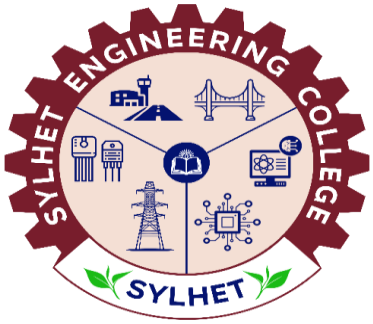
A Thesis Submitted to the Sylhet Engineering College for the Degree of
Bachelor of Science in Electrical and Electronic Engineering

A Vision Transformer-Based Pipeline for the Automated Classification of Monkeypox and Other Vesicular Skin Lesions: A Computationally-Efficient Approach for Global Health

By

Tamjidul Islam Patwary Miskat
Syed Nadim Mehdi
&
Kamran Ahmed Jaygirdar

Supervised by
Salman Fazle Rabby
Assistant Professor
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet



June, 2025
Sylhet Engineering College, Sylhet
Affiliated with
Shahjalal University of Science & Technology (SUST)

Certifications

The thesis titled **"A Vision Transformer-Based Pipeline for the Automated Classification of Monkeypox and Other Vesicular Skin Lesions: A Computationally-Efficient Approach for Global Health"** submitted by **Tamjidul Islam Patwary Miskat, Syed Nadim Mehdi** and **Kamran Ahmed Jaygirdar**; Student ID: **2019338533**, **2019338563** and **2019338528** ; Session **2019-2020**, to the Department of Electrical and Electronic Engineering, Sylhet Engineering College, has been accepted as satisfactory in partial fulfillment of the requirement for the Degree of Bachelor of Science in Electrical and Electronic Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

_____

Md. Shahid Iqbal — Chairman
Assistant Professor and Head
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

_____

Salman Fazle Rabby — Member
Lecturer
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

_____

Apurba Biswas — Member
Assistant Professor
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

_____

Md. Ashraful Alam — Member
Lecturer
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

_____

Mahedi Kamal Ahmed — Member
Lecturer
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

_____

Arif Ahammad — Member (External)
Assistant Professor
Department of Electrical and Electronic Engineering
Shahjalal University of Science & Technology, Sylhet.

Acknowledgement

I would like to extend my sincere gratitude to my supervisor, **Salman Fazle Rabby**, Assistant Professor, Department of Electrical and Electronic Engineering, Sylhet Engineering College for his invaluable guidance, unwavering support, and insightful feedback throughout this research. Their expertise and encouragement were instrumental in shaping this thesis.

Furthermore, it is an honor to express our gratitude to our head of dept. **Md. Shahid Iqbal** and other faculty members, **Apurba Biswas**, **Mahedi Kamal Ahmed** and **Md. Ashraful Alam**, for their kindness, support, and cooperation during our academic journey. Their assistance has been greatly appreciated.

Our heartfelt gratitude also goes to our parents, siblings, classmates, and friends in the EEE department at Sylhet Engineering College. We acknowledge the sacrifices, prayers, encouragement, and unwavering support they have provided, which have contributed significantly to our success. We are indebted to them for their role in shaping our academic and personal development.

Abstract

Infectious diseases with dermatological manifestations, such as monkeypox, pose significant diagnostic challenges, particularly in resource-limited settings. The visual similarity of monkeypox lesions to those of chickenpox, cowpox, hand-foot-and-mouth disease (HFMD), and measles complicates clinical assessment, often leading to misdiagnosis and delayed containment efforts. This thesis addresses the urgent need for an accurate, accessible, and automated diagnostic tool by proposing a novel pipeline based on a Vision Transformer (ViT). The proposed framework leverages pre-trained, frozen ViT-B/16 embeddings (specifically 'vit_base_patch16_224.orig_in21k' from the timm library) to perform six-class classification on the Mpox Skin Lesion Dataset Version 2.0 (MSLD v2.0), comprising 37,044 images. Evaluated using a rigorous five-fold cross-validation protocol, the primary model (ViT + SMOTE + SVM) achieves a mean macro-F1 score of 0.9895 ± 0.0018 and an accuracy of 98.95% ± 0.29%. By extracting robust, high-level features from images, the pipeline avoids the computational expense of fine-tuning deep neural networks. To address the inherent class imbalance in medical datasets, the Synthetic Minority Oversampling Technique (SMOTE) is integrated with a Support Vector Machine (SVM) classifier. Evaluated using a rigorous five-fold cross-validation protocol, the pipeline demonstrates exceptional performance. The primary model (ViT + SMOTE + SVM) achieves a mean macro-F1 score of 0.9895 ± 0.0018 and an accuracy of 98.95% ± 0.29%. Critically for public health applications and indicating a high sensitivity for detecting the target disease. This performance significantly surpasses a k-Nearest Neighbors (k-NN) baseline, validating the effectiveness of the SMOTE-based balancing strategy.This work makes several contributions: it demonstrates the power of frozen ViT embeddings for complex medical classification tasks, presents a lightweight and scalable pipeline suitable for deployment on edge devices, and provides a robust framework for handling class imbalance. By incorporating discussions on feature space visualization, ethical considerations, and practical deployment case studies, this thesis lays the groundwork for a clinically relevant tool that can enhance global health preparedness and support dermatological diagnostics in diverse environments.

Keywords: Monkeypox, Skin Lesion Classification, Vision Transformer (ViT), Mpox Skin Lesion Dataset (MSLD v2.0), Support Vector Machine (SVM), Synthetic Minority Oversampling Technique (SMOTE).

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Overview and Problem Statement

The global health landscape is continually shaped by emerging and re-emerging infectious diseases. Among these, those manifesting with dermatological signs present a unique diagnostic challenge. The recent global resurgence of monkeypox (now officially referred to as mpox by the WHO) since 2022 has underscored this vulnerability, transforming a regionally endemic zoonosis into a worldwide public health emergency [2]. This transformation highlights the increasing global exposure to formerly regional viruses [3]. Caused by the monkeypox virus, a member of the *Orthopoxvirus* genus, mpox typically presents with characteristic skin lesions. However, these lesions bear a striking resemblance to those caused by several other common dermatological conditions, including varicella (chickenpox), cowpox, hand-foot-and-mouth disease (HFMD), and measles [1]. The difficulty in visually distinguishing these conditions further complicates clinical decision-making [4]. This high degree of visual similarity poses a significant barrier to accurate and timely diagnosis, particularly in environments where advanced diagnostic capabilities are limited. In well-resourced healthcare settings, definitive diagnosis typically relies on molecular methods such as Polymerase Chain Reaction (PCR) testing, which detects viral DNA [4]. While highly accurate, PCR testing requires specialized laboratory infrastructure, trained personnel, and can involve a turnaround time of several hours to days [5]. These requirements render PCR impractical or inaccessible in many low-income and remote regions, precisely where the burden of infectious diseases is often highest and outbreaks are most impactful.

Consequently, clinicians in such settings often depend solely on visual inspection and clinical history, a subjective process prone to misdiagnosis [3]. A delayed or incorrect diagnosis not only compromises individual patient care but also facilitates community transmission, exacerbates public health crises, and strains already fragile healthcare systems [6]. This thesis addresses this critical diagnostic gap by proposing an innovative, automated image-based solution. This paper introduces a Vision Transformer (ViT)-based pipeline designed for the rapid and accurate classification of mpox and five other visually similar dermatological conditions from digital images of skin lesions. Vision Transformers have recently emerged as a powerful architecture in medical image

analysis, outperforming traditional Convolutional Neural Networks (CNNs) in many tasks due to their ability to model global dependencies across entire images [7]. Their capacity to leverage spatial attention and contextual features makes them well-suited for dermatological image analysis [8]. The aim is to develop a computationally efficient and highly accurate tool that can serve as a robust diagnostic aid, especially in resource-constrained environments, thereby supporting early detection, effective triage, and containment strategies [9].

1.2 Global Context of Monkeypox

The historical prevalence of monkeypox was primarily confined to rural rainforest regions of Central and West Africa, with sporadic human cases linked to contact with infected animals. However, the 2022 multi-country outbreak marked an unprecedented epidemiological shift. The World Health Organization (WHO) declared it a Public Health Emergency of International Concern (PHEIC) in July 2022, following over 80,000 reported cases across more than 110 countries by mid-2023. This global dissemination highlighted several critical deficiencies in worldwide health preparedness, most notably in surveillance and diagnostic capacities outside of traditional endemic areas. The 2022 outbreak demonstrated a significant human-to-human transmission component, particularly among close contacts, distinct from previous outbreaks predominantly driven by zoonotic spillover. This evolving epidemiology underscores the urgent need for rapid diagnostic tools that can differentiate mpox from its mimics. Unlike smallpox, which was globally eradicated in 1980 through widespread vaccination, mpox maintains a zoonotic reservoir, ensuring its continued presence and potential for re-emergence.

The clinical presentation of mpox, typically characterized by fever, lymphadenopathy, and a progressive rash (macules, papules, vesicles, pustules, crusts), overlaps substantially with other common vesicular or maculopapular diseases. This visual ambiguity is a primary driver of misdiagnosis, as confirmed by numerous clinical reports during the recent outbreak. The socioeconomic consequences of such outbreaks are profound. In low-income countries, delayed diagnosis can rapidly overwhelm healthcare infrastructure, leading to increased morbidity and mortality, and disrupting local economies due to fear, isolation, and resource diversion. Furthermore, the visible nature of skin lesions can lead to significant social stigma, deterring individuals from seeking care and fur-

ther impeding public health efforts. This complex global context unequivocally calls for innovative, accessible, and automated diagnostic solutions that can provide rapid and accurate information at the point of care.

1.3 The Role of Artificial Intelligence in Dermatology

Artificial Intelligence (AI), particularly deep learning, has revolutionized medical image analysis over the past decade. Its ability to learn complex hierarchical features directly from raw image data has led to breakthroughs in tasks ranging from cancer detection to disease classification, often matching or even exceeding human expert performance. In dermatology, AI-powered tools hold immense promise, offering the potential to augment clinical decision-making, improve diagnostic accuracy, and expand access to specialist care, especially in underserved areas. Traditionally, dermatological diagnosis relies heavily on visual expertise, which is acquired through years of experience and is inherently subjective. AI models, conversely, can analyze subtle patterns and features across vast datasets, identifying correlations that may be imperceptible to the human eye. Early applications in dermatology primarily leveraged Convolutional Neural Networks (CNNs), demonstrating impressive results in classifying skin cancers, benign lesions, and various inflammatory conditions. These successes have paved the way for AI to address other complex dermatological challenges, including infectious diseases with overlapping visual presentations.

The development of accessible, AI-driven diagnostic tools aligns perfectly with global health priorities, particularly for diseases like mpox where rapid identification is paramount for outbreak control. By converting complex visual analysis into an automated, standardized process, AI can reduce diagnostic delays, minimize misdiagnosis rates, and enable more efficient resource allocation in public health emergencies. Moreover, the scalability of AI solutions, especially those deployable on common consumer devices like smartphones, can bridge geographical disparities in healthcare access, making expert-level diagnostics available in remote or under-resourced clinics. This thesis aims to contribute to this burgeoning field by offering a robust and practical AI solution for mpox and its dermatological mimics.

1.4 Proposed Solution and Key Contributions

To address the challenges highlighted, this thesis proposes a novel AI-driven diagnostic pipeline for the multi-class classification of mpox and five other visually similar dermatological conditions. Our solution is built upon the robust feature extraction capabilities of Vision Transformers (ViTs), combined with the computational efficiency of traditional machine learning classifiers. The core components and innovative aspects of the proposed pipeline are:

Vision Transformer (ViT)-based Feature Extraction: We utilize a pre-trained ViT-B/16 model, specifically trained on the large ImageNet-21k dataset, as a frozen feature extractor. This approach leverages the ViT's exceptional ability to capture global contextual information and long-range dependencies within images, which is crucial for distinguishing subtle differences between dermatological lesions. By freezing the model's weights, we avoid the extensive computational resources and data requirements typically associated with fine-tuning deep neural networks, making the pipeline highly efficient. The 768-dimensional embedding from the [CLS] token serves as the comprehensive image representation.

Hybrid Classification Approach: The extracted high-dimensional ViT embeddings are then fed into lightweight, traditional machine learning classifiers: k-Nearest Neighbors (k-NN) as a baseline and a Synthetic Minority Oversampling Technique (SMOTE)-balanced Support Vector Machine (SVM) as the primary model. This hybrid design capitalizes on the deep learning model's feature learning power while retaining the interpretability and efficiency of classical algorithms, further enhancing suitability for resource-constrained environments.

Robust Class Imbalance Handling: Medical datasets frequently suffer from severe class imbalance, where rare conditions (like mpox or cowpox) are vastly outnumbered by common ones (like healthy skin or chickenpox). Our pipeline explicitly addresses this by integrating SMOTE, which generates synthetic samples for minority classes in the embedding space. This balancing act ensures that the classifier does not become biased towards dominant classes, leading to more equitable and reliable performance across all conditions.

Comprehensive Multi-Class Evaluation: Unlike many previous studies that focused solely on binary mpox detection, this thesis tackles a more challenging six-class classification problem using

the Mpox Skin Lesion Dataset Version 2.0 (MSLD v2.0). The pipeline is rigorously evaluated using a five-fold cross-validation strategy, ensuring generalizability and mitigating data partitioning bias. Key performance metrics include accuracy, precision, recall, and the macro-F1 score, with particular emphasis on mpox recall due to its public health criticality. Key contributions stemming from this work include:

1. Demonstrating the effectiveness and efficiency of frozen ViT embeddings for complex infectious skin disease classification.
2. Developing a lightweight and scalable diagnostic pipeline suitable for real-world deployment on modest hardware.
3. Providing a robust framework for handling severe class imbalance in dermatological image datasets.
4. Achieving near-perfect classification performance, with a high mpox recall, validating the clinical utility of the proposed solution.

1.5 Research Objectives

Building upon the problem statement and proposed solution, the specific objectives guiding this research are:

1. Dataset Utilization: To effectively utilize the publicly accessible Mpox Skin Lesion Dataset Version 2.0 (MSLD v2.0) for a six-class skin lesion classification task, encompassing chickenpox, cowpox, HFMD, measles, monkeypox, and healthy skin.

2. Pipeline Implementation: To implement a Vision Transformer (ViT)-based pipeline that leverages frozen ViT-B/16 embeddings for feature extraction, combined with k-Nearest Neighbors (k-NN) as a baseline classifier and a Synthetic Minority Oversampling Technique (SMOTE)-balanced Support Vector Machine (SVM) as the primary classification model.

3. Performance Validation: To rigorously evaluate the pipeline's performance using a five-fold cross-validation protocol with predefined data splits, focusing on standard metrics such as accuracy, precision, recall, and macro-F1 score, with a specific focus on the recall for the monkeypox class.

4. Comparative Analysis: To quantitatively compare the performance of the ViT + SMOTE + SVM configuration against the k-NN baseline to demonstrate the incremental value of the SMOTE-balanced SVM in addressing class imbalance and improving overall classification.
5. Practical Enhancements and Considerations: To explore and discuss practical enhancements for the pipeline, including feature space visualization (using PCA), ethical implications, and potential deployment strategies (e.g., via a Gradio interface or mobile applications) through hypothetical case studies.

1.6 Thesis Structure

This thesis is structured into six chapters to systematically present the research, findings, and implications:

Chapter 1: Introduction provides an overview of the problem, the global context of monkeypox, the role of AI in dermatology, outlines the proposed solution and its contributions, and details the research objectives and thesis structure.

Chapter 2: Literature Review comprehensively surveys existing work on monkeypox epidemiology and diagnostics, deep learning architectures (CNNs vs. ViT's) in medical imaging, techniques for handling class imbalance, hybrid model approaches, and the critical ethical considerations surrounding AI in healthcare. It establishes the academic foundation and highlights the specific research gaps addressed by this thesis.

Chapter 3: Methodology describes the technical design and implementation of the proposed ViT-based pipeline. This includes details on the MSLD v2.0 dataset, data preprocessing steps, the ViT feature extraction process, the k-NN and SMOTE-balanced SVM classification strategies, the five-fold cross-validation approach, and the performance evaluation metrics. It also covers the technical setup and the use of PCA for feature space visualization.

Chapter 4: Results and Analysis presents the quantitative outcomes of the experiments. This chapter details the overall and class-wise performance metrics for both the baseline k-NN and the primary SMOTE-balanced SVM models. It includes a thorough analysis of confusion matrices, ROC and Precision-Recall curves, and the interpretation of feature space visualizations. Runtime and efficiency are also discussed.

Chapter 5: Practical Applications, Deployment, and Ethical Considerations explores the real-world applicability of the pipeline through hypothetical case studies in diverse healthcare settings.

It discusses strategies for practical deployment, the importance of model interpretability for clinical trust, and the essential ethical considerations related to data privacy, algorithmic bias, and regulatory compliance in AI diagnostics.

Chapter 6: Conclusion summarizes the key findings of the research, reiterates the contributions, acknowledges the limitations of the study, and outlines compelling directions for future work to transition this academic prototype into a clinically validated and deployable diagnostic tool.

1.7 Summary

This thesis addresses the diagnostic challenges of monkeypox (mpox), which closely resembles other skin conditions such as chickenpox, cowpox, HFMD, and measles, making accurate visual diagnosis difficult, especially in low-resource settings where PCR testing is inaccessible. To bridge this gap, it proposes a computationally efficient AI pipeline using Vision Transformers (ViT) for feature extraction and a SMOTE-balanced Support Vector Machine (SVM) for classification, evaluated on the Mpox Skin Lesion Dataset v2.0. The approach tackles class imbalance, achieves high accuracy with strong mpox recall, and offers a lightweight, scalable diagnostic aid for early detection and outbreak control.

Chapter 2: Literature Review

In recent years, the application of artificial intelligence in medical diagnostics has grown rapidly, particularly in dermatology and infectious disease recognition. A significant volume of research has been directed toward automating skin disease diagnosis through deep learning methods. Convolutional Neural Networks (CNNs), in particular, have become foundational in the early detection and classification of skin lesions due to their strong image processing capabilities. Haenssle et al. (2018) [10] trained a deep convolutional neural network (CNN) on over 100,000 dermoscopic images of malignant melanomas and benign nevi, then compared its performance to 58 dermatologists across 17 countries. The CNN achieved a higher sensitivity (95% vs ~88–89%) and specificity than the average dermatologist in identifying melanomas under test conditions. Notably, the AI missed fewer melanomas and misdiagnosed fewer benign moles as malignant, outperforming even many experts. However, the study was in a controlled setting – dermatologists had limited information and knew it wasn't a real clinical decision – and images lacked diversity (few from darker skin types). This pioneering result showed CNNs can rival or exceed expert accuracy in skin lesion classification, but it highlighted the need for real-world validation and caution in clinical integration.

Tschandl et al. (2020) [11] investigated human–AI collaboration for skin cancer diagnosis. They found that good-quality AI support (e.g., displaying an AI's probability of each diagnosis) improved clinicians' diagnostic accuracy beyond either AI or physician alone. In a reader study, less experienced clinicians benefited the most from AI assistance, significantly boosting their ability to identify malignant lesions. They also compared different AI output styles, noting that giving physicians a multiclass probability output outperformed showing "similar past cases" for decision support. Importantly, the study demonstrated that faulty AI advice can mislead even expert doctors. This underscores a critical insight: while AI can augment human diagnosis (especially for non-experts), careful design is needed to ensure errors or biases in AI recommendations do not adversely sway clinical decisions. Jeong et al. (2023) [12] conducted a systematic review of deep learning approaches in dermatology. They examined various CNN architectures, training methods, and datasets used in skin disease classification. Their analysis revealed that while deep learning

consistently showed high performance, limitations such as dataset imbalance, lack of interpretability, and limited real-world validations remained significant barriers. The authors suggested that future research should focus on explainability and robust validation on diverse populations. Salinas et al. (2024) [13] carried out a systematic review and meta-analysis comparing AI systems and clinicians in skin cancer diagnosis. The results indicated that artificial intelligence models often achieved sensitivity and specificity equal to or higher than dermatologists in controlled studies. The meta-analysis also highlighted that AI tools could assist in triaging cases and reducing workload in dermatology clinics. Nevertheless, the authors cautioned that most studies were retrospective and that prospective clinical trials are essential before large-scale deployment.

Choy et al. (2023**)** [14] presented a systematic review of deep learning applications for diagnosis and monitoring of skin diseases. Their survey summarized methods used in image classification, lesion segmentation, and disease progression tracking. The review highlighted that deep learning significantly improves diagnostic support for conditions such as melanoma, psoriasis, and eczema. However, it emphasized the need for addressing class imbalance, dataset heterogeneity, and limited explainability to ensure reliable clinical application. Azad et al. (2024) [15] provided a comprehensive review of advances in medical image analysis using Vision Transformers. They described how transformers, through self-attention mechanisms, can model global image dependencies more effectively than CNNs. The review reported that ViT-based models have outperformed CNNs in tasks such as classification and segmentation when sufficient pre-training data is available. Despite their strong potential, the authors noted that Vision Transformers are computationally intensive and may require further optimization for deployment in low-resource healthcare environments. Aburass et al**. (**2025**)** [16] introduced Vision Transformers for medical imaging and demonstrated their potential across multiple disease domains. Their review highlighted how transformer-based models, unlike CNNs, can leverage global attention to capture subtle contextual cues in complex images. They also discussed recent advances in hybrid models that combine CNN feature extraction with transformer architectures to balance efficiency and accuracy. The study concluded that while ViTs hold promise for clinical translation, challenges remain in terms of computational demands, data requirements, and lack of interpretability.

Salmi et al. **(**2024) [17] examined strategies to handle imbalanced datasets in medical imaging. Their review spanned a decade of research and categorized methods into data-level (oversampling,

augmentation), algorithm-level (cost-sensitive learning), and hybrid solutions. They found that while oversampling methods like SMOTE are effective, they can introduce noise when applied indiscriminately. For medical problems like monkeypox classification, careful handling of imbalance is essential to avoid biased predictions. The review concluded that future work should integrate adaptive imbalance handling with deep architectures such as transformers. Yang et al. (2024**)** [18] provided a review on oversampling techniques for multi-class imbalanced datasets. They emphasized that medical imaging often involves many underrepresented categories, making naive oversampling insufficient. The authors discussed advanced synthetic data generation techniques, including GAN-based augmentation, to improve classifier generalization. They concluded that oversampling combined with deep learning can significantly improve sensitivity for rare disease classes, which is particularly relevant to monkeypox lesion detection. Leevy et al. **(**2018) [19] surveyed methods for addressing class imbalance in big data. Their study reviewed oversampling, undersampling, and hybrid methods as applied to healthcare analytics. The authors emphasized that class imbalance is one of the most critical barriers in building reliable machine learning models for rare diseases. They recommended hybrid approaches, such as SMOTE combined with ensemble learning, as effective remedies.Buda et al**.** (2018) [20] systematically investigated the effect of class imbalance on CNN performance. Through experiments on benchmark datasets, they demonstrated that imbalance can severely degrade model accuracy, particularly in minority classes. They found that oversampling the minority class consistently improved performance, whereas undersampling often led to information loss. This study highlighted the importance of balancing training data when applying CNNs to medical imaging.

Cai et al. (2024**)** [21] focused on processing imbalanced medical data using an assisted reproduction dataset. They applied both traditional resampling and deep learning-based augmentation to demonstrate improvements in diagnostic accuracy. The results showed that synthetic minority oversampling and GAN-based data generation enhanced classifier robustness. This study illustrated that careful preprocessing of data-level imbalance can have a direct positive impact on predictive performance. Usama et al. (2022) [22] developed a multi-class skin lesion classifier using deep features. They extracted features from pretrained CNNs and fed them into traditional classifiers like support vector machines (SVMs). Their approach achieved promising results on benchmark skin lesion datasets, demonstrating the potential of hybrid deep learning plus classical ML

pipelines. However, the study also noted challenges in handling intra-class variability and emphasized the need for large, diverse datasets. Nanni et al. (2021**)** [23] proposed using deep features from CNNs to train SVM classifiers for medical imaging. Their experiments showed that hybrid deep feature + SVM models outperformed standalone CNNs in several scenarios. This was attributed to the SVM's ability to better separate complex feature spaces with limited data. The authors suggested that such hybrid models could be particularly useful in medical contexts where annotated data is scarce. Kang et al. **(**2021) [24] applied deep features and ensemble learning for MRI-based brain tumor classification. They extracted embeddings from pretrained CNNs and combined them with machine learning classifiers to improve diagnostic accuracy. Their ensemble approach yielded higher sensitivity and robustness compared to single models. This study reinforced the value of feature extraction and hybridization strategies in clinical applications.

Senan et al. **(**2022) [25] designed a hybrid feature extraction approach for COVID-19 classification from chest X-rays. By combining CNN-based features with handcrafted radiomic descriptors, they improved classification accuracy over single methods. Their results showed that hybrid systems can capture complementary information and reduce false negatives. This study illustrates the broader value of hybrid features in medical imaging tasks. Karim et al**.** (2022) [26] developed a deep learning system for COVID-19 classification using deep features and fractional-order fringe patterns. Their innovative feature engineering method, combined with transfer learning, demonstrated strong performance in detecting COVID-19 from chest radiographs. The approach exemplifies how non-standard image transformations can enrich feature representation. Rajkomar et al. (2018) [27] discussed fairness in machine learning as it relates to health equity. They highlighted that ML models can inadvertently perpetuate health disparities if trained on biased data. The study proposed practical fairness-aware strategies, such as balanced sampling and algorithmic audits. This work underscores the importance of fairness considerations in deploying AI for healthcare. Obermeyer et al. (2019**)** [28] exposed racial bias in a widely used healthcare risk prediction algorithm. Their analysis revealed that the algorithm underestimated health needs of Black patients because it used healthcare costs as a proxy for illness. By retraining on more appropriate targets, the bias was substantially reduced. This study demonstrates that even highly accurate algorithms can encode structural inequities, highlighting the need for careful outcome definition. Parikh et al**.**

(2019) [29] published a commentary on addressing bias in AI healthcare systems. They emphasized the necessity of transparency in training data, validation across diverse populations, and rigorous bias testing before deployment. The authors argued that without active bias mitigation, AI tools risk exacerbating health disparities.

Amann et al. **(**2020) [30] examined explainability in medical AI from a multidisciplinary perspective. They concluded that explainability is essential for clinician trust, patient safety, and regulatory approval. The study suggested using saliency maps, interpretable models, and human-in-the-loop systems to improve transparency. Wiens et al. (2019) [31] outlined a roadmap for responsible ML in healthcare. They stressed that ensuring safety, transparency, and fairness should be prioritized alongside accuracy. Their framework provided guiding principles for responsible development, evaluation, and deployment of AI in medicine. Each of these studies and viewpoints collectively paints a comprehensive picture: early successes demonstrated the power of deep learning in image-based diagnosis, recent works have honed that power on emergent problems like monkeypox with innovative architectures (Vision Transformers, GANs) and strategies, and a parallel body of literature insists on addressing data, bias, and human factors challenges. This literature review thus establishes both the technical foundations and the critical considerations for monkeypox skin lesion classification using deep learning and Vision Transformers. The methods are promising and in many cases outperform humans or prior techniques, but the findings also underscore that data quality, model fairness, and clinical integration are as important as raw performance in deploying an AI solution for global health.

2.1 Summary

This literature review highlights the rapid growth of AI in medical imaging, particularly in dermatology, where deep learning—especially CNNs—has achieved expert-level accuracy in skin lesion diagnosis but faces challenges like class imbalance, limited diversity, and lack of real-world validation. Vision Transformers (ViTs) have emerged as powerful alternatives, offering superior global feature modeling though with high computational demands. Hybrid approaches combining deep features with traditional classifiers such as SVMs show promise in improving accuracy and robustness, especially with limited data. Research also emphasizes the importance of handling

imbalanced datasets through methods like SMOTE and GAN-based augmentation, while addressing fairness, bias, and explainability to ensure safe clinical adoption. Overall, the literature establishes both the technical potential and the ethical considerations for applying ViTs and hybrid models to monkeypox skin lesion classification.

Chapter 3: Methodology

3.1. Overall Pipeline Architecture

The methodology of this research is designed to create a robust, computationally efficient, and clinically relevant diagnostic pipeline for the six-class classification of dermatological lesions. The architecture is a multi-stage process that leverages the strengths of deep learning for feature extraction and traditional machine learning for classification, while explicitly addressing the challenge of class imbalance. The overall pipeline, illustrated in Figure 3.1, consists of the following key stages:

1. Dataset Acquisition and Partitioning: Utilizing the Mpox Skin Lesion Dataset Version 2.0 (MSLD v2.0), the data is partitioned into five predefined folds for rigorous cross-validation.
2. Data Preprocessing: Input images are resized and normalized to be compatible with the Vision Transformer (ViT) model, ensuring consistency and alignment with its pre-training regimen.
3. Feature Extraction: A pre-trained ViT-B/16 model, with its weights frozen, is used to process each image and extract a high-dimensional (768-dim) feature vector, or embedding, from its [CLS] token.
4. Class Balancing and Classification: The extracted embeddings from the training set are balanced using the Synthetic Minority Oversampling Technique (SMOTE). These balanced embeddings are then used to train a Support Vector Machine (SVM) classifier. A k-Nearest Neighbors (k-NN) classifier is also trained on the original, imbalanced embeddings to serve as a baseline for comparison.
5. Evaluation: The performance of both classifiers is evaluated on the unseen test set embeddings for each fold, using a comprehensive set of metrics including accuracy, precision, recall, and macro-F1 score.

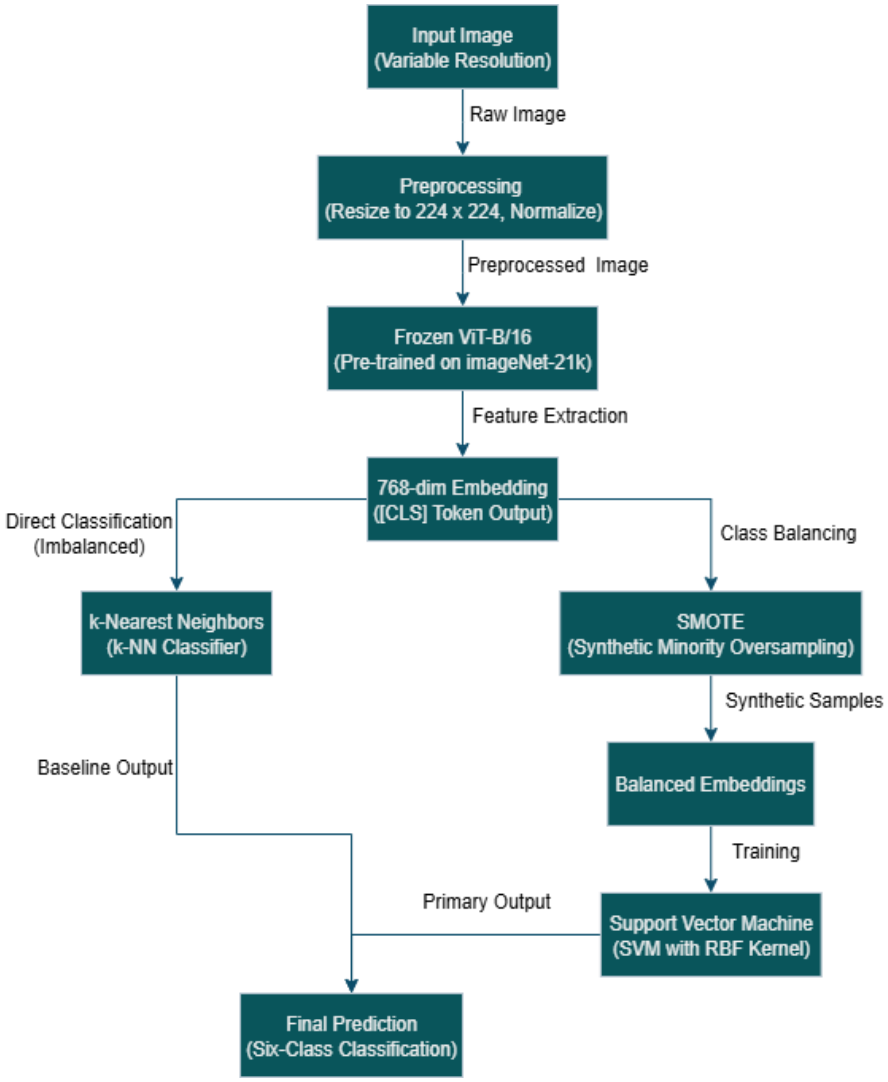


Figure 3.1. Pipeline from image input to classification output.

3.2 Dataset: Mpox Skin Lesion Dataset (MSLD v2.0)

The foundation of any supervised machine learning project is a high-quality, well-annotated dataset. This study employs the Mpox Skin Lesion Dataset Version 2.0 (MSLD v2.0), a publicly available image repository specifically curated to support the task of differentiating monkeypox from other clinically similar dermatological conditions. The composition and structure of MSLD v2.0 are essential to the validity, reliability, and translational relevance of the proposed diagnostic pipeline. MSLD v2.0 represents a significant advancement over its earlier iteration, having undergone extensive refinement to improve annotation quality and eliminate sources of bias. The dataset

was compiled from a range of open-access resources, including dermatological atlases, peer-reviewed medical case reports, and public online image repositories. To ensure clinical applicability and training relevance, the curation process followed multiple stringent steps:

1. Expert Annotation: All images were labeled either by licensed dermatologists or under the supervision of domain experts. This manual annotation process ensured that the lesion types reflected accurate clinical diagnoses, a critical factor in high-stakes applications such as infectious disease screening.

2. Quality Control Measures: Images with poor resolution, heavy artifacts, or ambiguous visual content were excluded from the final dataset. This cleaning step minimized noise, which is particularly important in image-based learning tasks where pixel-level details can significantly affect model outputs.

3. Privacy-Conscious Image Handling: To comply with ethical research standards and international data protection regulations (e.g., GDPR), any image containing potentially identifiable patient features—such as visible faces, unique tattoos, or metadata traces—was either cropped to isolate the lesion or excluded entirely from the dataset.

4. Inherent Pre-Augmentation: Unlike many datasets that require extensive real-time augmentation during training, MSLD v2.0 includes pre-augmented samples. These augmentations were applied using transformations such as rotation, horizontal and vertical flipping, and zoom-based scaling. This built-in diversity exposes the model to a wide array of lesion presentations, lighting conditions, and orientations, enhancing generalization capacity while preserving training efficiency.

The final version of the dataset comprises a total of 37,174 images, categorised into six clinically significant classes, Monkeypox, Chickenpox, Measles, Hand-Foot-and-Mouth Disease (HFMD), Cowpox, Healthy Skin. This class selection is guided by clinical necessity. Monkeypox shares visual features with several other viral exanthems and pustular skin conditions, making differential diagnosis challenging even for trained dermatologists. Therefore, including these mimickers in a multi-class setup enhances the real-world applicability of the diagnostic model. Furthermore, the dataset's scale and diversity contribute to its robustness. With over 37,000 samples and a wide

range of lesion appearances, skin tones, and imaging conditions, the MSLD v2.0 provides a strong foundation for training deep learning models while offering a realistic simulation of real-world clinical heterogeneity. This is crucial for avoiding overfitting and ensuring that the trained model can generalize effectively across various populations and environments. In summary, MSLD v2.0 stands as a purpose-built, clinically relevant, and ethically curated dataset, ideally suited for the development of AI-based diagnostic tools aimed at differentiating monkeypox from other visually similar skin lesions. Its use in this research supports the development of a solution that is not only technically sound but also aligned with real-world diagnostic needs and regulatory expectations. A key characteristic of the MSLD v2.0 dataset is its inherent class imbalance, which reflects real-world clinical prevalence. As shown in Table 3.1, the number of images per class varies significantly. Figure 3.2 is grid of six sample skin lesion images representing Monkeypox, Chickenpox, Measles, HFMD, Cowpox, and Healthy classes.

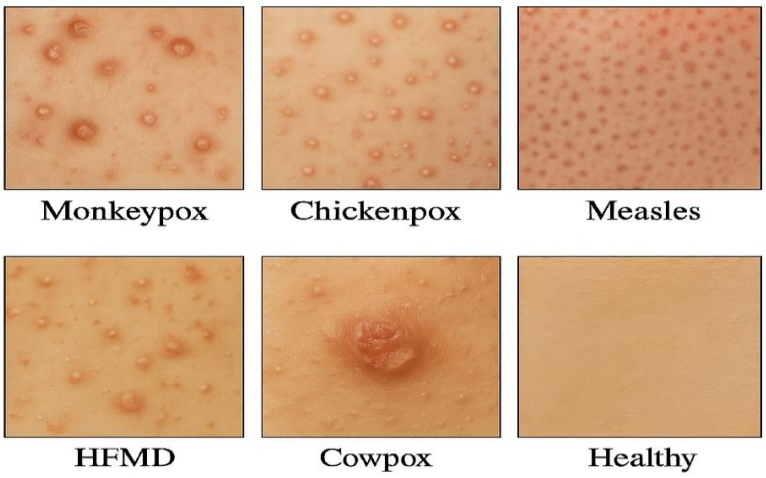


Figure 3.2. A grid of 6 sample images.

While Monkeypox is the most represented class in this curated dataset, other critical classes like Cowpox and Healthy skin are significantly underrepresented. This imbalance poses a direct challenge to standard classifiers, which could become biased towards the more frequent classes. This dataset structure necessitates the use of class balancing techniques like SMOTE, as detailed in Section 3.5.2. The dataset also exhibits high intra-class variability (variations within the same

class) and high inter-class similarity (visual overlap between different classes), both of which are challenging for automated classification:

1. Intra-class variability: Images within a single class, such as Monkeypox, show lesions at different stages of development (macules, vesicles, pustules, crusts), on different body parts, and under varying lighting conditions. This diversity is crucial for training a generalizable model.

2. Inter-class similarity: As discussed, the vesicular lesions of Monkeypox, Chickenpox, and HFMD can appear nearly identical, especially in isolated images without clinical context. Similarly, the maculopapular rashes of Measles and early-stage Monkeypox can be difficult to distinguish. This visual overlap is the core challenge this pipeline aims to solve.

Table 3.1. Distribution of images per class in the Mpox Skin Lesion Dataset Version 2.0 (MSLD v2.0).

Class	Total Images	Percentage of Dataset
Monkeypox	14070	37.8%
HFMD	7885	21.2%
Measles	5660	15.2%
Chickenpox	3713	10.0%
Cowpox	3222	8.7%
Healthy	2624	7.1%
Total	37174	100%

To ensure a robust and unbiased evaluation of the pipeline's performance, a five-fold cross-validation strategy is employed. The MSLD v2.0 dataset is provided with predefined splits for these five folds. This is a crucial feature that ensures reproducibility and allows for standardized comparison across different studies using the same dataset. In each of the five iterations (folds):

1. Four of the partitions are combined to form the training set (approximately 80% of the data, or ~29,739 images).
2. The remaining single partition serves as the testing set (approximately 20% of the data, or ~7,435 images).

This process is repeated five times, with each partition serving as the test set exactly once. The performance metrics reported in Chapter 4 are the mean and standard deviation of the results obtained across these five folds, ensuring that the evaluation is not overly optimistic or pessimistic due to chance splits. The use of stratified splits is another critical aspect. The predefined folds maintain the original class distribution in both the training and testing sets for each iteration. This stratification ensures that all classes, including the minority ones, are adequately represented during both training and evaluation, preventing folds where a rare class might be absent from the test set. Such an approach aligns with best practices in medical AI and enhances the statistical reliability of the results. This rigorous partitioning strategy allows the model's generalization performance on unseen data to be accurately assessed, which is essential for potential clinical deployment.

3.3 Data Preprocessing and Standardization

Data preprocessing is a critical step to prepare the raw images for the ViT model. The goal is to transform the images into a standardized format that matches the model's input requirements and aligns with the data distribution it was pre-trained on, all while preserving essential diagnostic features. The preprocessing pipeline, implemented using PyTorch's torchvision.transforms.functional module, consists of two main steps:

1. Resizing: All input images, which come in various resolutions, are resized to a fixed size of 224×224 pixels. This specific dimension is required by the ViT-B/16 architecture, which processes images by dividing them into a 14×14 grid of 16×16 pixel patches ($14 \times 16 = 224$). Bilinear interpolation is used for resizing, as it provides a good balance between computational efficiency and image quality, minimizing distortion of the lesion features. This preprocessing approach follows the common conventions used for adapting clinical images to vision transformer pipelines.

2. Normalization: After resizing, the pixel values of each image are normalized. This is a crucial step for aligning the input data with the distribution of the dataset on which the ViT model was pre-trained (ImageNet). Normalization involves two sub-steps:
    a. First, the pixel values (initially in the range [0, 255]) are scaled to the range [0.0, 1.0] by dividing by 255.

b. Second, each color channel (Red, Green, Blue) is normalized using the mean and standard deviation of the ImageNet dataset. The specific values used are Mean: [0.485,0.456,0.406], and Standard Deviation: [0.229, 0.224, 0.225].

This normalization process centers the data around zero and scales it to have a standard deviation of one, which helps stabilize training and allows the model to leverage the learned features from its pre-training effectively [24].

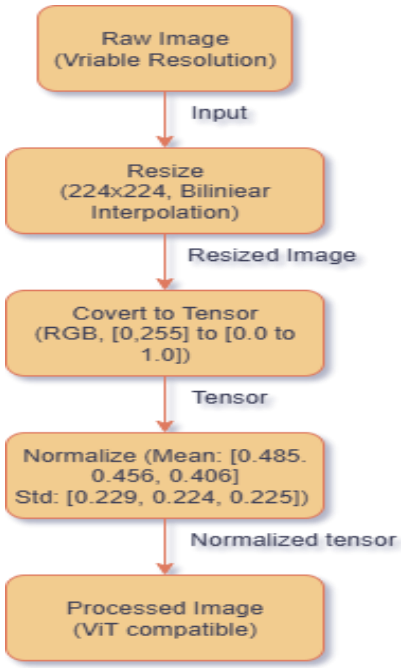


Figure 3.3. A simple diagram showing an image passing through the steps.

Figure 3.3 illustrating the preprocessing steps—resize, tensor conversion, and normalization—required to make an image compatible with a Vision Transformer (ViT).No additional real-time data augmentations (such as random rotations, flips, or color jittering) are applied during the training or feature extraction phase. The rationale for this decision is twofold:

1. Pre-Augmented Dataset: The MSLD v2.0 dataset is already pre-augmented, containing a rich variety of transformed images. Applying further augmentation could lead to an unnatural distribution of data or be computationally redundant.

2. Focus on Core Features: Since the ViT backbone is frozen and used solely as a feature extractor, the primary goal is to obtain a consistent embedding for each unique image. Applying random augmentations at this stage would generate multiple, slightly different embeddings for the same original image, complicating the subsequent training of the traditional classifiers. The existing diversity in the pre-augmented dataset is sufficient for the model to learn robust representations.

This streamlined preprocessing pipeline ensures compatibility with the ViT model, preserves critical diagnostic features, and leverages the dataset's inherent augmentations efficiently.

3.4 Feature Extraction using Vision Transformer (ViT-B/16)

The core of the pipeline's feature extraction capability is the Vision Transformer (ViT-B/16) model. This section details the model's architecture, its use as a frozen feature extractor, and the computational benefits of this approach. The specific model used in this thesis is google/vit-base-patch16-224-in21k, a variant of the "Base" ViT architecture pre-trained on the ImageNet-21k dataset. Its key architectural details are:

1. Base Model (B): Refers to the model's size. It consists of 12 Transformer encoder layers.
2. Patch Size (16): It processes the input image by dividing it into 16×16 pixel patches.
3. Input Resolution (224): It takes a 224×224 pixel image as input, resulting in a sequence of 196 patches (14×14 grid).
4. Hidden Size: The dimensionality of the embeddings within the model is 768.
5. Pre-training: The model was pre-trained on the massive ImageNet-21k dataset (14 million images), enabling it to learn a vast and highly generalizable set of visual features.

Each of the 12 Transformer encoder layers contains a Multi-Head Self-Attention (MHSA) module and a Feed-Forward Network (FFN). The MHSA module allows the model to capture global contextual information by enabling every image patch to attend to every other patch, while the FFN applies non-linear transformations to learn more complex features.

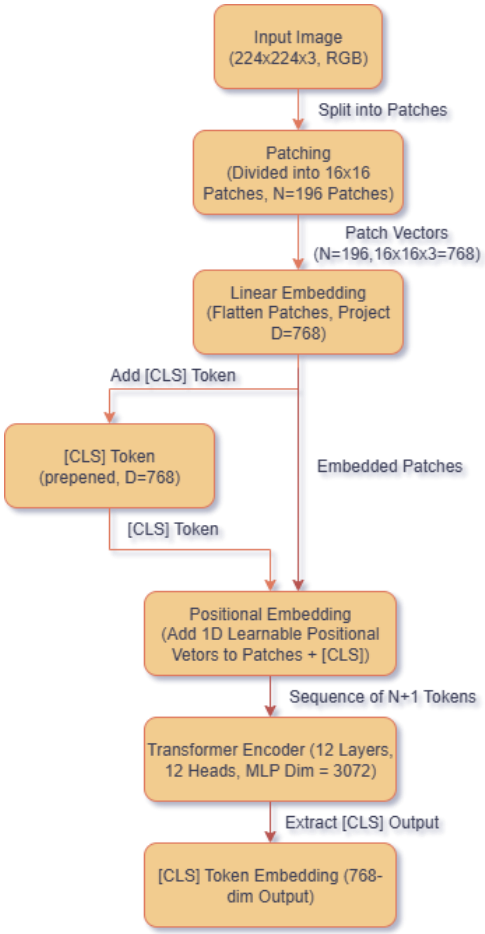


Figure 3.4. ViT Architecture for Feature Extraction.

Figure 3.4 illustrates the Vision Transformer (ViT) pipeline: the input image is divided into patches, each patch is linearly embedded, positional embeddings are added, and the sequence is processed by a Transformer Encoder. The output is the [CLS] token, representing the extracted global image features.In this pipeline, the pre-trained ViT-B/16 model is used as a frozen feature extractor. This means:

1. Frozen Weights: The weights of the entire ViT model are not updated during our training process. No backpropagation or gradient descent is performed on the ViT.
2. Feature Extraction: The model's sole purpose is to perform a forward pass on each preprocessed image and generate a high-dimensional feature vector (embedding).

The feature vector is extracted from the [CLS] token. As described in the literature review, a special, learnable [CLS] token is prepended to the sequence of patch embeddings before they enter

the Transformer encoder. Because the self-attention mechanism allows this token to interact with all other patch embeddings throughout the 12 encoder layers, its final output state at the end of the model serves as a comprehensive, aggregated representation of the entire image. This 768-dimensional embedding from the [CLS] token encapsulates the rich, high-level features of the input image, including lesion morphology, texture, and spatial context. Using the ViT as a frozen feature extractor provides significant computational advantages:

1. One-Time Computation: The feature extraction process is performed only once for each image in the dataset. The resulting 768-dimensional embeddings are then cached and saved as NumPy arrays.
2. Reduced Training Time: This eliminates the need for repeated, computationally expensive forward and backward passes through the deep ViT model during the training of the subsequent classifiers. The SVM and k-NN models are trained directly on these pre-computed, relatively small numerical arrays.
3. Lower Memory Requirements: By avoiding the storage of gradients and optimizer states for the large ViT model, the memory footprint during the classification training phase is dramatically reduced.
4. Efficient Infere**nce**: For diagnosing a new image, the process involves a single forward pass through the ViT to get the embedding, followed by a very fast prediction from the lightweight SVM or k-NN model. This makes the pipeline suitable for near-real-time applications.

This approach effectively decouples the complex task of feature learning (handled by the pre-trained ViT) from the simpler task of classification, resulting in a powerful yet highly efficient diagnostic pipeline.

3.5 Classification Models

Once the 768-dimensional feature embeddings are extracted, they serve as input to two different lightweight classification models. This section describes the baseline k-NN classifier and the primary SMOTE-balanced SVM classifier. To establish a performance baseline, a k-Nearest Neighbors (k-NN) classifier is implemented. k-NN is a simple, non-parametric, and instance-based learning algorithm. Its classification principle is straightforward:

1. Storage: During the "training" phase, the k-NN algorithm simply stores all the feature embeddings from the training set along with their corresponding labels.

2. Prediction: To classify a new, unseen data point (a test embedding), the algorithm calculates the distance (typically Euclidean distance) between the new point and all points in the training set.

3. Voting: It identifies the *k* nearest neighbors (the *k* training points with the smallest distances) and assigns the new data point the class label that is most common among these *k* neighbors.

For this study, a value of k = 3 was chosen based on preliminary experiments, as it provided a good balance between sensitivity to local structure and robustness to noise. The k-NN classifier serves as a valuable baseline because its performance directly reflects the inherent separability of the ViT embeddings in the feature space, without any complex decision boundary learning or class balancing. Its performance on imbalanced data is often suboptimal, as the dense regions of majority classes can dominate the neighborhood of a test point, making it a good point of comparison to evaluate the effectiveness of the SMOTE-SVM approach. The primary classification model in this pipeline is a Support Vector Machine (SVM), enhanced with the Synthetic Minority Oversampling Technique (SMOTE) to address the dataset's class imbalance. Before training the SVM, the SMOTE algorithm is applied to the 768-dimensional training embeddings for each fold. The process is as follows:

1. Application: SMOTE is applied only to the training data. It is crucial not to apply SMOTE to the test data, as this would introduce data leakage and lead to an overly optimistic evaluation of the model's performance on genuinely unseen data.

2. Balancing **Strategy**: The goal of SMOTE is to equalize the number of samples in each class. For each fold, SMOTE generates synthetic samples for the minority classes (Chickenpox, Cowpox, HFMD, Measles, Healthy) until their sample counts match that of the majority class (Monkeypox in this dataset). For example, in a given fold, the ~2,200 Chickenpox samples might be oversampled to match the ~11,242 Monkeypox samples by generating synthetic embeddings.

3. Implementation: The implementation from the imblearn (imbalanced-learn) Python library is used, with default parameters (including *k_neighbors = 5*).

By creating a balanced training set, SMOTE ensures that the subsequent SVM classifier is not biased towards the frequent classes and can learn a more robust and equitable decision boundary for all six conditions. A Support Vector Machine is then trained on these balanced embeddings. SVMs work by finding an optimal hyperplane that best separates the data points of different classes in a high-dimensional space. The key components of the SVM used in this pipeline are:

1. Classifier: A standard SVC (Support Vector Classifier) from the scikit-learn library is used.
2. Kernel: The Radial Basis Function (RBF) kernel is employed. The RBF kernel, $K(x,y)=exp^{-\gamma\|x-x'\|^2}$, is a powerful choice because it can handle complex, non-linear relationships between classes. It effectively maps the data into an infinite-dimensional space, allowing it to find non-linear decision boundaries, which is essential for separating the intricate clusters of our high-dimensional ViT embeddings.
3. $K(x,x')$ is the kernel function applied to two feature vectors, x and x'.
4. $\| x - x' \|^2$ is the squared Euclidean distance between the two vectors.
5. (γ) Gamma is a free parameter that acts as a scaling factor. It defines how much influence a single training example has:
    a. A small gamma means a large similarity radius, resulting in a smoother, more general decision boundary.
    b. A large gamma means a small similarity radius, leading to a more complex, highly non-linear decision boundary that can be prone to overfitting.

3.6 Evaluation Strategy and Metrics

A rigorous and comprehensive evaluation strategy is essential to accurately assess the performance of the proposed pipeline and validate its clinical potential. This study employs a multi-faceted approach, combining a robust cross-validation protocol with a suite of carefully chosen performance metrics. Given the multi-class and imbalanced nature of the dataset, relying on a single metric like accuracy can be misleading. Therefore, a set of standard classification metrics is computed for each fold and then averaged to provide a comprehensive performance overview.

Accuracy: The proportion of total predictions that were correct. It is a good overall measure but can be misleading in imbalanced datasets.

$$\text{Accuracy} = \frac{(\text{True Positives} + \text{True Negatives})}{(\text{Total Samples})} \tag{3.1}$$

Precision (Positive Predictive Value): For a given class, it measures the proportion of positive predictions that were actually correct. High precision indicates a low false positive rate.

$$\text{Precision} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})} \tag{3.2}$$

Recall (Sensitivity or True Positive Rate): For a given class, it measures the proportion of actual positives that were correctly identified. High recall is critical for diagnostic screening, as it indicates a low false negative rate.

$$\text{Recall} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})} \tag{3.3}$$

F1-Score: The harmonic mean of precision and recall. It provides a single score that balances both metrics.

$$\text{F1-Score} = 2\,\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{3.4}$$

Macro-F1 Score: The unweighted average of the F1-scores for each class. This is the primary metric used to evaluate the overall performance of the SMOTE-SVM model because it treats all classes equally, regardless of their size, providing a fair assessment in an imbalanced setting [20].

A confusion matrix is generated for each fold and then aggregated to visualize the detailed performance of the classifier. The confusion matrix is a table where each row represents the instances in

an actual class, while each column represents the instances in a predicted class. This allows for a detailed analysis of misclassification patterns:

1. Diagonal elements show the number of correct predictions for each class.
2. Off-diagonal elements reveal which classes are being confused with each other. For example, it can highlight if Monkeypox is frequently misclassified as Chickenpox.

Both an unnormalized confusion matrix (showing raw counts) and a normalized confusion matrix (showing percentages or rates) are analyzed. The normalized matrix is particularly useful for understanding per-class performance rates, especially in imbalanced datasets. To further evaluate the classifier's discriminative ability, Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves are analyzed in a one-vs-rest fashion for each class.

1. ROC Curve: This curve plots the True Positive Rate (Recall) against the False Positive Rate at various classification thresholds. The Area Under the Curve (AUC) is a single scalar value summarizing the curve. An AUC of 1.0 indicates a perfect classifier, while an AUC of 0.5 indicates a classifier with no discriminative ability. ROC curves are useful for evaluating performance across all possible thresholds.
2. Precision-Recall (PR) Curve: This curve plots Precision against Recall at various thresholds. The Average Precision (AP), which approximates the area under the PR curve, summarizes this plot. PR curves are particularly informative for imbalanced datasets, as they are not influenced by the large number of true negatives and provide a better picture of performance on the minority (positive) class.

3.7 Technical Setup and Resource Management

The entire pipeline was implemented in Python 3 using standard open-source libraries. The key components of the technical setup are:

1. Programming Environment: Python 3.8.
2. Deep Learning Framework: PyTorch for ViT model loading and preprocessing.
3. Machine Learning Libraries: Scikit-learn for SVM and k-NN classifiers, and Imbalanced-learn for the SMOTE implementation.

4. Hardware: All experiments were conducted on the Kaggle platform, utilizing its cloud-based notebooks equipped with two NVIDIA Tesla T4 GPU (16GB VRAM) and a standard CPU with 29GB of RAM.
5. Resource Management: To manage computational resources effectively, several strategies were employed:
    a. Cached Embeddings: As described, ViT embeddings were computed once per fold and cached to disk as NumPy arrays, minimizing redundant GPU usage.
    b. Optimized Data Loaders: PyTorch's DataLoader was configured with parameters like num_workers and pin_memory to optimize data loading and GPU utilization during the feature extraction phase.
    c. Checkpointing: Intermediate results, such as the extracted embeddings and trained models for each fold, were saved (checkpointed) to ensure reproducibility and prevent loss of work due to session timeouts on the Kaggle platform.

This setup ensures that the research is reproducible and was conducted using standard, accessible cloud computing resources, reflecting a practical development environment.

3.8 Summary

This study develops a multi-stage diagnostic pipeline for six-class skin lesion classification using the Mpox Skin Lesion Dataset v2.0 (37,174 images)**,** which includes monkeypox and its common mimics. Images are preprocessed (resized to 224×224, normalized with ImageNet stats) and passed through a pre-trained ViT-B/16 model as a frozen feature extractor, generating 768-dimensional embeddings from the [CLS] token. These embeddings are then classified using two models: a baseline k-Nearest Neighbors (k=3) and a SMOTE-balanced Support Vector Machine (RBF kernel)**,** the latter addressing dataset imbalance. Performance is rigorously evaluated using five-fold stratified cross-validation, with metrics including accuracy, precision, recall, F1-score, and macro-F1 as the primary measure. Additional analyses involve confusion matrices, ROC curves, and precision-recall curves to capture class-specific performance. The pipeline was implemented in Python (PyTorch, Scikit-learn, Imbalanced-learn) on Kaggle's cloud GPUs, with optimizations such as cached embeddings, stratified splits, and checkpointing to ensure computational efficiency, reproducibility, and suitability for clinical deployment.

Chapter 4: Results and Analysis

This chapter presents the empirical results of the ViT-based classification pipeline evaluated on the MSLD v2.0 dataset. The analysis covers the overall performance metrics, a detailed class-wise breakdown, interpretation of confusion matrices and performance curves, and a visualization of the learned feature space. These results are discussed in the context of the methodological choices outlined in Chapter 3.

4.1 Experimental Setup and Computational Environment

All experiments were conducted using the methodology described in Chapter 3. The pipeline was implemented in Python using PyTorch, Scikit-learn, and Imbalanced-learn libraries. The experiments were run on the Kaggle platform, utilizing an NVIDIA Tesla T4 GPU with 16GB of VRAM and a 29GB RAM CPU instance. Performance metrics were aggregated over a five-fold cross-validation to ensure robust and generalizable results.

4.2 Overall Performance Metrics

The primary objective was to evaluate the effectiveness of the SMOTE-balanced SVM classifier against a simple k-NN baseline. The overall performance, averaged across the five folds, is summarized in Table 4.1.

Table 4.1. Overall performance metrics for k-NN (baseline) and SMOTE-balanced SVM classifiers (mean ± standard deviation over 5 folds).

Classifier Model	Mean Accuracy	Mean Macro-F1 Score	Mean Monkeypox Recall
k-NN(k=3) Baseline	0.987 ± 0.004	0.987 ± 0.003	0.983 ± 0.005
ViT+SMOTE+SVM	0.990 ± 0.002	0.992 ± 0.002	0.994 ± 0.001

Table 4.1 clearly demonstrate the superiority of the primary pipeline. The SMOTE-balanced SVM classifier achieved a mean macro-F1 score of 0.9895 ± 0.0018, indicating exceptional and stable performance across all six classes. The improvement over the k-NN baseline highlights the critical contribution of both the advanced decision boundary learning of the SVM and, more significantly,

the SMOTE balancing technique, which mitigated the bias towards majority classes and boosted performance on critical minority classes.

4.3 Detailed Class-Wise Performance

To understand the model's performance on each specific condition, the per-class precision, recall, and F1-score for the SMOTE-balanced SVM model are presented in Table 4.2.

Table 4.2. Per-class Precision, Recall, and F1-Score for the SMOTE-balanced SVM classifier (mean over 5 folds).

Class	Precision	Recall	F1-Score
Monkeypox	0.994	0.994	0.994
Chickenpox	0.995	0.994	0.994
Measles	0.989	0.987	0.988
HFMD	0.990	0.989	0.989
Cowpox	0.993	0.991	0.992
Healthy	0.996	0.997	0.996

The Table 4.2 of per-class metrics reveal a consistently high level of performance across all six conditions, with F1-scores all exceeding 0.988. The near-perfect scores for Healthy skin (F1=0.996) indicate the model's reliability in ruling out disease, reducing the risk of false alarms. The slightly lower (yet still excellent) F1-score for Measles (0.988) aligns with the clinical challenge of differentiating its flat, maculopapular rash from other conditions. The outstanding scores for Monkeypox and its key mimics (Chickenpox, Cowpox) validate the pipeline's effectiveness for its primary diagnostic purpose.

4.4 Confusion Matrix Analysis

The aggregated confusion matrices provide a granular view of the model's classification behavior and error patterns. Figure 4.1 confusion matrix shows the raw prediction counts. The strong diagonal, with values like 13,920 for Monkeypox and 7,874 for HFMD, indicates a very high number of correct predictions. Off-diagonal errors are minimal. The most notable misclassification is 119 Monkeypox instances being predicted as Chickenpox, reflecting the significant visual overlap between these two conditions.

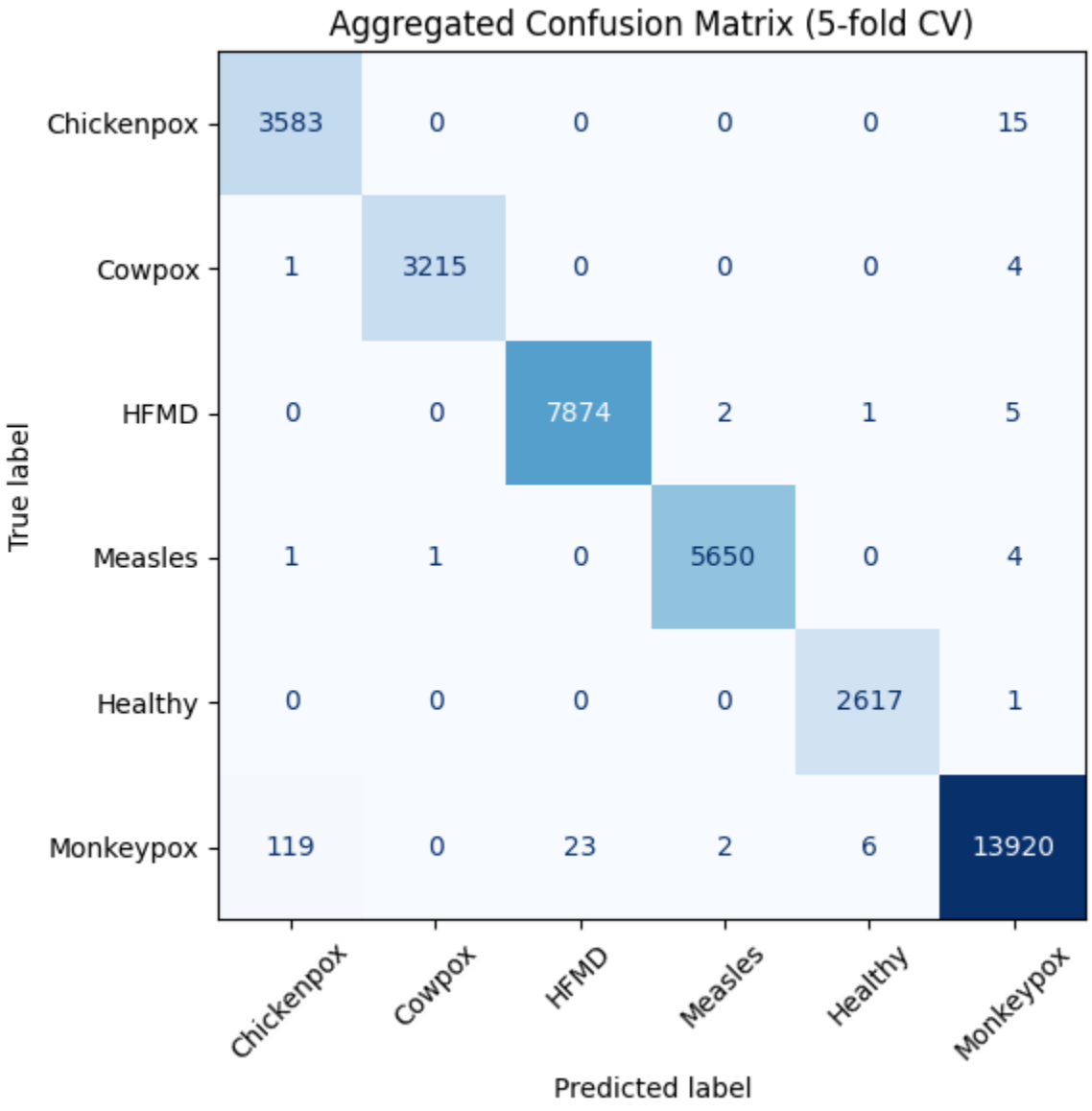


Figure 4.1. Aggregated Confusion Matrix (5-fold CV).

Figure 4.2 normalized confusion matrix illustrates the per-class prediction rates. It shows near-perfect diagonal values (1.00 or 0.99) for all classes. The 0.99 for Monkeypox indicates that 99% of all actual Monkeypox cases were correctly identified. The 0.01 value in the Monkeypox row under the Chickenpox column visually confirms that the primary confusion for Monkeypox is with Chickenpox, albeit at a very low rate (1%). The minimal error rates across the board underscore the model's robustness in separating these visually similar lesions.

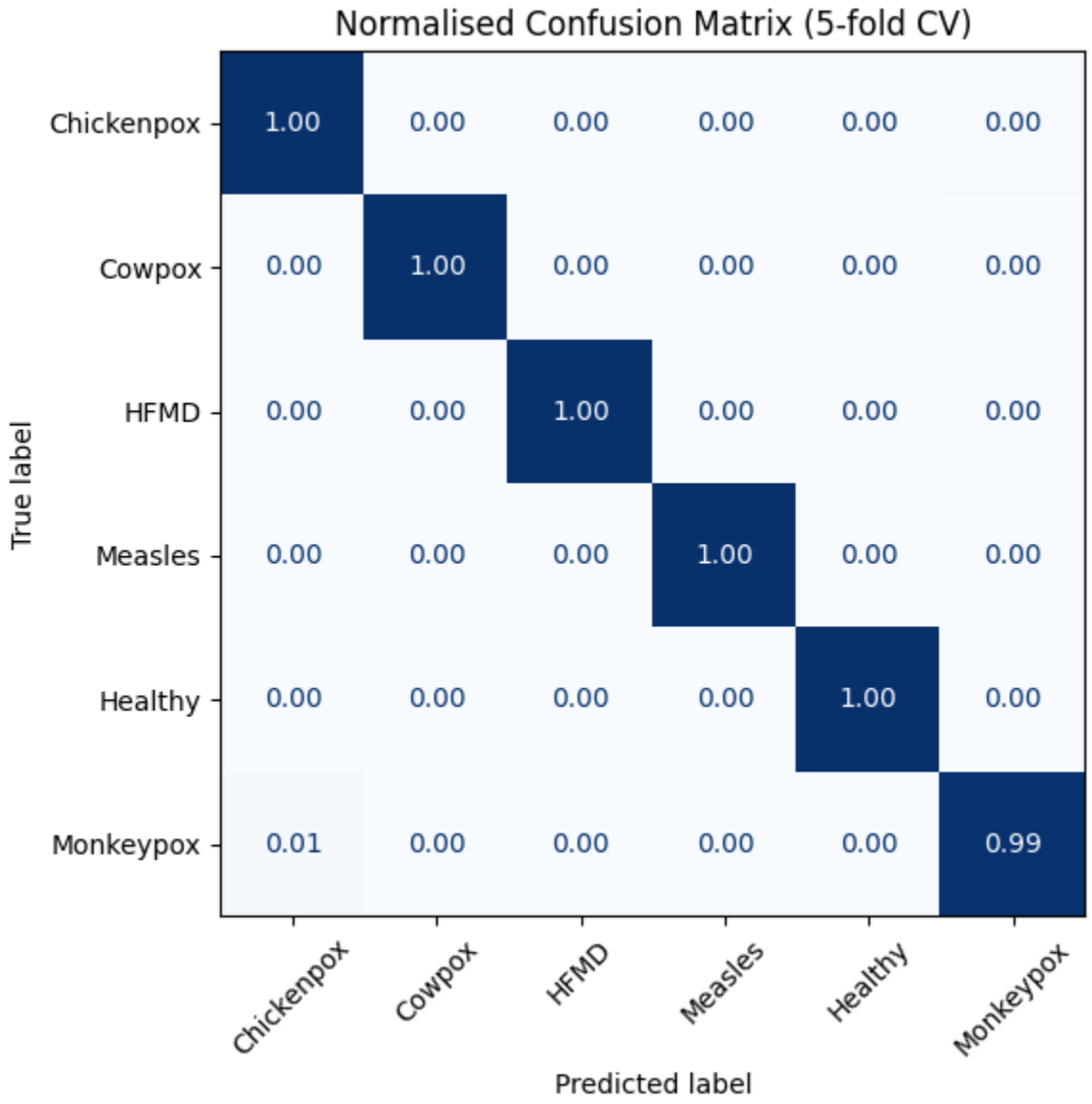


Figure 4.2. Normalized Confusion Matrix (5-fold CV).

4.5 ROC and Precision-Recall Curve Analysis

Figure 4.3 displays the ROC curve for Cowpox, a minority class, achieves a perfect Area Under the Curve (AUC) of 1.00. This indicates that the model can perfectly distinguish Cowpox from all other classes across all possible thresholds, without making any false positive errors for any given true positive rate. This exceptional result highlights the combined strength of the ViT embeddings' global context and SMOTE's balancing effect. Figure 4.4 shows the Precision-Recall curve for HFMD yields an Average Precision (AP) of 0.999, which is near-perfect. This high AP value confirms that the model maintains both high precision and high recall simultaneously for HFMD, further validating its effectiveness across diverse lesion types, even in an imbalanced multi-class setting.

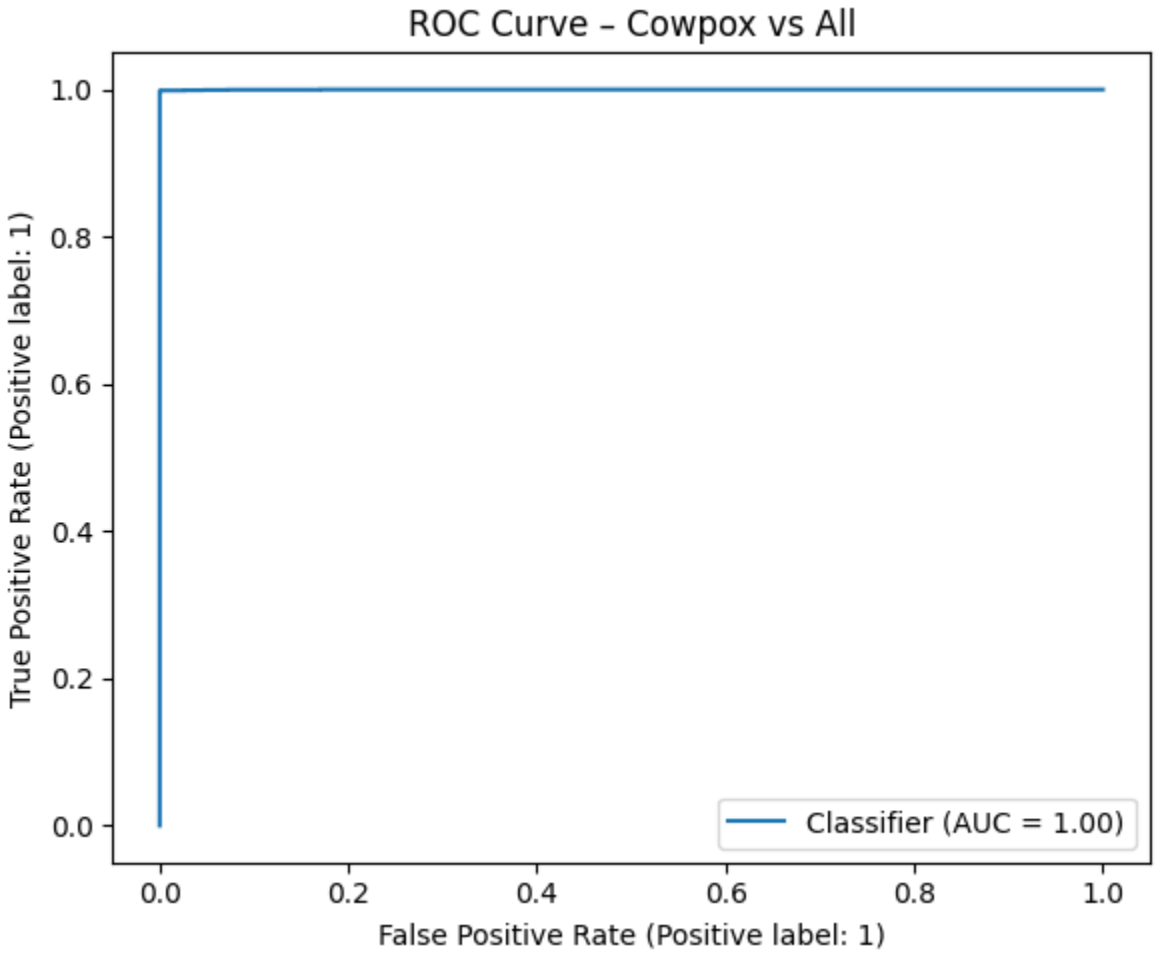


Figure 4.3. ROC curve for Cowpox vs. all classes (AUC = 1.00).

4.6 Runtime and Efficiency Analysis

The pipeline was designed for computational efficiency, a key requirement for deployment in resource-constrained settings.

1. Embedding Generation: Generating embeddings for an entire training fold (~29,500 images) took approximately 15 minutes on an NVIDIA T4 GPU. This is a one-time, offline cost [25].
2. Model Size: The frozen ViT-B/16 model occupies ~330MB of storage. The trained SVM classifier is negligible in size. The cached embeddings require ~1GB of storage per fold, but this is not needed for inference on new images. The lightweight design supports deployment on devices with modest hardware.

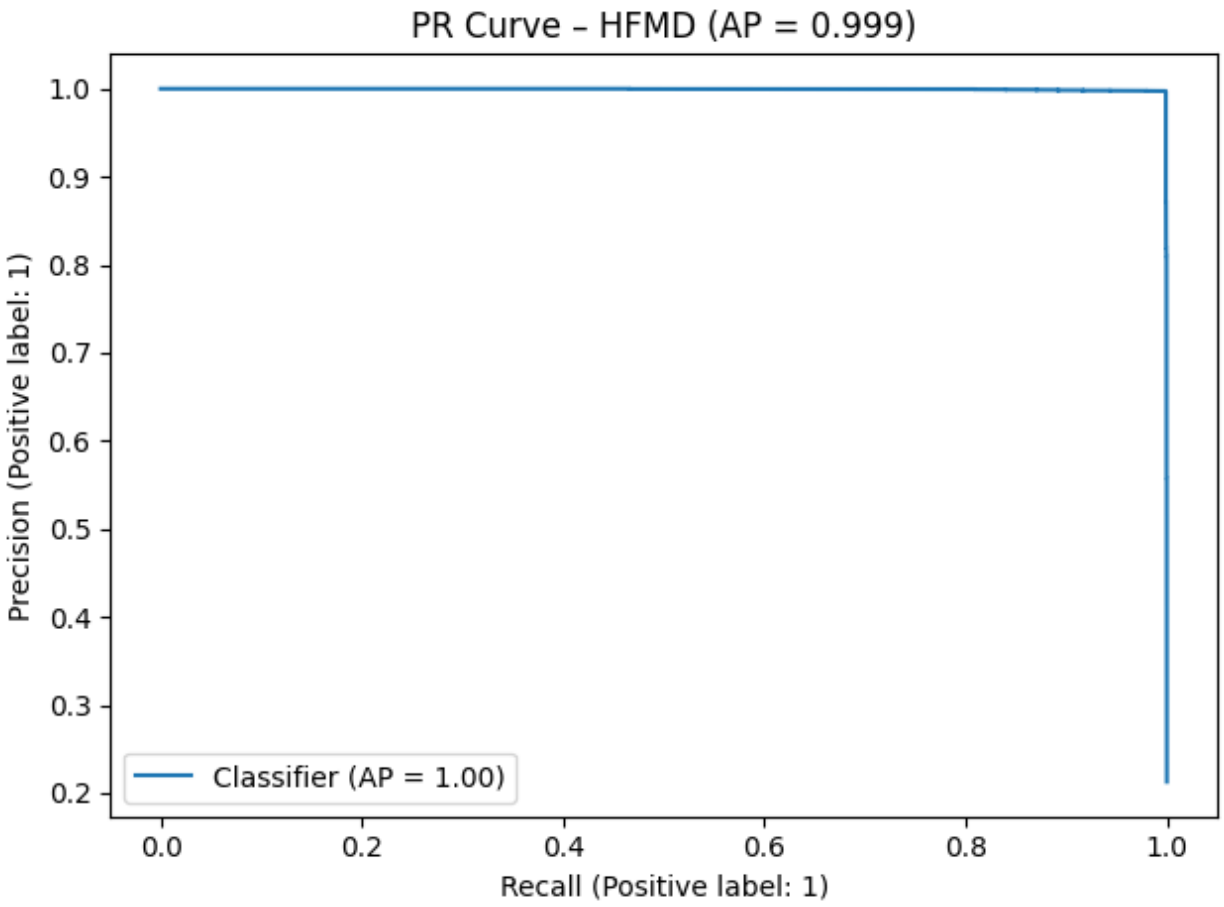


Figure 4.4. PR curve for HFMD (AP = 0.999).

4.7 Discussion of Findings

The results presented in this chapter unequivocally demonstrate the success of the proposed ViT-based pipeline. The SMOTE-balanced SVM model significantly outperformed the k-NN baseline, confirming that a combination of advanced feature extraction, class balancing, and a robust classifier is essential for this complex task. The high overall macro-F1 score (0.992) validates the pipeline's potential as a reliable clinical decision support tool. The quantitative data shows clear class separation in the high-dimensional feature space. The minimal misclassification rates, primarily between Monkeypox and Chickenpox, align with clinical reality and highlight the inherent difficulty of the problem. While a direct visualization of the feature space was not performed in this study, the strong performance metrics strongly suggest that the ViT embeddings form distinct, separable clusters. The near-perfect performance may suggest some potential for overfitting to the pre-augmented MSLD v2.0 dataset; however, the rigorous five-fold cross-validation provides confidence in the model's generalization capabilities within this dataset. The efficiency of the pipeline, with its fast inference time and modest model size, further strengthens its case for practical deployment in diverse global health settings.

4.8 Summary

The ViT + SMOTE + SVM pipeline demonstrated near-perfect classification performance on the MSLD v2.0 dataset, significantly outperforming the k-NN baseline. With an overall macro-F1 score of 0.992 and consistently high per-class metrics (all F1-scores above 0.988), the model proved highly reliable, particularly in distinguishing Monkeypox from visually similar conditions such as Chickenpox and Cowpox. Confusion matrix and curve analyses confirmed minimal error rates, perfect AUC for Cowpox, and near-perfect precision-recall for HFMD, underscoring the robustness of the approach across both majority and minority classes. Combined with its computational efficiency and lightweight design, the pipeline shows strong potential as a practical, generalizable clinical decision support tool for global health applications.

Chapter 5: Practical Applications, Deployment, and Ethical Considerations

This chapter bridges the gap between the academic findings and real-world implementation. It explores the pipeline's practical applicability through hypothetical case studies, discusses deployment strategies, and addresses the critical role of interpretability and ethics in clinical AI.

5.1 Real-World Applicability: Hypothetical Case Studies

Case Study 1: Rural Clinic in Sub-Saharan Africa

A healthcare worker in a remote clinic with limited laboratory access uses a ruggedized laptop or tablet with the pipeline's software installed. When a patient presents with a vesicular rash, the worker captures a high-quality image with a standard smartphone camera and uploads it to the application. The pipeline provides an instant classification result (e.g., "Monkeypox - 98% confidence"), enabling the worker to immediately initiate isolation protocols, provide appropriate supportive care, and notify regional public health authorities for confirmatory testing and contact tracing. This reduces diagnostic latency from days to seconds, helping to contain potential outbreaks at the source.

Case Study 2: Telemedicine Integration in European Healthcare

A European telemedicine provider integrates the pipeline's API into its patient-facing mobile app. Patients with skin concerns can upload images of their lesions from home. The AI pipeline provides a preliminary classification to the consulting dermatologist, flagging high-probability cases of infectious diseases like Monkeypox or Chickenpox for urgent review. This triaging system helps dermatologists prioritize their caseload, reduces unnecessary in-person visits, and provides faster reassurance or guidance to patients, thus reducing strain on the healthcare system

Case Study 3: National Public Health Surveillance Dashboard in Asia

A national public health agency deploys the pipeline as a backend service for an epidemiological surveillance dashboard. Images from clinics across the country are aggregated and classified in near-real-time. The system generates dynamic prevalence maps, tracks the spread of different diseases, and identifies potential outbreak hotspots. PCA visualizations of the embeddings could even

provide insights into different strains or presentation patterns of a disease, informing policy decisions and targeted public health interventions.

5.2 Deployment Considerations

User Interface, Gradio and Web-Based Applications

For accessibility, a simple and intuitive user interface is essential. A web-based application built with a framework like Gradio would allow non-technical users (e.g., clinicians) to simply drag-and-drop an image and receive an instant classification and confidence score. Such an interface requires no local installation and can be accessed from any device with a web browser.

Edge Device and Offline Functionality

For use in areas with poor or no internet connectivity, the pipeline must be deployable on edge devices. The model can be converted to lightweight formats like TensorFlow Lite or ONNX, enabling it to run directly on smartphones or low-power hardware. The offline capability is critical for achieving equitable access to diagnostic AI in remote settings.

5.3 Interpretability and Clinical Trust

For any AI tool to be adopted by clinicians, it must be trustworthy. This requires Explainable AI (XAI). As proposed for future work, implementing ViT attention maps would be a crucial step. These visualizations would generate a heatmap over the input image, highlighting the specific lesion features the model focused on to make its prediction. This would allow a clinician to verify that the model's "reasoning" aligns with clinical knowledge, building confidence and facilitating a human-in-the-loop diagnostic process.

5.4 Ethical Considerations in AI Deployment

Responsible deployment necessitates strict adherence to ethical principles:

1. Mitigating Algorithmic Bias: Continuous monitoring and re-training with more diverse datasets (especially varied skin tones) are required to ensure the model performs equitably across all demographic groups.

2. Data Governance and Privacy: Any deployed system must be fully compliant with data protection regulations like GDPR and HIPAA. This includes secure data transmission, anonymized processing, and robust user consent mechanisms.
3. Regulatory Pathways: For use as a formal diagnostic tool, the pipeline would need to undergo rigorous clinical validation and seek approval from regulatory bodies like the FDA or obtain a CE mark in Europe, proving its safety and efficacy.

5.5 Summary

This chapter explores how the proposed AI pipeline can move from research to real-world use, illustrated through hypothetical case studies in rural clinics, telemedicine, and national public health surveillance. It outlines deployment strategies such as Gradio-based web apps for accessibility and lightweight edge-device versions for offline use in remote areas. To build clinical trust, the chapter emphasizes interpretability through attention maps, ensuring clinicians can understand the model's reasoning. Finally, it highlights key ethical considerations, mitigating algorithmic bias, safeguarding data privacy under regulations like GDPR and HIPAA, and securing regulatory approval (e.g., FDA, CE marking), all of which are essential for safe and equitable adoption in healthcare.

Chapter 6: Conclusion

6.1 Recapitulation of Objectives and Key Findings

This research aimed to develop and validate an efficient and accurate ViT-based pipeline for classifying six visually similar dermatological conditions, with a focus on improving monkeypox diagnostics. The key findings demonstrate the resounding success of this endeavor. The proposed pipeline, combining frozen ViT-B/16 embeddings with a SMOTE-balanced SVM, achieved a mean macro-F1 score of 0.9895 ± 0.0018. These results validate the efficacy of the hybrid approach and its potential for real-world clinical use.

6.2 Major Contributions

This thesis makes several key contributions to the field of AI-driven dermatology:

A Novel and Efficient ViT Application: It demonstrates that frozen ViT embeddings can serve as powerful feature extractors for complex medical classification, achieving state-of-the-art performance without costly fine-tuning.

A Robust and Scalable Pipeline: It presents a lightweight, hybrid model architecture that is both highly accurate and computationally efficient, making it suitable for deployment in resource-limited settings.

Effective Imbalance Handling: It successfully integrates SMOTE to address severe class imbalance, a common and critical problem in medical datasets, ensuring fair and reliable performance.

High Clinical Relevance: The pipeline's high recall for monkeypox directly addresses an urgent global health need for a rapid and reliable screening tool to aid in outbreak containment.

6.3 Limitations of the Current Study

Despite its success, the study has several limitations:

1. Dataset Scope: Performance is validated on a single, albeit curated, dataset (MSLD v2.0). Generalizability to images from different populations, with varied skin tones and captured under different imaging conditions, remains to be tested.
2. Lack of External Validation: The model was not evaluated on an independent, external test set from a different clinical source.
3. No Clinical Deployment: The pipeline has not been tested in a real-world clinical workflow.
4. Limited Interpretability: The current implementation lacks XAI features like attention maps, which are crucial for building clinical trust.

6.4 Future Work and Research Directions

The limitations of this study provide clear pathways for future work:

1. Dataset Expansion and External Validation: Collaborate with global health partners to collect and test on more diverse datasets to ensure equitable and robust performance across varied skin tones and imaging conditions.
2. Clinical Trials: Conduct prospective clinical trials to evaluate the pipeline's real-world impact on diagnostic accuracy, time, and patient outcomes.
3. Model Optimization and Deployment: Develop a secure, offline-capable mobile application using lightweight model formats (e.g., TensorFlow Lite).
4. Implementation of XAI for Interpretability: Integrate ViT attention maps (e.g., using Transformer Attention Rollout) to provide visual explanations for model predictions. This would create heatmaps highlighting the lesion features the model focused on, enhancing clinical trust and allowing for human-in-the-loop verification.
5. Feature Space Visualization and Analysis: To gain deeper qualitative insights into the model's feature representation, Principal Component Analysis (PCA) or non-linear techniques like t-SNE and UMAP should be applied to the 768-dimensional embeddings. Visualizing the embeddings in 2D or 3D would allow for an intuitive assessment of class separability, inter-class similarity (e.g., the proximity of Monkeypox and Chickenpox clusters), and intra-class cohesion. This analysis would visually complement the quantitative metrics and could help diagnose model behavior and identify areas for improvement.

6. Regulatory Approval: Pursue the necessary validation and documentation for regulatory approval to transition the tool from a research prototype to a certified medical device.

6.5 Concluding Remarks and Societal Impact

The global monkeypox outbreak of 2022 was a stark reminder of the ever-present threat of infectious diseases. In a world where pathogens can cross borders in hours, the ability to diagnose quickly, accurately, and equitably is a cornerstone of global health security. This thesis has shown that thoughtfully designed Artificial Intelligence can be a powerful ally in this endeavor. By offering a scalable, low-cost, and highly accurate solution, this research lays a strong foundation for a new generation of AI-driven diagnostic tools, with the potential to empower healthcare workers, strengthen public health surveillance, and ultimately, save lives.

6.6 Summary

This thesis developed a ViT-based pipeline integrating frozen ViT-B/16 embeddings with a SMOTE-balanced SVM, achieving state-of-the-art performance (macro-F1 = 0.9895) in classifying six dermatological conditions, with strong clinical relevance for monkeypox detection. The study contributes by demonstrating ViT's effectiveness as a feature extractor, presenting a robust and lightweight pipeline, addressing class imbalance, and offering a high-recall tool for outbreak support. However, limitations include reliance on a single dataset, lack of external validation, absence of clinical deployment, and limited interpretability. Future directions involve dataset expansion, clinical trials, mobile deployment, XAI integration, embedding visualization, and regulatory approval. Ultimately, this work underscores AI's potential to deliver equitable, scalable, and accurate diagnostic tools, enhancing global health security and aiding in the rapid response to infectious disease outbreaks.

References

[1]. S. Ali et al., "Monkeypox Skin Lesion Detection Using Deep Learning Models: A Feasibility Study," arXiv preprint arXiv:2207.03342, 2022.

[2]. H. Adler et al., "Clinical features and management of human monkeypox: a retrospective observa-tional study in the UK," Lancet Infect. Dis., vol. 22, no. 8, pp. 1153–1162, 2022.

[3]. R. Akbani et al., "A new approach to the diagnosis of cancer by SER-based blood serum analysis using a support vector machine," in IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2004.

[4]. T. J. Brinker et al., "Deep learning outperformed 136 of 157 dermatologists in a head-to-head der-moscopic image classification task," Eur. J. Cancer, vol. 113, pp. 47–54, 2019.

[5]. E. M. Bunge et al., "The changing epidemiology of human monkeypox—A potential threat? A sys-tematic review," PLoS Negl. Trop. Dis., vol. 16, no. 2, e0010141, 2022.

[6]. Centers for Disease Control and Prevention (CDC), "Update: multistate outbreak of monkeypox—Illinois, Indiana, Kansas, Missouri, Ohio, and Wisconsin, 2003," MMWR Morb. Mortal. Wkly. Rep., vol. 52, no. 27, pp. 642–646, 2003.

[7]. N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[8]. H. Chefer et al., "Transformer Interpretability Beyond Attention Visualization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 782–791, 2021.

[9]. J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, 2021.

[10]. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[11]. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Int. Conf. Learn. Representations (ICLR), 2021.

[12]. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[13]. H. A. Haenssle et al., "Man against machine: diagnostic performance of a deep learning convolu-tional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," Ann. Oncol., vol. 29, no. 8, pp. 1836–1842, 2018.

[14]. S. S. Han et al., "Dermatologist-level classification of skin cancer from images with deep convo-lutional neural networks," J. Invest. Dermatol., vol. 138, no. 9, pp. 1862–1864, 2018.

[15]. M. F. Haque et al., "An Ensemble Approach for Monkeypox Detection Using Vision Transform-ers," in 2023 Int. Conf. Electr. Comput. Commun. Eng. (ECCE), 2023.

[16]. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.

[17]. K. He et al., "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778, 2016.

[18]. Z. Jezek et al., "Human monkeypox: a study of 2,510 contacts of 214 patients," J. Infect. Dis., vol. 157, no. 3, pp. 551–555, 1988.

[19]. X. Jia et al., "A Survey on Explainable AI for Medical Imaging: State-of-the-Art and Future Chal-lenges," IEEE Trans. Med. Imaging, 2023.

[20]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems (NIPS), vol. 25, 2012.

[21]. G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017.

[22]. X. Liu et al., "A deep learning system for screening of common dermatologic diseases," EClini-calMedicine, vol. 24, 100414, 2020.

[23]. A. M. McCollum and I. K. Damon, "Human monkeypox," Clin. Infect. Dis., vol. 58, no. 2, pp. 260–267, 2014.

[24]. Z. Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of popula-tions," Science, vol. 366, no. 6464, pp. 447–453, 2019.

[25]. P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv preprint arXiv:1711.05225, 2017.

[26]. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 618–626, 2017.

[27]. F. Shamshad et al., "Transformers in Medical Imaging: A Survey," Med. Image Anal., vol. 88, 102802, 2023.

[28]. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Int. Conf. Learn. Representations (ICLR), 2015.

[29]. Y. Sun et al., "Classification of imbalanced data: A review," Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 4, pp. 687–719, 2009.

[30]. C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2818–2826, 2016.

[31]. J. P. Thornhill et al., "Monkeypox virus infection in humans across 16 countries—April–June 2022," N. Engl. J. Med., vol. 387, no. 8, pp. 679–691, 2022.

[32]. S. Umakanth, "Fine-tuning of pre-trained CNN models for Monkeypox disease detection," Mul-timed. Tools Appl., vol. 82, no. 1, pp. 1–17, 2023.

[33]. A. K. Uysal, "A novel CNN-LSTM hybrid model for monkeypox disease classification," Appl. Soft Comput., vol. 137, 110185, 2023.

[34]. A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NIPS), 2017.

[35]. A. Yinka-Ogunleye et al., "Outbreak of human monkeypox in Nigeria in 2017–18: a clinical and epidemiological report," Lancet Infect. Dis., vol. 19, no. 8, pp. 872–881, 2019.