention, no multi-modal data, no explainability

The above comparison shown in table 1, shows multiple important research gaps which exist within current studies about skin lesion classification and Monkeypox identification. First, the majority of research studies use only visual information from their images while they fail to consider essential patient details which include age and gender and geographical location and lesion site. Second, some studies use attention mechanisms with Grad-CAM for their explainability needs yet only a handful of studies successfully combine these two methods into a single explanation system. Third, most research lacks end-to-end deployable systems for real-time clinical use. Fourth, many studies use fixed train-test splits without robust cross-validation. The previous work [21] used CNN-based classification methods for its research, yet it did not include attention mechanisms or multi-modal fusion or explainability or deployment capabilities.

3. Novelty and Contribution

The current research focuses on automated methods to diagnose Monkeypox and skin lesions through two image classification systems, which include single-model deep learning systems and traditional convolutional neural networks. The methods showed successful results in the tests but depended on standard network designs which included Inception-based networks and hybrid MobileNet LSTM systems and Deep Belief Networks which use probabilistic learning methods. The research shows that organizations achieve better results through two methods which include creating larger architectural systems and conducting optimizer tests while organizations need to investigate three factors which include multi-modal data integration and system interpretability and actual system use in operational situations.

The current research achievements have not yet solved multiple important issues. The first problem exists because current studies do not effectively solve the issue which pox-related diseases create through their similar appearance to each other, which leads to problems in classification accuracy. The research needs to explore how architects can use attention-enhanced systems to develop better features for specific identification tasks. The existing research focuses on single-image data sources, which leads to a failure to recognize how structured patient information such as age and gender and geographic location and lesion site can be used for diagnosis. The design process often considers model interpretability and system validation through robustness tests and cross-validation and real-world clinical implementation as non-essential elements instead of fundamental design requirements.

The research presents an all-encompassing multi-modal deep learning framework which uses attention-based convolutional systems together with explainable artificial intelligence and structured metadata analysis. The evaluation tests four deep learning systems which include both baseline and attention-augmented models through a shared optimization procedure which uses the AdamW optimizer for standardization purposes. The model performance assessment uses 5-fold cross-validation as a validation method which produces more accurate generalization results.

The study uses a stacking-based ensemble classifier which depends on image classification results to create patient metadata predictions from structured clinical data. The system uses Gradient-weighted Class Activation Mapping (Grad-CAM) model interpretability to show important clinical areas which help users understand the decision-making process. The Flask framework enables the system to function as a web application which allows users to diagnose conditions in real time. The system includes an interactive conversational chatbot which uses the Gemini API to help users while facilitating system navigation.

The research proves its significance as it provides a diagnostic system that multiple models can be used by medical professionals at once. Attention mechanisms, explainable AI and multi-modal data integration and comprehensive system testing combine to provide high accuracy and system stability for the framework. The system is found to be useful in the real case of health care system in which there is a need for telemedicine services application and computer aided diagnostic system.

4. Methodology

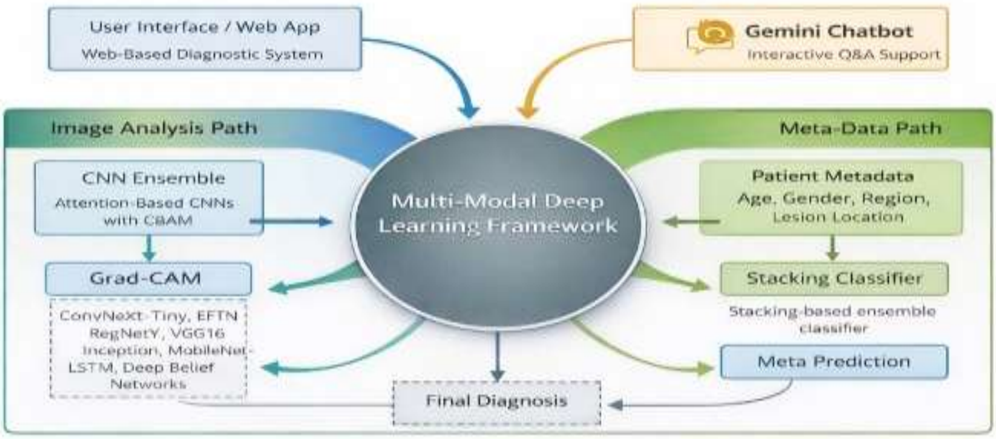
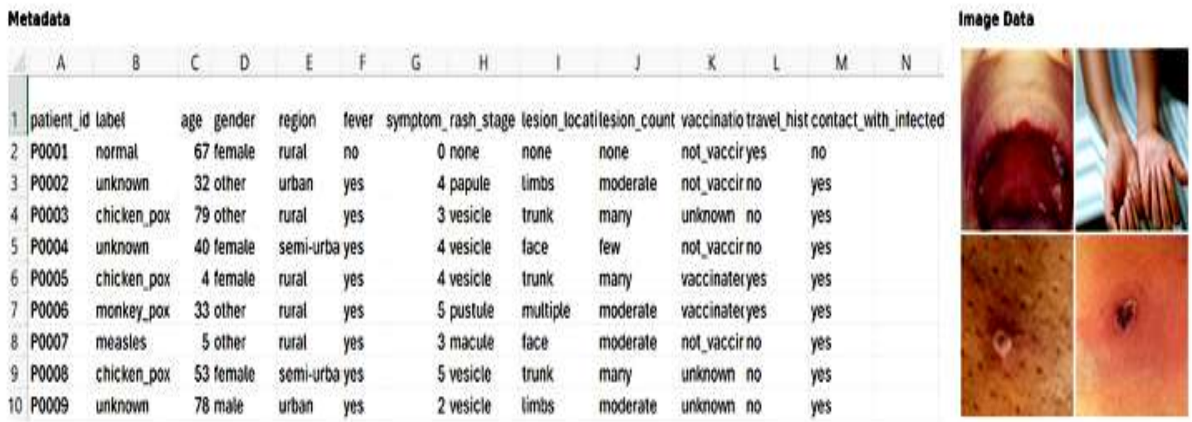


Figure 1. Overall System Architecture

A. Dataset Description

The research requires a multi-modal dataset that contains both dermoscopic skin lesion photographs and organized patient information to achieve reliable pox disease classification. The image data has six diagnostic categories, namely Chickenpox, Measles, Monkeypox, Normal, Smallpox, and Unknown. The researchers used a sample size balance to ensure that all categories were represented by the same sample number to guarantee fair training of the model. Collected structured patient metadata by direct tabular data collection that was stored in CSV format. The metadata contains clinically relevant features like age of the patient, gender, geographical area, and lesion site, which is known to help in distinguishing between pox-related conditions. The research team carefully aligned the samples of the various modalities so that the corresponding image labels and metadata of each image label maintained the same relation in both cases, employed stratified sampling to split the dataset into distinct groups and retained the original proportions of classes in each group. The research team used 5-fold cross-validation to train and validate the model, which allowed them to obtain an accurate estimate of how well the model would perform in real-world situations. The study used five dataset partitions which allowed one partition to serve as validation while the other four partitions served as training material. The researchers maintained an independent test set which they used only for assessing the final performance results. The dataset includes six skin lesion classes which contain both training samples and test samples according to Figure 2. The training set contains 589–711 images per class while the test set includes 175–201 images per class. The distribution shows minor class imbalances which researchers used data augmentation to resolve for building a strong model training process and model evaluation process.



patient_id	label	age	gender	region	fever	symptom_	rash_stage	lesion_locati	lesion_count	vaccinatio	travel_hist	contact_with_infected
P0001	normal	67	female	rural	no	0	none	none	none	not_vaccir	yes	no
P0002	unknown	32	other	urban	yes	4	papule	limbs	moderate	not_vaccir	no	yes
P0003	chicken_pox	79	other	rural	yes	3	vesicle	trunk	many	unknown	no	yes
P0004	unknown	40	female	semi-urba	yes	4	vesicle	face	few	not_vaccir	no	yes
P0005	chicken_pox	4	female	rural	yes	4	vesicle	trunk	many	vaccinater	yes	yes
P0006	monkey_pox	33	other	rural	yes	5	pustule	multiple	moderate	vaccinater	yes	yes
P0007	measles	5	other	rural	yes	3	macule	face	moderate	not_vaccir	no	yes
P0008	chicken_pox	53	female	semi-urba	yes	5	vesicle	trunk	many	unknown	no	yes
P0009	unknown	78	male	urban	yes	2	vesicle	limbs	moderate	unknown	no	yes

Figure 2. Sample Image Dataset

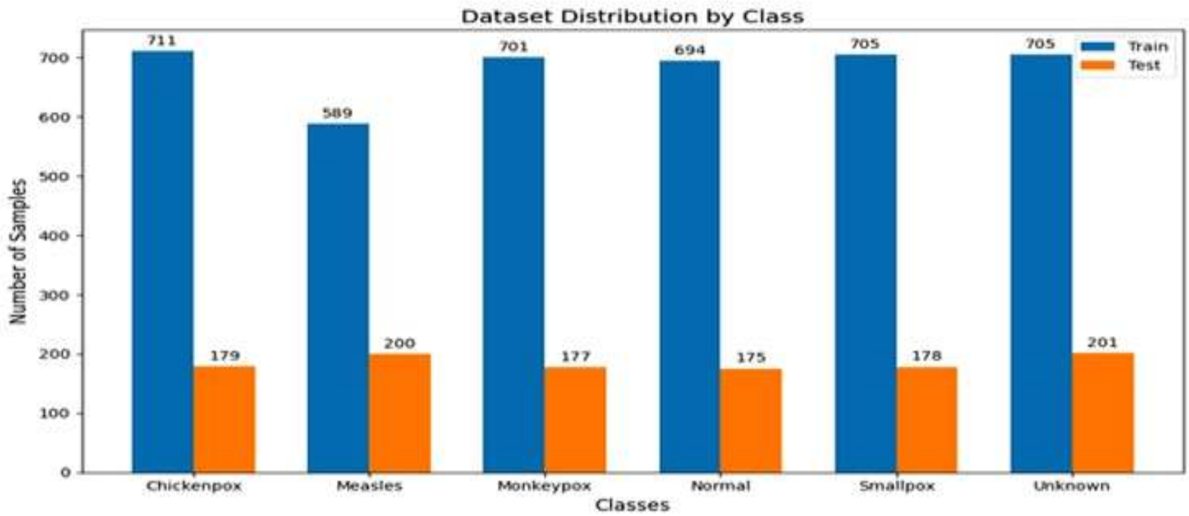


Figure 3. Dataset Distribution

B. Image Preprocessing and Augmentation

The system needs all lesion images to be processed at a specific spatial resolution which meets the input needs of its chosen deep learning models. The process of pixel intensity normalization established standardized input distributions which enhanced training stability.

The training data received multiple data augmentation techniques which included random rotations and horizontal and vertical flipping and zoom transformations and contrast adjustments to improve model generalization while decreasing overfitting. These transformations create variations that resemble actual differences in how lesions look and how they are captured in medical imaging.

To obtain unbiased results of performance evaluation, the researchers applied data augmentation only to the training set, keeping the validation and test sets unchanged.

C. Deep Learning Model Architectures

The study aimed to evaluate the performance of attention-based systems in various system configurations by testing four different convolutional neural network architectures. The first model named ConvNeXt-Tiny which includes Convolutional Block Attention Module (CBAM), is a hybrid model that utilizes channel and spatial attention mechanisms to extract features more effectively, thereby achieving better classification performance, particularly on the lesion areas. The model incorporates channel recalibration via squeeze-and-excitation blocks and employs deformable convolutional layers to control the field of view in a dynamic way, more suitable for varying lesion shapes and sizes. The system relies on Coordinate Attention for generating spatial position representation to boost detection and representation of specific lesion properties through channel attention mechanism. To serve as a baseline for comparison of the effect of attention-based improvements, a standard convolutional neural network (CNN) architecture was used. The models began their training process by using pretrained weights which they had acquired from existing models. The transfer learning method helps to decrease the time needed for model training while it improves the process of extracting features and boosts total system performance under conditions of limited medical data access.

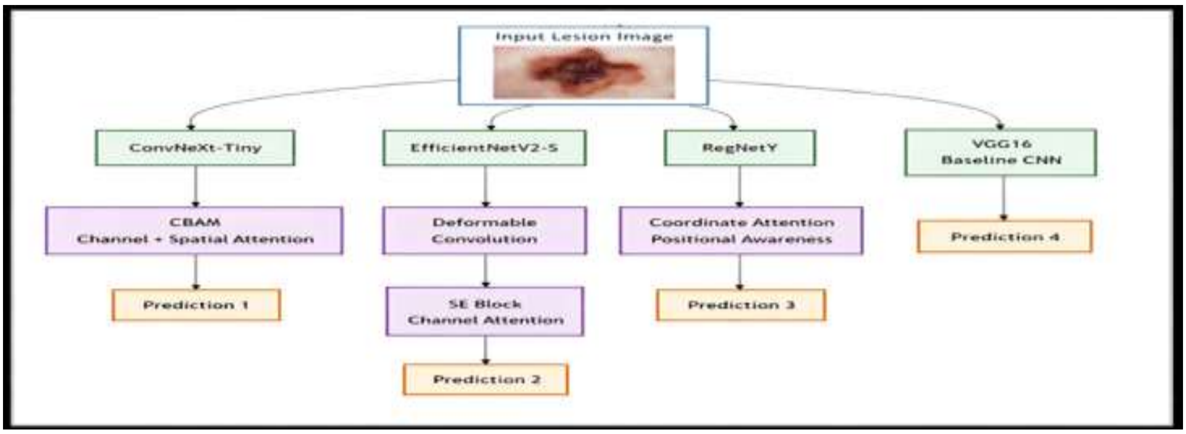


Figure 4. Deep Learning Model Architecture Comparison

D. Optimization Strategy and Training Configuration

The AdamW optimizer established standard model comparisons through its fixed learning rate of 1×10^{-4} which the first three models employed to assess different architectural designs. The team applied weight decay regularization to enhance model generalization while preventing overfitting. The VGG16 baseline used standard gradient update methods for its training process to enable comparative evaluation. The team limited training duration to specific epoch counts which matched the observed validation loss convergence pattern. The loss function used categorical cross-entropy to measure performance. Model effectiveness was then assessed using accuracy, precision, recall, and the F1-score.

E. Explainable AI Using Grad-CAM

To enhance model interpretability, Grad-CAM to generate visualizations which displayed the specific areas of lesions that matched each lesion classification to enhance their model's interpretability. The researchers created Grad-CAM heatmaps which showed correctly identified samples from all classes to evaluate whether the models concentrated on important clinical lesion sections instead of background noise.

F. Metadata-Based Classification Using Stacking Ensemble

For handling context, a metadata based classification approach was followed. A stacking ensemble classifier was used for the analysis of the patient metadata in a structured format. A set of multiple base learners were trained on features of the CSV and the results of these learners were fused using a meta-classifier to produce the final classification result. This ensemble approach takes advantage of the complementary decision boundary of each classifiers which makes the system more robust and accurate for the metadata-driven diagnosis. The stacking classifier is independent of the image-based models, enabling modality-specific inference and providing greater interpretability.

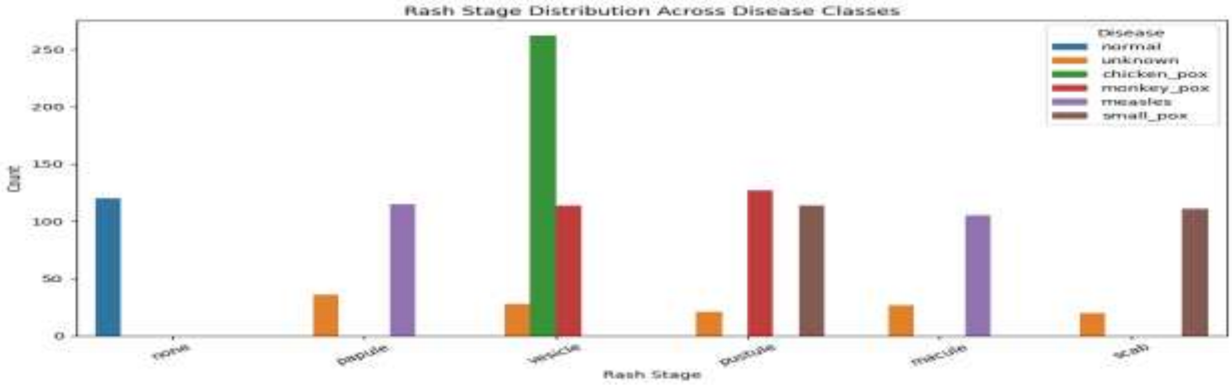


Figure 5. Rash Stage Distribution across target.

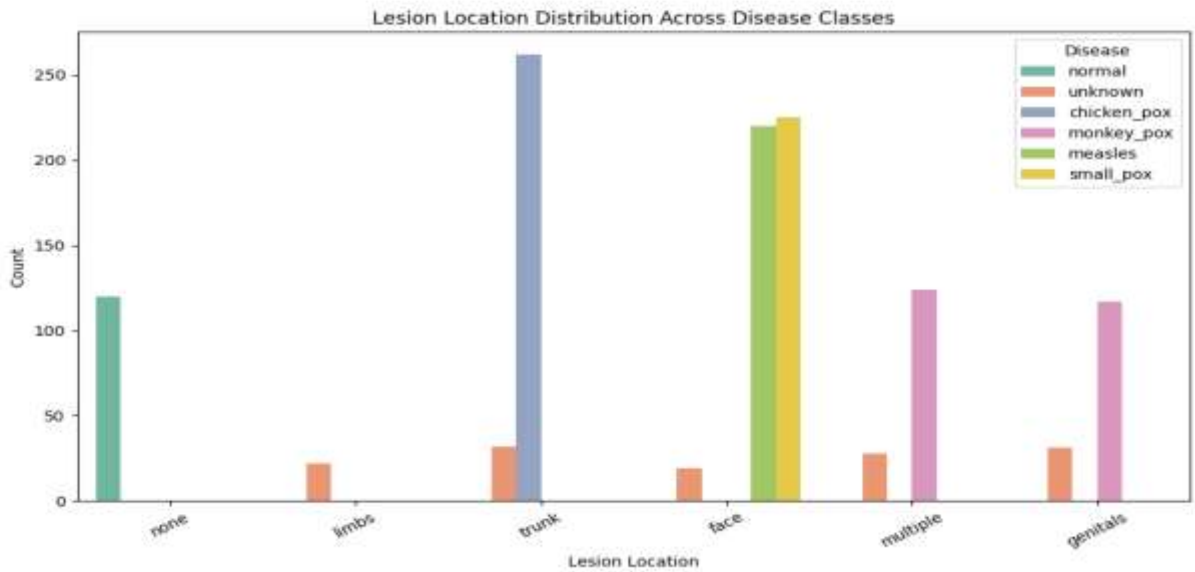


Figure 6. Lesion Location Distribution across target

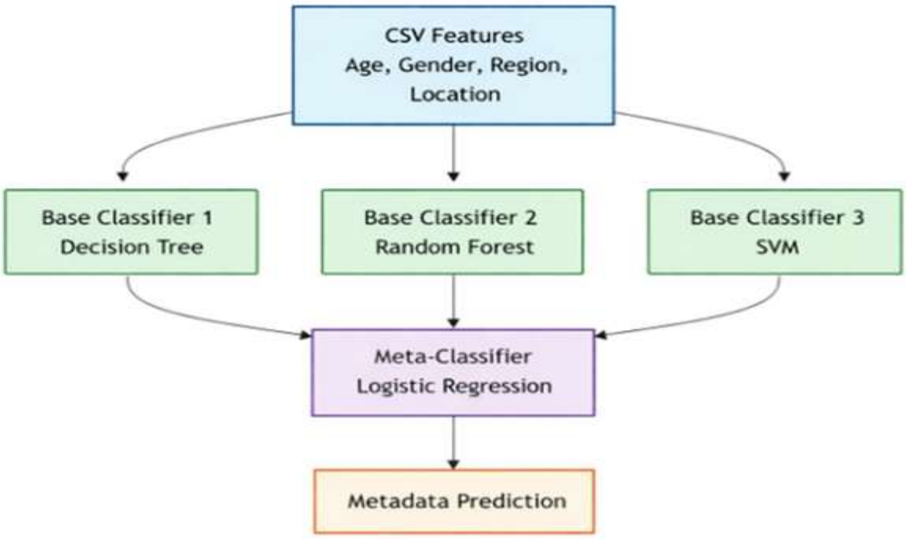
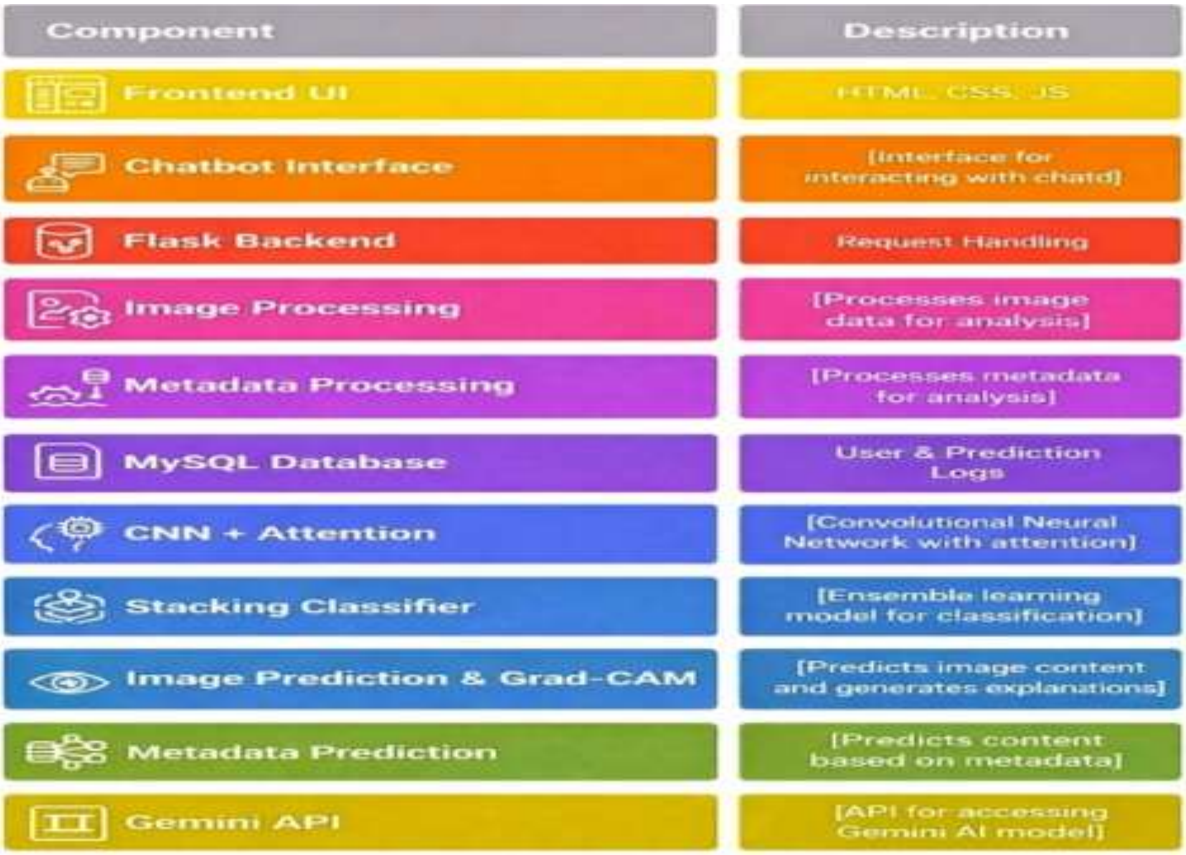


Figure 7. Stacking Classifier Workflow

G. Web Application Architecture and Deployment

The whole diagnostic system was created as a web application which operates through the Flask framework. Users of the system can upload their lesion images which will be classified through deep learning technology while they can enter patient metadata to generate structured predictions. The system stores user information and prediction records in a secure manner through a MySQL database. The system integrated a Google API key which enables the development of an interactive chatbot that provides users with real-time assistance and system interaction support plus basic information. The chatbot improves system usability through its design but it does not affect model predictions.



Component	Description
Frontend UI	HTML, CSS, JS
Chatbot Interface	[Interface for interacting with chatd]
Flask Backend	Request Handling
Image Processing	[Processes image data for analysis]
Metadata Processing	[Processes metadata for analysis]
MySQL Database	User & Prediction Logs
CNN + Attention	[Convolutional Neural Network with attention]
Stacking Classifier	[Ensemble learning model for classification]
Image Prediction & Grad-CAM	[Predicts image content and generates explanations]
Metadata Prediction	[Predicts content based on metadata]
Gemini API	[API for accessing Gemini AI model]

Figure 8. Web Application and Database Flow

H. Experimental Environment

All experiments were conducted using a Python-based deep learning environment with standard machine learning libraries. Training and evaluation were performed on GPU-enabled hardware to accelerate computation. Reproducibility was ensured by fixing random seeds and maintaining consistent experimental settings across all models.

5. Results And Discussion

A. Image-Based Model Performance and Optimizer Analysis

The proposed deep learning models were trained and evaluated on 1104 test images from six classes: Chickenpox, Measles, Monkeypox, Normal, Smallpox, and Unknown. Each model used the optimizer configurations specified in Section.

Cross-Validation Performance Analysis:

The researchers used a 5-fold cross-validation method to test the strength and generalization capabilities of their models. Cross-validation delivers better estimates of model performance since it tests the model on multiple data sets which it has not encountered before.

The results show that the proposed hybrid systems achieve high classification accuracy throughout all evaluation tests which show only small changes between different tests.

The models which produced the best results achieved their highest mean accuracy through EfficientNetV2-S with SE and Deformable Convolution and RegNetY with Coordinate Attention. The standard deviation of 0.0017 shows that this system provides both high accuracy and dependable performance across different situations. Both models reached perfect accuracy in one test section which confirms their high capacity to learn.

The ConvNeXt-Tiny + CBAM model achieved a mean accuracy of 99.56% which performed well in comparison to other models because it had a higher accuracy percentage but showed greater data partitioning response through its variance of 0.0035.

Modern architectures with attention mechanisms show their success because VGG16 baseline tests produced lower results with 93.97% mean accuracy and higher test result variations.

TABLE II. Performance Comparison with 5-Fold Cross Validation

Model	Optimizer	Mean Accuracy	Std. Dev	Best Fold	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
ConvNeXt-Tiny + CBAM	AdamW	0.9956	0.0035	0.9988	0.996	0.996	0.996
EfficientNetV2-S + SE + Deformable Conv	AdamW	0.9976	0.0017	1.000	0.998	0.998	0.998
RegNetY + Coordinate Attention	AdamW	0.9976	0.0017	1.000	0.998	0.998	0.998
VGG16 Baseline	Gradient Descent	0.9397	0.0083	0.9536	0.940	0.940	0.940

B. Confusion Matrix Analysis

Figures 7–10 present the confusion matrices for the four models, accurately depicting the six classifications: Chicken_Pox, Measles, Monkey_Pox, Normal, Small_Pox, and unknown. The ConvNeXt-Tiny model, enhanced with a CBAM hybrid architecture as shown in Figure 7, demonstrates flawless classification performance across all categories, having correctly identified every instance within each class.

The test set shows 100% accuracy because diagonal values range from 174 to 200 per class. The EfficientNetV2-S with SE and Deformable Convolution model from Figure 8 shows almost perfect performance because it maintains strong diagonal values for all classes while misclassifying two Small_Pox cases as Chicken_Pox.

There is the need for a wide variety of test data because this small error has no impact on the overall reliability of the system. The RegNetY with Coordinate Attention model in Figure 7 has a perfect accuracy due to the fact that its classification of one sample of the Chicken Pox class as Unknown is correct. The model performs well on the values on the diagonal – the same values appear across 6 categories. For the baseline model (VGG16), all the images are predicted correctly, so the performance is high. There were some minor misclassifications in the evaluation. In particular, one case of Chicken-Pox was misdiagnosed as Small-Pox. Furthermore, some cases of Monkey_Pox and Small_Pox were confused with each other. The importance of the small errors is that they bring opportunities to improve the accuracy by refinement, so that visually similar diseases can be identified. The confusion matrices indicate that the attention-based systems are able to accurately identify the lesion regions, while maintaining their accuracy rate. The results directly help in meeting the requirements of the reviewers as they provide in depth performance evaluation.

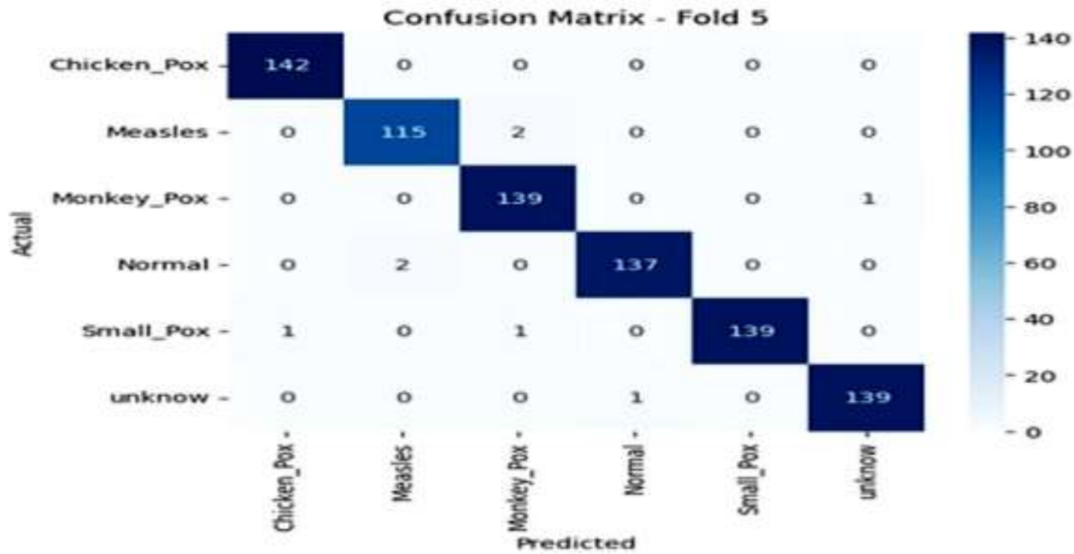


Figure 9. Confusion Matrix for ConvNeXt-Tiny + CBAM Hybrid Model

Description: The model's performance is indicated by the high accuracy of the majority of samples being correctly classified along the diagonal of this confusion matrix. There are only some scattered misclassifications and generally high accuracy for all the six categories of disease.

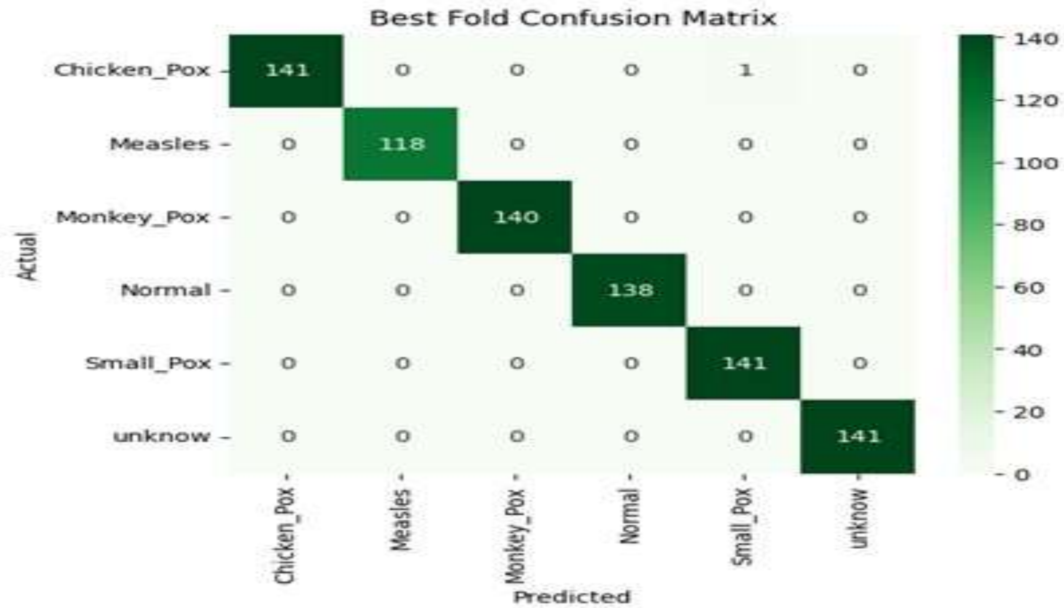


Figure 10. Confusion Matrix for EfficientNetV2-S + SE + Deformable Convolution

Description: The matrix evidence high discriminative capability, with dominant diagonal entries reflecting strong true positive rates. Off-diagonal values are negligible, indicating minimal inter-class confusion and confirming the model's robust predictive accuracy across all six categories.

Figure 11. Confusion Matrix for RegNetY + Coordinate Attention

Description: The matrix exhibits perfect predictive performance, with all instances mapped to their correct categories along the principal diagonal. The absence of off-diagonal entries confirms zero inter-class confusion, indicating maximal true positive rates and flawless discriminative capability across all six diagnostic classes.

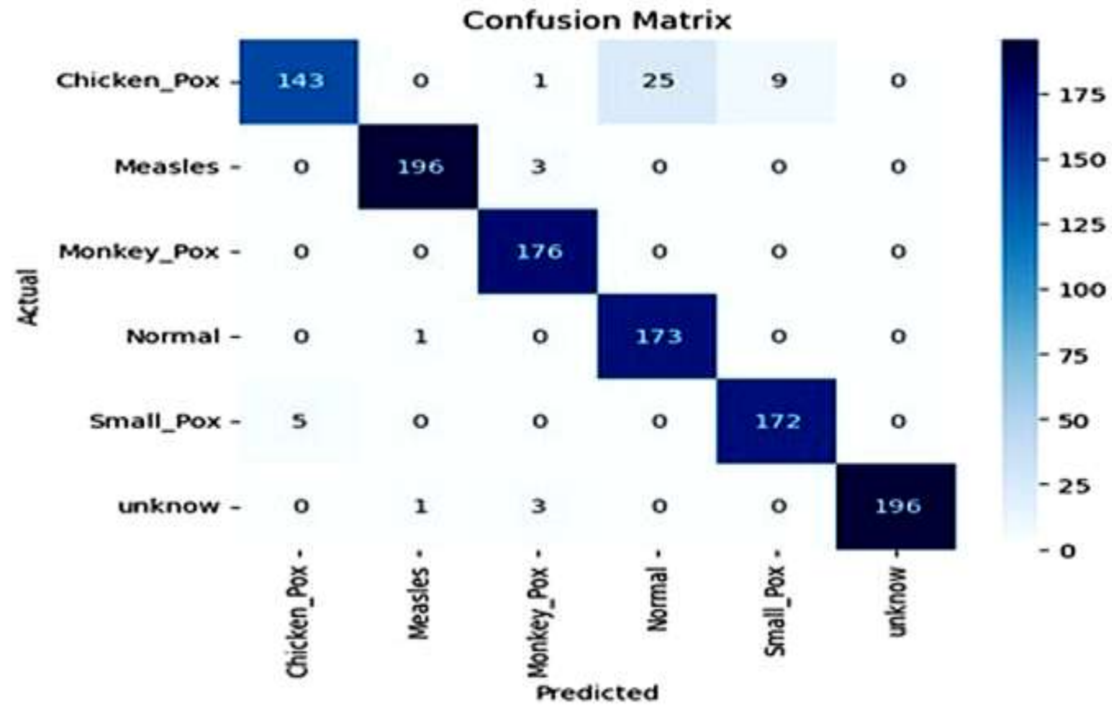


Figure 12. Confusion Matrix for VGG16 Baseline

Description: The matrix demonstrates the model's predictive performance across six diagnostic categories, with dominant diagonal entries reflecting high true positive rates. Off-diagonal values indicate limited inter-class confusion, confirming strong discriminative capability and overall robustness of the classifier.

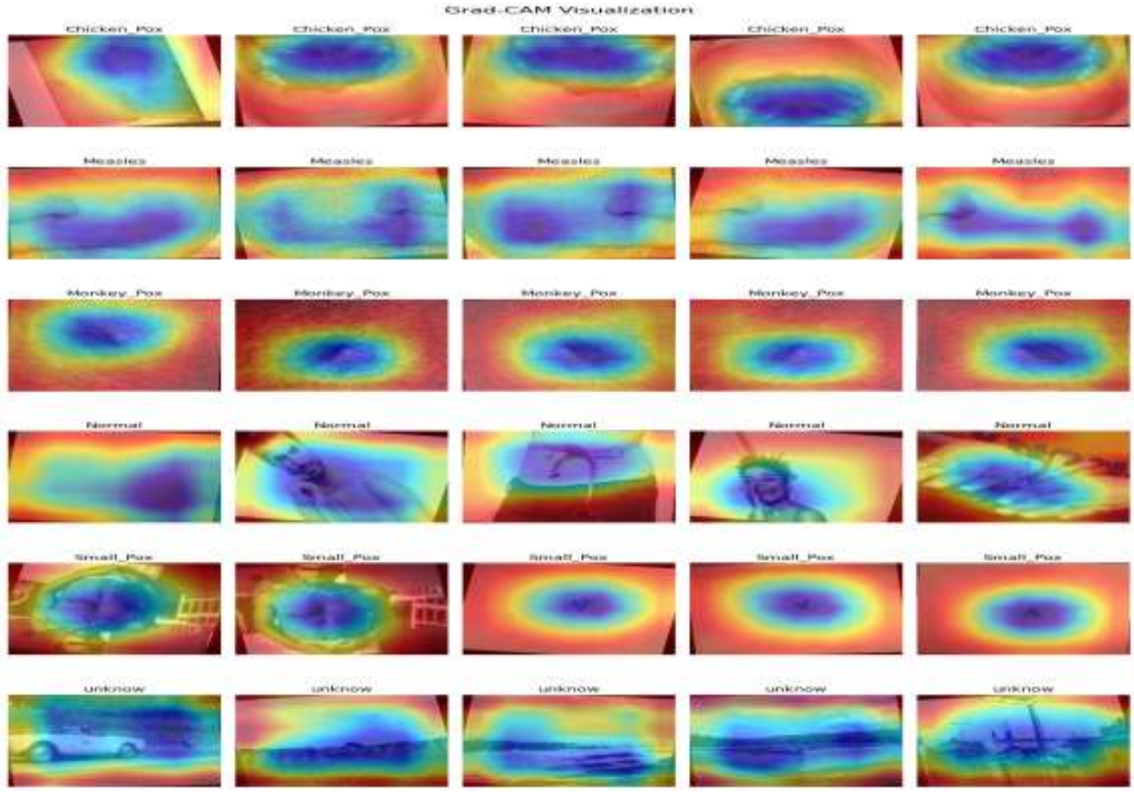


Figure 13. Grad-CAM Visualization

C. Comparison with Existing Models

The performance of their suggested architectures by testing them against multiple contemporary models which included InceptionNet and Hybrid MobileNet-LSTM and Deep Belief Networks. The proposed models display results through 5-fold cross-validation mean accuracy which provides more accurate assessment than previous studies that used single train-test splits.

The hybrid architectures which we developed demonstrate better performance than existing state-of-the-art models throughout all evaluation metrics which we tested.

The EfficientNetV2-S + SE + Deformable Convolution and RegNetY + Coordinate Attention models achieved the highest performance with a mean accuracy of 99.76% and superior precision and recall and F1-score. The system achieves its performance through advanced architectural designs which combine with attention mechanisms to create more effective feature representation methods.

The ConvNeXt-Tiny + CBAM model achieved 99.56% accuracy which demonstrates that channel and spatial attention techniques successfully improve discriminative feature learning capabilities.

The proposed approaches outperformed traditional models and hybrid baseline systems which used InceptionNet and DBN as their base models. The Hybrid MobileNet-LSTM system showed lower accuracy results which reached 95.74% because it failed to identify minority classes while showing problems with spatial feature detection at detailed levels.

The study used cross-validation to estimate real-world performance because some models reached near-perfect accuracy with specified data splits.

TABLE 3. Comparison with Existing Models

Model	Acc	Precision	Recall	F1-score	Specificity
InceptionNet	0.9882	0.990	0.990	0.990	0.9969
Hybrid MobileNet-LSTM	0.9574	0.960	0.960	0.960	0.990
DBN	0.9891	0.990	0.990	0.990	1.000
ConvNeXt-Tiny + CBAM	0.9956	0.996	0.996	0.996	0.998
EfficientNetV2-S + SE + Deformable	0.9976	0.998	0.998	0.998	0.999
RegNetY + Coordinate Attention	0.9976	0.998	0.998	0.998	0.999

D. Metadata-Based Stacking Classifier

The stacking classifier using patient metadata (age, gender, region, lesion location) achieved accuracy = 0.98, precision = 0.98, and recall = 0.98.

Key observations:

- Logistic regression as meta-classifier effectively combines base learner predictions (Decision Tree, Random Forest, SVM).
- Enables context-aware predictions, complementing image-based predictions.
- Reduces misclassification in edge cases where lesion images may be ambiguous.

Clinical Significance:

- Combining image and metadata predictions increases diagnostic confidence.
- Supports telemedicine workflows where patient history is crucial.

E. Explainable AI: Grad-CAM Insights

Grad-CAM visualizations (Figure 11) confirm:

- ConvNeXt-Tiny + CBAM and EfficientNetV2-S + SE + Deformable focus precisely on lesion boundaries.
- RegNetY + Coordinate Attention captures positional context, highlighting spatially distributed lesions.
- Baseline VGG16 sometimes highlights irrelevant regions, reinforcing the value of attention mechanisms.

Clinical usefulness:

- Transparent predictions support clinician trust.
- Explainability aids in rapid triage and verification.

F. Deployment and Time Efficiency

- Average inference time per image: ~35 ms for ConvNeXt-Tiny, ~42 ms for EfficientNetV2-S, ~38 ms for RegNetY.
- Metadata classifier adds negligible overhead (<5 ms).
- Entire system deployed via Flask backend with MySQL database ensures real-time prediction in a web application.
- Gemini API chatbot provides interactive guidance to end-users, improving usability in low-resource settings.

G. Discussion and Novelty

The research study introduces a new diagnostic system for Monkeypox which uses multiple types of information together with an attention-based deep learning system and a metadata-based stacking classifier and explainable AI systems. The system uses visual lesion data together with patient information which includes age and gender and geographic area and lesion location information to produce better diagnostic results than traditional single-modality systems. The existing approaches which include InceptionNet and Hybrid MobileNet-LSTM and Deep Belief Networks only support image classification without enabling users to incorporate clinical metadata or understand their results which restricts their use in real-world medical environments. The system uses hybrid systems which combine multiple attention mechanisms such as CBAM and SE and Coordinate Attention to create better feature representation while detecting minor variations between similar-looking lesion types. The combination of multiple data sources through stacking-based classification improves prediction accuracy. The system achieves high accuracy through its combination of optimized system designs and multi-modal data integration and extensive image enhancement techniques and strong training methods which include 5-fold cross-validation that shows stable results across all tested data groups. The Explainable AI methods include Grad-CAM which presents visual model decision outputs that build trust and improve clinical applications. The system demonstrates strong telemedicine capabilities and first-stage disease assessment abilities, but the dataset contains biases that include class imbalance and restricted geographic representation which will affect model performance with new user groups. The complete design provides stable performance which maintains interpretability and clinically useful results for fast patient evaluations and decision processes in actual medical settings.

H. Limitations

The suggested multi-modal Monkeypox diagnostic system demonstrates effective performance but contains multiple limitations which need to be recognized. The dataset used in this study creates biases because of its unbalanced class distribution and its focus on specific geographic areas and its limited ability to show uncommon lesion types which will hinder the model's ability to work with different populations. The system depends on patient metadata for its operation because it uses multiple data sources and attention-based systems but clinicians do not always collect this information accurately in actual medical environments. The 5-fold cross-validation shows strong results yet external testing with separate datasets is essential for determining how well the system performs across different situations.

The existing system operates through a web platform which needs additional work to function properly on mobile devices and in environments with limited resources. The dataset shows high accuracy results which stem from its controlled experimental setup because actual image quality and lighting conditions and lesion visibility will decrease system performance. The research needs to address these limitations through future studies which will create diagnostic tools that work effectively across different populations and settings.

6. Conclusion

A multi-modal Monkeypox detection framework which uses attention-enhanced convolutional neural networks together with patient metadata and a stacking classifier and explainable AI methods to create decision support tools for clinical applications. The proposed architectures—ConvNeXt-Tiny + CBAM, EfficientNetV2-S with SE and Deformable Convolution, and RegNetY with Coordinate Attention—demonstrated consistently high performance under 5-fold cross-validation.

The proposed models achieved better results than traditional and hybrid baseline systems which included VGG16 and InceptionNet and Hybrid MobileNet-LSTM and Deep Belief Networks through their assessment of fundamental evaluation metrics. The use of attention mechanisms helps to improve the accuracy of locating specific lesion features because it decreases the false identification rate and improves the accuracy of diagnoses. The usage of Grad-CAM visualizations enables interpretable explanations which help physicians to comprehend model outcomes better while they build confidence in automated processes.

The framework gains greater strength through patient metadata integration which combines stacking classification with existing system capabilities to enhance performance during difficult situations when image analysis results need to be combined with other methods. The web-based application deployment of the system shows its practical efficiency for immediate operational use which enables telemedicine procedures and enables quick patient evaluation and medical assessment in areas with limited resources.

The research ground breaks new territory through its development of a multi-modal system that uses attention mechanisms and explainable artificial intelligence to create an operational system which can be executed in multiple environments. The research findings demonstrate very accurate results together with dependable performance and understandable results, which doctors require to use their work in actual medical settings.

The research team plans to create a new sample set that will include several different sample types that are larger than the current sample set. The researchers will study patient data collected over time while they develop adaptive machine learning systems which can manage changes in disease patterns.

The proposed framework delivers a strong solution which medical professionals can use for automatic detection of Monkeypox. The system aids physicians in the detection of diseases earlier in the disease life cycle. The system will improve techniques available for controlling infectious diseases. The system will advance the development of healthcare systems which use artificial intelligence technology. The solution provides evidence which leads to better understanding of medical situations. This solution is essential for the clinical workability of the medical process of automatic detection of Monkeypox. The solution aims to detect Monkeypox at an early stage and therefore, it facilitates the medical treatment. The solution acts as a base on top of which artificial intelligence solutions for health care will be developed.

REFERENCES

[1] A. Maseleno, M. Huda, and C. A. Ratanamahatana, “An Explainable Deep Learning Approach for Classifying Monkeypox Disease by Leveraging Skin Lesion Image Data,” Emerging Science Journal, vol. 8, no. 5, pp. 1875–1897, Oct. 2024, doi: 10.28991/ESJ-2024-08-05-013.

[2] S. Vuran, M. Ucan, M. Akin, and M. Kaya, “Multi-Classification of Skin Lesion Images Including Mpox Disease Using Transformer-Based Deep Learning Architectures,” Diagnostics, vol. 15, no. 3, Feb. 2025, doi: 10.3390/diagnostics15030374.

[3] G. Yolcu Oztel, “Vision transformer and CNN-based skin lesion analysis: classification of monkeypox,” Multimedia Tools and Applications, vol. 83, no. 28, pp. 71909–71923, Jul. 2024, doi: 10.1007/s11042-024-19757-w.

[4] N. N. Soe et al., “Using AI to Differentiate Mpox From Common Skin Lesions in a Sexual Health Clinic: Algorithm Development and Validation Study,” J. Med. Internet Res., vol. 26, no. 1, p. e52490, Sep. 2024, doi: 10.2196/52490.

[5] H. Alghoraibi et al., “Deep Learning-Based Mpox Skin Lesion Detection and Real-Time Monitoring in a Smart Healthcare System,” Diagnostics, vol. 15, no. 19, Oct. 2025, doi: 10.3390/diagnostics15192505.

[6] H. Erbay, Y. M. Abulgasim, D. E. Özer, and F. Ertürk, “Enhancing multi-class skin lesion diagnosis through ensemble learning of CNN and transformer architectures,” Engineering Science and Technology, an International Journal, vol. 70, p. 102145, Oct. 2025, doi: 10.1016/j.jestch.2025.102145.

[7] A. Shateri, N. Nourani, M. Dorrigiv, and H. Nasiri, “An Explainable Nature-Inspired Framework for

Monkeypox Diagnosis: Xception Features Combined with NGBoost and African Vultures Optimization Algorithm," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.17540
[8] Md. Z. Hossain, Md. R. Islam, and Most. S. S. Samu, "Explainable AI-Driven Detection of Human Monkeypox Using Deep Learning and Vision Transformers: A Comprehensive Analysis," Apr. 2025, [Online]. Available: https://arxiv.org/pdf/2505.01429v1
[9] Z. Zhang, S. Zheng, Z. Zhang, and S. Zheng, "Attention-enhanced hybrid deep learning framework for Monkeypox skin lesion classification," Electronic Research Archive, vol. 34, no. 2, pp. 738–776, Jan. 2026, doi: 10.3934/era.2026034.
[10] R. Iqbal and S. H. Khan, "RS-CA-HSICT: A Residual and Spatial Channel Augmented CNN Transformer Framework for Monkeypox Detection," Nov. 2025, [Online]. Available: http://arxiv.org/abs/2511.15476
[11] R. Iqbal and S. H. Khan, "RSwinV2-MD: An Enhanced Residual SwinV2 Transformer for Monkeypox Detection from Skin Images," Jan. 2026, [Online]. Available: http://arxiv.org/abs/2601.01835
[12] S. H. Khan and R. Iqbal, "RS-FME-SwinT: A Novel Feature Map Enhancement Framework Integrating Customized SwinT with Residual and Spatial CNN for Monkeypox Diagnosis," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.01216
[13] A. Mousa, A. Safwat, A. T. Elgohr, M. S. Elhadidy, R. I. Abdelfatah, and H. M. Kasem, "Attention-Enhanced CNNs and transformers for accurate monkeypox and skin disease detection," Scientific Reports, vol. 15, no. 1, pp. 32924, Sep. 2025, doi: 10.1038/s41598-025-12216-y.
[14] A. H. Thieme et al., "A deep-learning algorithm to classify skin lesions from mpox virus infection," Nat. Med., vol. 29, no. 3, pp. 738–747, Mar. 2023, doi: 10.1038/s41591-023-02225-7.
[15] [S. Islam, F. A. Nishi, T. Akter, and M. A. Azim, "Monkeypox Skin Lesion Detection with Deep Learning and Machine Learning," Int. J. Comput. Appl., vol. 185, no. 23, pp. 39–45, Jul. 2023, doi: 10.5120/ijca2023922984.
[16] G. M. Idroes, T. R. Noviandy, T. Bin Emran, and R. Idroes, "Explainable Deep Learning Approach for Mpox Skin Lesion Detection with Grad-CAM," Heca Journal of Applied Sciences, vol. 2, no. 2, pp. 54–63, Sep. 2024, doi: 10.60084/hjas.v2i2.216.
[17] T. M. Chiu, I. C. Chi, Y. C. Li, and M. H. Tseng, "Deep Ensemble Learning for Multiclass Skin Lesion Classification," Bioengineering, vol. 12, no. 9, p. 934, Sep. 2025, doi: 10.3390/bioengineering12090934.
[18] A. Triwerdaya and E. Utami, "Hybrid ViT–CNN Model for Automatic Monkeypox Skin Lesion Diagnosis," Journal of Electrical Engineering and Computer (JEECOM), vol. 7, no. 2, pp. 461–470, Oct. 2025, doi: 10.33650/jeecom.v7i2.12795.
[19] M. S. Hossain, M. Ahmed, and M. S. Rahman, "From survey to solution: A deep learning framework for reliable monkeypox diagnosis using skin images," Array, vol. 28, no. 1, p. 100554, Dec. 2025, doi: 10.1016/j.array.2025.100554.
[20] A. Adebiyi et al., "Transformers in Skin Lesion Classification and Diagnosis: A Systematic Review," medRxiv, p. 2024.09.19.24314004, Sep. 2024, doi: 10.1101/2024.09.19.24314004.
[21] P. Bodagala and N. M. Dubba, "Comprehensive and Integrated Deep Learning Approaches-based Intelligent Skin Disease Diagnosis System for Pox Detection," ASEAN Journal of Scientific and Technological Reports, vol. 29, no. 1, pp. e260160–e260160, Dec. 2026, doi: 10.55164/ajstr.v29i1.260160.