A Deep Learning Framework for Arsenic Skin Disease Detection Leveraging Dual-Teacher Knowledge Distillation with a Depthwise-Separable Convolution and KAN-Based Lightweight Student Model

by

Md Humaion Kabir Mehedi
22266022

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
December, 2025

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at BRAC University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis partially contains material which has been accepted in a journal name "PLOS One".

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md Humaion Kabir Mehedi
22266022

Approval

The thesis titled “A Deep Learning Framework for Arsenic Skin Disease Detection Leveraging Dual-Teacher Knowledge Distillation with a Depthwise-Separable Convolution and KAN-Based Lightweight Student Model” submitted by

1. Md Humaion Kabir Mehedi (22266022)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science and Engineering on December 7, 2025.

Examining Committee:

Examiner:
(External)

Dr. Muhammad Firoz Mridha
Professor
Department of Computer Science
American International
University-Bangladesh

Examiner:
(Internal)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Arsenic poisoning in groundwater poses a major public health issue in Bangladesh. Chronic arsenic poisoning often leads to arsenicosis, a chronic disease characterized by cutaneous effects, including melanosis, leukocomelanosis and keratosis. Timely treatment of these lesions on the skin is important since prompt diagnosis and timely action are taken within the medical profession. However, this is challenging in rural areas. In most areas, there are no trained dermatologists, diagnoses tend to be subjective and there may be inadequate healthcare facilities. The proposed study involves the development of an explainable and lightweight deep learning framework for the automatic identification of skin diseases caused by arsenic, which will be computationally efficient and interpretable on a clinical scale. This system uses the dual-teacher knowledge distillation (KD) approach, in which two high-capacity models, InceptionV3 and Xception+InceptionM, are used to distill discriminative and contextual information to a small student model called Inception-Residual-KANNet (IR-KANNet). ArsenicSkinImageBD dataset was trained and validated using a stratified data split and 5-fold cross-validation to ensure class balance and model stability. The final evaluation found accuracy 0.9692, precision 0.9610, recall 0.9867 and F1-score 0.9737 on both infected and not-infected cases using stratified data split and a factor of 78.39% reduction in parameters over the teacher 1 and 35.29% in teacher 2 networks. These results indicate the appropriateness of the model for use in low-resource healthcare settings by providing convenient and reliable AI-based screening for arsenicosis detection. Grad-CAM++ and LIME make the model explainable and the resulting transparent heatmaps are consistent with the clinical regions in terms of lesions. This study is relevant for developing interpretable, efficient and domain-flexible medical AI models. It also provides a basis for further study of explainable knowledge distillation and edge-deployable solutions for the diagnosis of other dermatological diseases.

Keywords: Arsenic skin disease, Skin lesion classification, Inception-Residual-KANNet (IR-KANNet), Explainable deep learning, Dual Teacher Knowledge Distillation, Lightweight CNN, Kolmogorov–Arnold Network (KAN), LIME, Medical Image Analysis.

Dedication

It is our genuine gratefulness and warmest regard that we dedicate this work to all our loved ones for their love and inspiration.

Acknowledgement

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

CAD Computer-Aided Diagnosis

CNN Convolutional Neural Network

DL Deep Learning

$FLOPs$ Floating-Point Operations Per Second

$Grad - CAM + +$ Gradient-Weighted Class Activation Mapping++

$IR - KANNet$ Inception-Residual-KANNet

KAN Kolmogorov–Arnold Network

KD Knowledge Distillation

$LIME$ Local Interpretable Model-Agnostic Explanations

ROC Receiver Operating Characteristic

$SepConv$ Depthwise-Separable Convolution

XAI Explainable Artificial Intelligence

Chapter 1

Introduction

This chapter introduces the background and motivation behind the study, focusing on the widespread public health threat posed by arsenic-contaminated groundwater and its characteristic dermatological manifestations. It describes the problem of early diagnosis in resource-constrained rural environments and how recent innovations in deep learning can provide a solution for the automated detection of skin lesions. The problem statement, research gaps and objectives are also provided in this chapter, which guides the development of an explainable, lightweight diagnostic framework for arsenic-induced skin diseases.

1.1 Background

Arsenic contamination of groundwater continues to pose a significant community health concern in South Asia, especially in Bangladesh, India and Nepal, where tube wells are the main source of water used by millions of people. The World Health Organization (WHO) has highlighted the chronic consumption of arsenic-polluted water as one of the main environmental health risks [1]. A systematic review [2] indicated that chronic exposure of drinking water to asbestos and arsenic is directly linked to non-cancerous and cancerous alterations of the skin, with dose-sensitive increases in risk that are already observed even at low to moderate levels of exposure. These skin presentations tend to be the first and most noticeable indications of arsenic poisoning. Chronic arsenic poisoning is frequently found on the skin owing to its distinctive dermatologic manifestations of hyperkeratosis, raindrop hyperpigmentation and diffuse melanosis, which serve as some of the first clinical findings of arsenic toxicity [3]. Arsenic and microbial pollution of drinking water and industrial pollutants pose an alarming danger to human health in Bangladesh where the overwhelming contamination of drinking water remains a major problem. Chowdhury [4] identified people with arsenic-related skin diseases by a team and a medical group across 261 villages in 80 police stations within 33 districts. These districts were selected from 50 known arsenic-affected districts where groundwater arsenic levels were higher than 50 $\mu g/L$. The investigation found arsenic-affected patients in 222 villages under 69 police stations across 31 of the 33 surveyed districts. In total, 18,840 people, including children, were examined and 3,725 individuals with arsenic-related skin problems were identified in the study. Among these patients, 1,885 were men, 1,542 were women and 298 were children. If children were included, approximately 19.77% of all examined individuals had arsenical skin lesions. Separately, 24.52% of of 13,976 adults and 6.13% of 4,864 children were affected. The children were under the age of 11 years and had been drinking arsenic-contaminated water in the affected villages. Among them, 298 children (6.12%) had visible arsenic-related skin lesions, whereas the rate for adults was much higher at 24.47%. Between 1995 and 2000, they collected and tested 33,092 tube-well water samples from four major geomorphological regions of Bangladesh: Hill Tract, Table Land, Flood Plain and Deltaic areas, covering all 64 districts. They found that 60 districts had arsenic levels higher than the WHO drinking-water guideline of 10 $\mu g/L$ and 50 districts had levels above the national maximum limit of 50 $\mu g/L$. The level of arsenic in groundwater varies across the country: some areas have very low concentrations, some are almost free from contamination, while others show extremely high arsenic levels. The map of Figure 1.1 illustrate how arsenic is distributed across the affected districts and summarizes the overall groundwater arsenic status in all 64 districts of Bangladesh.

Image-based diagnosis of arsenic-induced skin diseases is crucial for early detection, especially in regions with basic healthcare infrastructure but a shortage of dermatologists. The latest developments in deep learning (DL) and computer-aided diagnosis (CAD) have made it possible to perform automated skin image analysis with high accuracy and scalability and to provide high-speed and consistent screening of skin images even with a mobile gadget. Convolutional neural network (CNN)-based [5], [6] models can be effectively used to identify and classify the dermatological features of arsenic poisoning and provide a cost-effective scalable resource that is understandable and applicable to public health in arsenic-affected regions. In recent years, DL has emerged as a transformative approach for the diagnosis of arsenic-induced skin conditions, offering an accurate and reliable alternative to conventional dermatological assessments. Transformer-based models show effectiveness in arsenic-induced skin lesion classification that could never be achieved in the study [7] using natural images with more than 86% diagnostic accuracy, surpassing the use of conventional CNNs, with results being interpretable as clinical images based on Grad-CAM visual explanations. A fusion-based DL model that reaches high sensitivity and specificity in detecting arsenic-related skin abnormalities with field data in Bangladesh, ArsenicNet [6]. The combination of these studies points to modern deep learning scaling the diagnostic gap in resource-deficient and arsenic-impacted regions, using lightweight and explainable architectures, as they make medical image analysis in a mobile platform reliable.

1.2 Motivation

AI has become a revolutionary power in healthcare, allowing for early disease detection, accurate diagnoses and large-scale decision-making in various fields. Specifically, the use of deep learning algorithms, particularly convolutional neural networks (CNNs), have transformed the process of skin image classification and reached impressive success rates in detecting lesions, pigmentation abnormalities and dermatological diseases with no less success than professional clinicians. Nevertheless, the majority of existing deep models are black-box and computationally intense, unable to be applied into real-world healthcare systems, especially in low resource or rural settings where arsenic contamination continues to be one of the greatest health concerns of the population. These models can be computationally expensive to in-

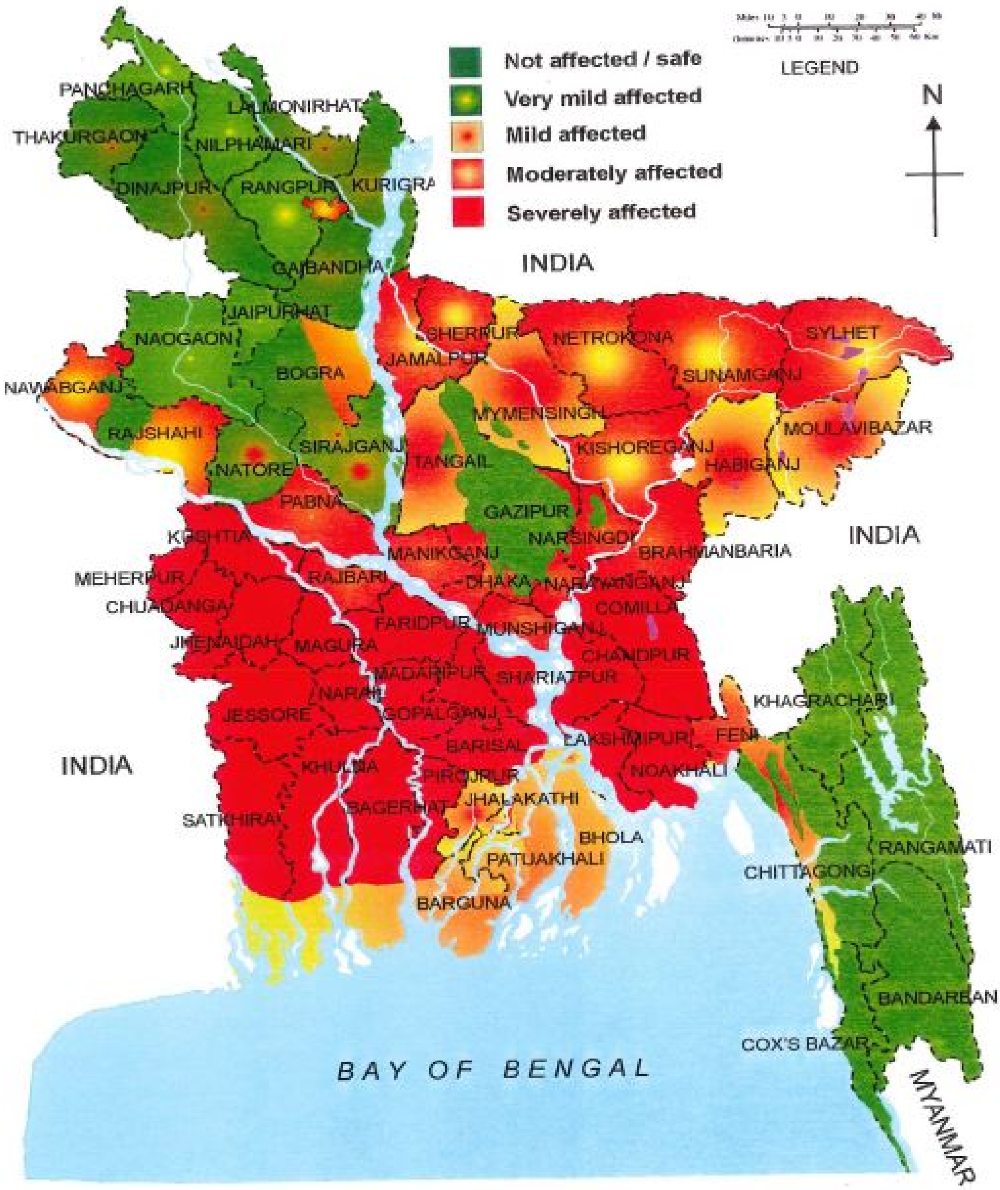


Figure 1.1: Geographical distribution of arsenic-affected groundwater regions across Bangladesh [4].

fer and many of them are not interpretable, damaging the clinician trust required and limiting their use in community-wide inquiries. This lack of clarity shows that there is a dire necessity to have a lightweight, effective and explainable deep learning model that can effectively identify arsenic-related skin lesions without sacrificing clarity and availability to medical professionals. Such an interpretable AI system would help advance the process of early diagnosis, decrease the possibility of disease progression and enhance overall population health outcomes in arsenic-affected areas.

1.3 Problem Statement

Arsenic poisoning of groundwater is still one of the worst problems for the environment and people in Bangladesh and other third-world countries. Persistent consumption of arsenic-polluted water leads to arsenicosis, which first exhibits obvious skin lesions, melanosis, leucomelanosis and keratosis, mostly on the palms, soles and trunk, among other parts [8], [9]. These initial skin symptoms become the initial clinical signs of toxicity of arsenic in the body and may develop to Bowen disease and squamous cell carcinoma when not properly treated [10].
Although the early diagnosis of arsenic-induced skin diseases is clinically significant, the diagnosis of arsenic-induced skin illnesses in rural or resource-limited regions is challenging because of the following drawbacks:

- **Shortage of trained dermatologists:**

  Many arsenic-contaminated areas do not have access to qualified doctors who can distinguish arsenic-related lesions from other dermatological consequences.

- **Subjectivity and variability in clinical observation:** Manual visualization is biased and relies on the experience of clinicians, lighting conditions and morphology of the lesions, resulting in inconsistent or slow diagnostic results.

- **Limited access to laboratory testing and imaging tools:**

  Biopsies, spectroscopy, or dermoscopic imaging are confirmatory tests that are invasive, expensive and uncommon in the primary care units of rural populations.

- **Complexity of existing deep learning models:**

  Although recent CNN architectures (e.g., VGG19, ResNet152V2, InceptionV3, Xception and EfficientNetV2B0) have proven to perform well across general medical imaging, they tend to be computationally expensive, memory-intensive and not interpretable; therefore, they cannot be deployed on low-power clinical or mobile systems.

- **Need for efficient and trustworthy AI systems:**

  Medical AI applications should also be capable of producing explainable predictions with high predictive accuracy, such that model outputs are labeled by medically meaningful visual features and are computationally viable to support real-time applications.

A balanced combination of diagnostic accuracy, computational efficiency and interpretability of a specific deep learning framework optimized to detect arsenic-induced skin diseases has yet to be achieved. Currently available high-capacity models have proven to be accurate but cannot be used in low-resource healthcare settings because they are complex, hardware-demanding and based on opaque prediction logic.
Therefore, this study fills an important research gap by creating a lightweight and interpretable deep learning model that can effectively classify arsenic lesions on human skin accurately, transparently, with low computational cost and is clinically useful. The proposed model will enable credible AI-assisted screening in rural health settings, where access to professional dermatological care and diagnostic infrastructure is inadequate.
This study assumes the organization of three closely connected subproblems, each responding to a core issue in the creation of an interpretable, efficient and implementable deep learning framework to identify arsenic-induced skin illness.

- **Performance–Efficiency Trade-off:** One of the main challenges is creating a lightweight model that can be as accurate as a large-capacity CNN like InceptionV3 and Xception and be computationally efficient enough to be used in screening in real time. The parameters stored in traditional deep models can reach tens of millions, substantially increasing inference latency and energy consumption and thereby limiting their deployment on mobile and low-resource clinical devices. Thus, this subproblem examines how architectural optimizations (e.g., depthwise-separable convolutions, multi-scale feature extractions and multi-scale attention mechanisms) can alleviate the computational load without negatively impacting the classification performance. It aims to find the best accuracy-complexity trade-off, which allows for strong inference under realistic conditions.

- **Knowledge Transfer for Low-Resource Learning:**

  Since the availability of labeled skin images of people affected by arsenic is relatively low, there is a risk of overfitting and ineffective generalization associated with direct training of deep models. This study uses knowledge distillation (KD) as an approach to discriminative knowledge transfer between large, pretrained teacher networks and a small student model to alleviate this problem. The aim is to make the lightweight model suitably predict with the level of accuracy of teachers via learning by scaling temperature and use of soft-target loss and weighted soft-hard loss as well as optimization by cross-validation.

- **Model Explainability for Clinical Adoption:** For a clinically reliable AI-assisted diagnosis, the prediction must be understandable to medical professionals and must have a visual explanation. This subproblem aims to integrate explainable AI (XAI) techniques, including Grad-CAM++ and LIME, into the KD framework to illustrate lesion-relevant areas that correspond to dermatological arguments. The component facilitates both the accuracy and transparency and interpretability of predictions by aligning the saliency maps in the model with medically anatomized pigmentation and keratotic localities to identify medically anatomized zones and improve prediction accuracy. This interpretability brings about a feeling of confidence among clinicians to enable the adoption of artificial intelligence-based arsenic screening instruments in rural and low-resource healthcare settings.

1.4 Research Objectives

This study primarily aimed to design a deep learning–based classifier and evaluate its explainability for classifying arsenic-induced skin ailments from digital images. The proposed framework leverages knowledge distillation (KD) to transfer knowledge of high-capacity teacher networks to a lightweight student model, thereby achieving an optimal trade-off between predictive accuracy, computational efficiency and interpretability. In addition, along with the stratified data split, k-fold cross-validation was used as the primary evaluation protocol to ensure statistical robustness and enhance the generalizability of the model.

- **Development of Image-Based Classification Framework:**
  The initial aim was to create a standardized image-based classification pipeline that could differentiate between normal and arsenic-affected skin images. This pipeline was used as the base pipeline for further model training, evaluation and explainability analysis. To provide consistency and reproducibility, every image will be subjected to a uniform image preprocessing procedure that includes resizing to 224 x 224, normalizing the intensity and label-preserving data augmentation strategies, including random rotations and horizontal/vertical flips. The augmentations are aimed at overfitting reduction and generalization of the interpretation in settings with small and heterogeneous medical imaging data. It also requires a reproducible and stable baseline classification framework that can provide a reliable fold-wise performance that can serve as a controlled place to benchmark models and make comparisons with them.

- **Design of Lightweight Student Network:**
  To develop a computationally efficient convolutional neural network with an appropriate accuracy-efficiency trade-off that can be deployed on a mobile or edge device. The architecture includes depthwise separable convolutions, residual blocks and compact Inception modules to reduce the number of parameters while preserving the feature learning capacity. Non-linear representation learning is further enhanced through Kolmogorov–Arnold Network (KAN) modules applied to the high-level feature representations. The adjustment of the width and depth multipliers of the network is aimed at reducing the number of parameters and floating-point operations per second (FLOPs) without affecting the diagnostic accuracy. The objective is to simplify the models by at least 60-80% compared to more traditional CNN baseline models like InceptionV3, VGG19, or EfficientNetV2B0 and compete effectively.

  A compact CNN with an optimal trade-off between computation and diagnostic performance was designed to enable potential deployment in fieldable robotic systems for arsenic lesion screening.

- **The Implementation of Knowledge Distillation (KD):**
  To perform a knowledge distillation mechanism by enabling the transfer of latent knowledge between a high-capacity teacher model and a lightweight

student model. The teacher network will be InceptionV3 and hybrid Xception+InceptionM, as they have already been demonstrated to be effective in complicated image classification medical tasks. The KD process utilizes temperature-scaled soft targets and a weighted composite loss function that combines soft (generated by the teacher) and hard (ground-truth) labels. The temperature (T) and the weighting coefficient were tuned empirically through multiple training runs and the combination that yielded the best cross-validation performance of the student model was selected.

- **The Integration of Explainable Artificial Intelligence (XAI):**
  The incorporation of explainable artificial intelligence (XAI) techniques will make the classification task simpler and more transparent. Gradient-weighted class activation mapping (Grad-CAM++) and local interpretable model-agnostic explanations (LIME) were used to create visual-saliency maps that indicate the image regions with the strongest impact on the decision-making process of the model. To ensure that these interpretability tools reflect the existence of prognostically and relevant lesion areas, namely hyperkeratosis, pigmentary clustering and diffuse melanosis, qualitative evaluation should be performed in line with accepted dermatological patterns.

  An interpretable model that can create clinically relevant visual explanations can lead to better medical trust, transparency and the possibility of real-world adoption as part of the diagnostic process.

- **Fixed Stratified Data Split and Cross-Validation:**
  A comprehensive and statistically rigorous assessment protocol was developed based on a fixed stratified data split (80% for training, 10% for validation and 10% for testing) and ($k = 5$) fold cross-validation for the assessment strategy. Under the cross-validation protocol, every sample is used in training and validation, which results in a strong approximation of model generalization but reduces sampling bias owing to the fixed train/validation/test split. Performance will be measured using typical evaluation measures, such as accuracy, precision, recall, F1-score, receiver operating characteristic curve (ROC) and confusion matrix for both protocols.

This study will provide a programmatic fulfillment of the following goals: The proposed research will provide a lightweight, explainable and statistically confident deep learning system for the automated detection of arsenic-induced skin diseases. The proposed framework aims to bridge the extensive gap between high-performance AI studies and the shortcomings in high-resource regions where arsenic has a significant impact and clinical usability, such as high-quality explanations.

1.5 Contributions

This research is an original contribution to the field of explainable and efficient deep learning for the analysis of medical images, especially skin diseases caused by arsenic. The study incorporates architectural novelty, knowledge distillation systems and explainability; thus, the next generation of lightweight, transparent and statistically validated diagnostic artificial intelligence is one step ahead.

1. **Dual-Teacher Knowledge Distillation Framework:** Development of a dual-teacher knowledge distillation (KD) framework, where two large teacher models are employed collectively to oversee a small student network.

    - **Teacher 1:** The hybrid Xception+InceptionM model has depthwise-separable fine-grained patterns and multi-scale spatial differences. The reason for using it is that such a fusion architecture always provides high discrimination of delicate lesion textures; therefore, it is a perfect high-quality source of knowledge to be used in distillation.
    - **Teacher 2:** InceptionV3 is known to have a hierarchical vision of the texture, color and morphology of medical images. It is applied since its stable and well-generalized feature extraction is complementary to Teacher 1, which allows the student to be trained on a more comprehensive and robust set of lesion characteristics.

    The soft probabilistic outputs of the teachers are confidence-weighted and dynamically averaged to produce enriched soft targets for distillation. The KD loss functional is an adaptation of the Kullback-Leibler divergence (and categorical cross-entropy) weighted by an adaptive α weight, which has temperature scaling. The knowledge-transfer system in this ensemble allows the student to achieve the accuracy of the teacher at a greatly reduced computational cost and can be deployed

2. **Lightweight Student Architecture:** This study designed a custom compact convolutional neural network named Inception-Residual-KANNet (IR-KANNet). The architecture also uses several efficiency-based and attention-focusing elements. SepConv is a deep separable convolution that minimizes the overall quantity of parameters and the number of floating operations. Small Inception blocks enable feature extraction in a multi-scale manner. Spillover relationships enhance the stability of the gradients during training. Kolmogorov–Arnold Network (KAN) blocks use attention in the latent space that is learned and of Gaussian basis. Overfitting is reduced and at the same time, there is an improvement in generalization with the help of layer normalization and dropout.

3. **Explainable AI Integration:** This was performed using two complementary Explainable AI (XAI) strategies to explain the behavior of the model at the feature-map and pixel levels:

    - Grad-CAM++ produces class-specific activation heatmaps in order to visualize salient regions of lesions at the convolutional feature scale.
    - LIME offers local interpretability based on instances, which can obtain fine-grained explanations with systematic perturbations of input pixels and measuring their effect on classification.

    This visual analysis confirmed that the teacher and student models paid consistent attention to clinically accessible areas, such as hyperpigmented and keratotic areas, which supports the concept of transparency and clinical reliability.

4. **Experimental Evaluation Strategy:** To achieve reproducibility, the framework uses a stratified data split (80% for training, 10% for validation and 10% for testing) and ($k = 5$) fold cross-validation for the assessment strategy. accuracy, precision, recall, F1-score, receiver operating characteristic curve (ROC) and confusion matrix. In addition, we conducted a detailed ablation study to visualize the importance of each component in the proposed framework.

This study proposes a framework of interconnected methodologies to integrate architectural compression, dual-teacher knowledge distillation and explainable AI into one unified diagnostic tool. The suggested system promotes the state-of-the-art in lightweight and transparent medical imaging interventions classified with the help of the frontier of deep learning, both in terms of its approach and potential usefulness in medical drained rural healthcare systems and other areas of AI-useful performance.

1.6 Thesis Organization

The proposed thesis is divided into several chapters to make the proposal of a new framework structured and coherent. Chapter 1 presents the background, motivation, problem formulation and research objectives regarding the detection of skin diseases caused by arsenic. Chapter 2 reviews the literature on dermatological manifestations of arsenic, conventional machine learning methods, the newest deep learning models, lightweight architectures, knowledge distillation and explainable AI techniques. Chapter 3 presents the proposed methodology, which involves the general workflow, dataset preparation, selection of teacher and student models, IR-KANNet architecture design and strategy of knowledge distillation using two teachers. In Chapter 4, the results and discussion are introduced, the experimental environment (hardware and software), supervision, values of hyperparameters, performance analysis between the teacher and student models, cross-validation results, robustness testing, statistical analysis and visualizations of explainability. Chapter 6 presents the constraints of the present work, as well as the prospects for further research, including dataset scaling, novel approaches to distillation and actual edge implementation. Finally, Chapter 7 concludes the thesis by summarizing the major contributions, revealing the importance of the proposed framework and supporting its further use in medical image analysis.

Chapter 2

Literature Review

This literature review summarizes some of the important foundations and advancements that motivated this thesis. It starts with a description of the public health context of arsenic exposure and some of the characteristic dermatological features of exposure. Next, it surveys the application of traditional machine learning and modern deep learning on medical imaging with an emphasis on the diagnosis. The review then focuses on lightweight and attention-based architectures, which are designed for efficiency and success with strengthening features and knowledge distillation as a way to compress models with less of loss of accuracy. Explainable AI methods are influencing to make clinical trust and sometimes it's discussable. Finally, the section brings together observed limitations in order to define the research gap dealt with in this work.

2.1 Arsenic Exposure and Dermatological Manifestations

Groundwater arsenic pollution is one of the most serious environmental and public health problems of South Asia, especially Bangladesh, where more than 20 million people drink arsenic-contaminated water of 10 ug/L [11]. Chronic ingestion of arsenic-contaminated water causes arsenicosis, a multisystem disease that causes oxidative stress and cellular damage to the liver, lungs, kidneys and skin. Among its various pathological effects, dermatological manifestations such as hyperpigmentation and keratosis are the earliest and most obvious biomarkers of chronic exposure to arsenic. Cutaneous signs are usually precursors of systemic diseases and offer crucial diagnostic opportunities for early treatment. However, manual diagnosis of arsenicosis by visual inspection or biopsy is highly subjective and depends on the expertise of the dermatologist, which is rare in rural and non-resource areas. Consequently, automatic image-based diagnosis systems aimed at objectively identifying arsenic-induced lesions are of interest to lithologists using deep learning and computer vision. Such models have the potential to significantly impact community-level screening by facilitating cost-effective, large-scale and early detection of arsenicosis to reduce its long-term public health burden. Such models can significantly impact community-level screening by facilitating cost-effective, large-scale and early detection of arsenicosis to reduce its long-term public health burden (see Figure 2.1).
Because of the difficulties in diagnosing arsenicosis patients, deep learning frame-

Figure 2.1: Representative clinical signs of chronic arsenic poisoning on the palms, including diffuse pigmentation and hyperkeratosis lesions.

works have been used in some studies to automatically classify skin lesions for different purposes, focusing on accuracy, generalizability and computational efficiency. Jobayer et al. [12] proposed ARS-CNNSA, which is a CNN model augmented by self-attention for accurate feature extraction and prevention of overfitting. Therefore, compared to well-known models such as VGG16, MobileNetV2 and InceptionV3, it reached 91% accuracy and 90% F1-score while outperforming conventional pre-trained networks while keeping the models small. In response to the need for optimization in terms of accuracy and efficiency, Saha et al. [13] used EfficientNetB1 for the automatic classification of healthy and arsenic-affected skin and achieved 94.69% accuracy. Additionally, their framework achieved robustness to different skin textures after extensive data augmentation, which served as a solid basis for implementation in the real world. Following the flexibility of transfer learning, Aakash et al. To address this defect, Newaz et al. [7] presented a transformer-augmented learning pipeline fed by more than 11,000 mobile phone-captured images with 20 lesions of arsenic. Their comparative analysis with other models showed that the Swin transformer with LIME and Grad-CAM for model explainability attained the highest performance (86% accuracy). Importantly, their web-based diagnostic tool confirmed that real-time community-based screening was feasible using devices capable of being accessed by everyone, all the time, thus allowing for the convergence of laboratory- and community-based health care.
Further to these developments, another breakthrough contribution from Emu et al. [14] is the rollout of the first dataset on arsenic skin disease, called the Arsenic-SkinImageBD Disease Dataset, including 8,892 balanced images after augmentation from four arsenic-affected areas of Bangladesh. This dataset overcomes one of the bottlenecks for the generalization of deep learning models in arsenicosis research, which is the data scarcity barrier. Although its annotations do not consider the guidance of subclass labels, it offers an important ideal groundwork for future integration into models, as well as for benchmarking and future AI-based clinical studies. Collectively, these studies constitute a continuum of holistic development within the field of dermatological research with respect to arsenic, ranging from data generation and data augmentation to attention-based augmentation, transfer learning

and transformer-based models. Together, they provide a convergence of fragments of accuracy, interpretability and accessibility and establish the basis for a series of scalable, artificial intelligence-based arsenicosis diagnostic systems for transformative public health screening of all forms of rural low-income environments. These comparative insights are presented in Table 2.1, which provides an overview of the reviewed models, their performance and their limitations.

Table 2.1: Summary of reviewed studies on AI-based arsenicosis detection.

Reference	**Model**	**Results**	**Limitations**
[12]	ARS-CNNSA (CNN + Self-Attention)	Achieved 91% accuracy and 90% F1-score; outperformed VGG16, MobileNetV2 and InceptionV3 with smaller model size.	Region-specific dataset; limited generalization across populations.
[13]	EfficientNetB1	94.69% accuracy; demonstrated robustness to diverse skin textures through extensive augmentation.	Limited dataset diversity; lack of cross-domain validation.
[7]	Transformer-based pipeline (Swin Transformer + LIME + Grad-CAM)	86% accuracy; interpretable framework; enabled real-time web-based community screening.	Moderate accuracy; dependent on quality of mobile-captured images.
[14]	ArsenicSkinImageBD Dataset	Balanced dataset from four regions; provided foundation for generalizable arsenicosis models.	Lacks subclass annotations; limited lesion-specific granularity.

2.2 Traditional Machine Learning (ML) and Deep Learning in Medical Imaging

Early studies on detection of arsenic-induced skin lesions were mainly based on classical machine learning algorithms that used manually designed features of images to classify the skin lesions into affected and healthy regions. Frequently, feature descriptors such as the Gray-Level Co-Occurrence Matrix (GLCM), local binary patterns (LBP) and color histogram descriptors were combined with classifiers such as support vector machine (SVM), K-nearest neighbor (KNN) algorithm and random forest. These strategies and dictionaries laid the basis for improving the automated analysis of arsenic lesions by emphasizing the statistical characterization of texture and feature-based classification efficiency. Both Almustafa et al. [15] and Setegn and Dejene [16] addressed the problem of detecting skin diseases with automated systems using the classic machine learning pipeline, but with different modalities in data and focus on models. Almustafa et al. relied on histopathological biochemical features (i.e., biopsy), where handcrafted creations such as erythema, itching and the Koebner phenomenon were critical predictors. Their comparison of multiple algorithms, including random forest and logistic regression, proved that the SGD Classifier was the most effective (with an accuracy of almost 99%).
In a similar quest for diagnostics, Setegn and Dejene focused on image datasets of dermatology, but had a similar framework to Hajii's focus on model benchmarking and interpretability. Their experiments on random forest, decision tree,

KNN, logistic regression and SVM reaffirmed the power of tree methods, as random forest achieved 97.9% accuracy. The methodological similarities between the two studies, in terms of manual feature extraction, the need for limited and often imbalanced datasets and the composition of the task for stable classical classifiers, point to a common limitation of scalability and representation power. Going beyond these limitations, Balasundaram et al. [17] combined traditional ML and deep learning by introducing a genetic algorithm-optimized stacking ensemble that combined ResNet50V2, DenseNet121 and CNN architectures against DermNet and HAM10000 data. Their hybrid design achieved 91.73% accuracy in HAM10000, which could be considered as generalizing and adaptability, which had problems before with handcrafted systems. Taken together, these studies show the evolution of feature-engineered, algorithmic experimentation advances to the development of optimization-based hybrid ensembles, a general attribute of the evolution of machine learning in dermatological diagnostics from manually tuned algorithms for diagnosis to data-rich models that are self-adjusting for prediction.

In the past few years, the use of machine learning techniques for environmental and health-related diagnostics has shown sufficient versatility to solve problems in many areas (from the modeling of community awareness to disease prediction to real-time monitoring). Singh et al. [18] and Lombard et al. [19] have some methodological parallels in utilizing machine learning approaches in arsenic-risk prediction at different scales and settings. Singh et al. predicted arsenic awareness among the populations of India affected by contamination in rural areas using logistic regression and six classical classifiers, such as SVM and random forest, based on socioeconomic, demographic, and behavioral data. Their models showed that education, caste, occupation, sanitation habits and social trust were the major determining factors, with computational methods such as SVM and random forest performing the best. Similarly, Lombard et al. used boosted regression tree (BRT) and random forest classification (RFC) to predict the probability of contamination in groundwater wells in the United States using hydrogeological and climatic features. Both studies took advantage of the non-linear ability of random forest to learn: Singh's to render patterns of awareness of behavior and Lombard's to model geochemical risk gradients. While Singh's work emphasized social answers to arsenic, Lombard's large-scale framework responded to environmental foreseers, stressing the reactivity of ML in socio-environmental well-being modeling, regardless of limitations of scale notes together with regional change.

Transitioning into environmental to biomedical fields, Balaji et al. [20] and Souri et al. [21] addressed ML integration for healthcare diagnostics and monitoring that share a common objective of better detection of early stage disease, as well as both prevention and personalized healthcare implementation. In the skin lesion segmentation task, Balaji et al. developed a dynamic graph cut-based segmentation framework with naive bayes classifier for skin lesion categorization using ISIC 2017 dataset with more than 91% accuracy for each type of skin lesion compared to FCN and SegNet architectures. Their approach involved accurate feature extraction and probabilistic classification, which was used to deal with complex dermatological imagery. Complementing this image-based focus, Souri et al. created an IoT-based student health monitoring system based on sensor data and ML algorithms, for which SVM achieved 99.1% accuracy, which was better than Decision Tree, random forest and Multilayer Perceptron. Both studies showed the potential of combining ML with

domain-specific systems: Balaji used hybrid segmentation for visual diagnostics and Souri used predictive analytics for real-time physiological monitoring. Together, these four studies demonstrate the broad adoption of machine learning in public health, from social and environmental predictive health to sensor-based personalized healthcare, which is revealed to be an adaptive movement towards data-driven, context-aware intelligent systems for public and self-informed health.
Recent refinements in medical diagnostics have led to the synthesis of machine learning, deep learning and explainable AI (XAI) frameworks, all of which seek to trade off between predictive quality and interpretation. This dual focus is common to Arravalli et al. [22] and Dharmarathne et al. [23], which, however, have different diseases and datasets. Arravalli et al. constructed a stacking model with random forest, logistic regression, CatBoost, LightGBM and XGBoost to predict breast cancer based on UCTH dataset. They also used five XAI models, including SHAP, LIME, Eli5, QLattice and Anchor, to explain important predictors such as tumor size and involved nodes, with an F1-score of 84. Similarly, Dharmarathne et al. focused on interpretability in the diagnosis of diabetes using SHAP-based feature explanation on the Pima Indians dataset, with glucose, BMI and age as the most dominant predictors, with an XGB of 80 and 77 percent training and testing, respectively. The two articles show how essential transparency is in clinical AI but have the same drawback of small, localized datasets, which limits generalization. Although these articles emphasize the interpretability and moderate sizes of datasets, Rai and Yoo [24] worked in a more inclusive meta-analytic framework, analyzing 130 articles (2018-2023) on brain, lung, skin and breast cancers and comparing machine learning with deep learning results. Their results showed that the ML models not only showed near-perfect accuracies (as high as 99.89), but deep learning was always better at feature abstraction and had 100 percent accuracy in certain situations. Variability between datasets, however, particularly in the case of skin cancer (28.8%), was an indication of the necessity of standardized, hybrid ML-DL systems, just as the hybrid aspirations of Arravalli and Dharmarathne.
Building on this frontier of hybridization, Gupta et al. [25] addressed the problem of thyroid disease detection by training a framework of ML that was optimized with differential evolution (DE) to use cGANs to augment their data in ten thyroid classes. Their DE-tuned AdaBoost achieved the highest accuracy of 99.8%, but at the expense of the complexity of computations. Collectively, these research works reveal a progressive fusion of ML, DL and XAI paradigms, from interpretable ensemble (Arravalli, Dharmarathne) and evaluative synthesis (Rai and Yoo), to evolutionary optimization (Gupta), which demonstrate a unary trend of explainability, adaptivity and data efficiency of healthcare intelligence ready for clinical deployment to scale. Based on the literature considered, it can be readily seen that the initial research in the field of medical and environmental diagnostics, including arsenic awareness modeling for dermatological and clinical disease classification, was controlled by conventional machine learning and ensemble-based frameworks. The approaches proved highly interpretable and had early diagnostic achievement, but in many cases, they had difficulties with the scalability of data, representational richness and fitting nonlinear data patterns.

Table 2.2: Summary of reviewed works on traditional ML for medical and environmental diagnostics

Reference	Model	Results	Limitations
[15]	Classical ML (SGD, RF, LR) with handcrafted histopathological features	SGD Classifier achieved ∼99% accuracy using biopsy-based features (erythema, itching, Koebner phenomenon).	Limited data; handcrafted features; poor scalability and representation power.
[16]	RF, DT, KNN, LR, SVM on dermatological images	Random Forest achieved 97.9% accuracy; validated the efficiency of tree-based models.	Small, imbalanced dataset; limited generalization; manual feature dependence.
[17]	Genetic Algorithm–Optimized Stacking Ensemble (ResNet50V2 + DenseNet121 + CNN)	91.73% accuracy on HAM10000; improved adaptability and generalization compared to handcrafted systems.	High computational cost; dataset-specific performance.
[18]	Logistic Regression, SVM, RF for arsenic awareness modeling	Identified education, caste and sanitation as key factors; RF and SVM performed best.	Socio-behavioral bias; regional specificity; lacks medical imaging integration.
[19]	Boosted Regression Tree (BRT) + Random Forest for groundwater contamination	Accurately predicted contamination probabilities using hydrogeological and climatic data.	Regional bias; limited transferability to other geochemical zones.
[20]	Dynamic Graph Cut segmentation + Naive Bayes (ISIC 2017 dataset)	>91% accuracy; improved lesion segmentation over FCN and SegNet architectures.	Dataset limited to specific lesions; lacks multi-modal validation.
[21]	IoT-based health monitoring system + SVM	Achieved 99.1% accuracy; superior to DT, RF and MLP; strong real-time prediction capability.	Limited to student datasets; lacks clinical-level validation.
[22]	Stacking Ensemble (RF, LR, CatBoost, LightGBM, XGB) + XAI (SHAP, LIME, Eli5, QLattice, Anchor)	F1-score = 84; key predictors: tumor size, involved nodes; interpretable breast cancer model.	Small dataset; limited external validation.
[23]	XGBoost + SHAP for diabetes diagnosis	Training accuracy 80%, test accuracy 77%; glucose and BMI most influential features.	Dataset limited (Pima Indians); lacks population diversity.
[24]	Meta-analysis (130 studies) comparing ML vs DL in cancer prediction	DL models achieved up to 100% accuracy; ML up to 99.89%; highlighted hybrid model need.	Dataset variability (skin cancer 28.8%); inconsistent evaluation metrics.
[25]	DE-optimized ML framework + cGAN data augmentation + AdaBoost	Achieved 99.8% accuracy for thyroid classification; robust synthetic data generation.	High computational complexity; limited interpretability.

These limitations have inherently prompted the shift of paradigm to deep learning structures, capable of learning hierarchical and abstract visual representations and

hence provided more robust, generalizable and clinically reliable systems to detect arsenic lesions. These findings are summarized in a systematically planned way in Table 2.2, which can identify the comparable performances and limitations of traditional ML techniques.
As the weaknesses of classical machine learning became apparent, including restrictions on hand author features, inadequate scalability and deficient ability to adapt to the intricate visual structures of medical and dermatological imagery, deep learning has become a disruptive substitute in medical and dermatological imaging. In contrast to classical models, deep neural networks automatically derive hierarchical and abstract representations of raw data, allowing for more accurate, robust and extrapolation to a broad range of imaging conditions. Their capacity to identify detailed spatial and textural characteristics has contributed significantly to lesion detection, segmentation and classification. As a result, deep learning has emerged as the motor of current diagnostic intelligence, evoking increased clinical integration and research interest. The recent advances in dermatological deep learning have shown the increasing complexity and versatility of convolutional and transfer learning models on automated diagnosis.
The vision of developing a skin disease detector based on a multi-model or ensemble-based deep framework is common to Badr et al. [26], Akram et al. [27] and Dhanushraju et al. [28], as each of them aims at achieving a higher sensitivity and scalability to various datasets. Badr and Akram both applied transfer learning to reuse a trained convolutional network: the model of Badr with the Xception architecture to classify five major skin conditions and the ensemble of SkinMarkNet by Inception, Xception and ResNet to classify lesions of monkeypox. Although the method used by Badr resulted in a 95% accuracy and a 99.4% AUROC, the method used by Akram resulted in 90.62% accuracy, which shows that the transfer learning approach can be modified to use in broad as well as specific dermatology. However, both were limited in the same way in terms of data reliance and calculation intensity, highlighting that model scalability remains a problem. Along this same developmental path, Dhanushraju et al. devised a multi-level multi-class deep learning model that combined the information of multiple sources, such as HAM10000, ISIC, Arsenic Healthy Skin and DermNet and augmented its dataset to over 12,000 images. Their deep hierarchical structure was 91.2% accurate and 95.36% specific, fine-grained lesion differentiation but required large amounts of computational resources. Together, these studies reflect the developments in the design and use of deep learning for dermatological imaging from transfer learning-based classification to hierarchical multi-level settings, indicating that while modern-day models provide better precision and adaptability, they maintain a balance between performance, efficiency and data diversity as the foundation for more performance-efficient, clinically deployable dermatological diagnostic systems.

Deep learning in dermatological and other fields of medical imaging is also slowly evolving towards more adaptive, explainable and contextual systems, as in several recent works by Girmaw [29], Radhika et al. [30], Yang et al. [31] and Cao et al. [32]. Both researchers, Girmaw and Radhika, emphasized the need to improve classification accuracy by optimizing the model. However, Girmaw's transfer learning method was also applied in the animal skin disease recognition problem across cattle, sheep and goats, resulting in 99.01% accuracy using the EfficientNetB7 model. Coincidentally, Radhika's Self-Reliant ResNet (SR-ResNet) improves melanoma detection by Zoutendijk's acceleration for faster, stable convergence by processing 10,000 dermoscopic images. Each demonstrates the ability of deep learning to learn complex skin texture patterns that work outside the limits of handcrafted features that can be used to analyze spatial information. Multi-modal and also generative paradigms: Yang's feature fusion, anatomical metadata and fuzzy ensemble aggregation with MobileNetV3-Large to boost inflammatory skin disease diagnosis (F1=0.8) emphasize the rightness and interpretability, context sensitivity; Cao's GAN-based MIM (occluded image reconstruction) model for mpox detection: AUROC 0.8237, proves to be robust to noisy clinical examples. A common feature of all four studies is the effective deployment of deep learning to improve diagnostic accuracy and generalizability, whether via transfer learning, optimization, ensemble fusion, or generative reconstruction.

However, they also have intrinsic issues, such as computational cost, data dependence and cross-domain generalization as their constraints. Collectively, they represent the shift towards multi-modality classifiers from single ones to more holistic, data-driven architectures that can capture the complexity and context to truly advance dermatological intelligence from being scalability-agnostic and clinically uninterpretable to being real-time and clinically interpretive. In the past few years, the application of deep convolutional kernels to attain greater generalization, accuracy and efficiency in classification has become a constant in dermatological imaging research, including skin cancer classification. Ashfaq et al. [33], Naeem et al. [34], Ozdemir and Pacal [35] are all included in this direction through hybrid, attention-based and CNN-based models with a focus on dermoscopic image classification. Both Ashfaq and Naeem used large-scale ISIC datasets by introducing multi-class lesion classification and efficient feature representation. Cfy's CNN generated by Ashfaq trained on over 25,000 ISIC Challenge images was able to distinguish between eight lesion classes with 90.18% accuracy with good precision, although the sensitivity was moderate (55.32%), which suggested that there were still certain difficulties detecting subtle lesions. Building on this framework, Naeem et al. proposed SkinDWNet with depth-wise dilated convolutions (DDCs) and residual blocks with unused features (FRBs) and gradient boosting for a more optimized classification. Their hybrid easy-to-use clinical system, SkinDWNet + GB, reported 97.04% and outperformed CNN-classical models and other previous state-of-the-art models, proving the capabilities of using architectural depth and feature refinement despite the increased computational cost. In addition, Ozdemir and Pacal extended this work further by integrating ConvNeXtV2 blocks with separable self-attention layers for a balance between local texture information and global region attention, achieving an accuracy of 93.48% and an F1-score of 91.82% on the ISIC 2019 data. Consequently, while the paper by Ashfaq and Naeem tackled how to maximize the discriminative capacity by maximizing convolutional depth and hybridization, Ozdemir and Pacal focused on contextual awareness and efficiency. Together, these papers tell the story of the evolution toward the next deep learning best architecture –using hybrid/attention-based models-that steadily improves over past progress towards clinical resilience for skin cancer classification and diagnosis while preserving interpretability despite class imbalance. The recent developments in hybrid deep learning have speeded up the evolution of the dermatological image classification, with an increasing focus on complementary model fusion (CMF) and

Table 2.3: Summary of deep learning and hybrid models in dermatological diagnostics

Reference	Model	Results	Limitations
[26]	Xception-based Transfer Learning CNN	95% accuracy, 99.4% AUROC; classified five major skin conditions effectively.	High computational cost and dataset-specific dependency.
[27]	SkinMarkNet (Inception + Xception + ResNet Ensemble)	90.62% accuracy for monkeypox lesion classification.	Data reliance; computationally intensive; limited scalability.
[28]	Multi-level Multi-class Deep Model (HAM10000, ISIC, Arsenic Skin, DermNet)	91.2% accuracy, 95.36% specificity; improved lesion differentiation.	Requires large computational resources; data-heavy framework.
[29]	EfficientNetB7-based Transfer Learning	99.01% accuracy in animal skin disease recognition across species.	Domain-specific dataset; limited generalization to human dermatology.
[30]	Self-Reliant ResNet (SR-ResNet) with accelerated convergence	Improved melanoma detection with 10,000 dermoscopic images; fast, stable training.	Dataset confined to specific lesion types; lacks interpretability layer.
[31]	MobileNetV3-Large + feature fusion + fuzzy ensemble	F1-score $\approx$ 0.8 for inflammatory skin disease; emphasized interpretability and context sensitivity.	Moderate performance; ensemble fusion increases model complexity.
[32]	GAN-based MIM (occluded image reconstruction)	AUROC = 0.8237; robust against noisy and occluded clinical images.	High training cost; limited dataset diversity.
[33]	CNN trained on 25K ISIC Challenge images	90.18% accuracy; strong precision; moderate sensitivity (55.32%).	Struggles with subtle lesion detection; class imbalance.
[34]	SkinDWNet (DDC + FRB) + Gradient Boosting	97.04% accuracy; surpassed prior CNNs and SOTA models.	Computationally intensive; complex architecture.
[35]	Hybrid (InceptionV3 + DenseNet121, weighted sum rule)	92.27% accuracy; >92% sensitivity & specificity.	High-resolution input cost; multiple inference overhead.
[36]	Hybrid (VGG19 + ResNet18 + MobileNetV2 + SVM/DT/NB/KNN)	ResNet18+MobileNetV2-SVM achieved 92.87% accuracy (ISIC 2018).	Computationally heavy; scalability limited by model chaining.

multi-stage feature refinement (SR) in the context of robust melanoma and skin cancer classification. Three studies on hybrid structures combining convolutional backbones with improved fusion methods to improve diagnostic accuracy have been conducted: Akter et al. [36], Shakya et al. [37] and Mateen et al. [38]. Both Akter and Shakya used two-stage models that used pre-training network penetrations with either classical or modern classifiers. Akter combined InceptionV3 and DenseNet121 into a weighted sum rule, obtaining 92.27% accuracy and over 92% sensitivity and specificity and verified that taking advantage of the similarity features of different models can provide better performance. Similarly, hybrid models using VGG19, ResNet18 and MobileNetV2 combined with SVM, Decision Tree, Naive Bayes and KNN were proposed by Shakya et al., where ResNet18-MobileNetV2-SVM was deemed the best model with 92.87% accuracy on the ISIC 2018 dataset. In addition to showing improved lesion discrimination, both studies underline that the joint embedding of various feature spaces comes at the cost of increased computational cost posed by high-resolution input as well as multiple sequential inferences per model. Going further, Mateen and other peers included a triple-hybrid architecture of U-Net, Inception-ResNet-v2 and Vision Transformer for segmentation and classification on the ISIC 2020 data achieved 98.65% accuracy with 99.2% sensitivity - far better than previous versions of hybrids. Additionally, a Vision Transformer was added for contextual understanding, which contributed to improvements in the HAM10000 validation. Together, these studies show that deep learning is now moving towards integrative, multi-model architecture combining convolutional precision with transformer-level contextual reasoning, a third key current concern: scalability (accuracy, computational efficiency and clinical applicability) of dermatological AI systems. These various advancements are systematically compiled in Table 2.3 and provide a good comparison of model accuracies and constraints.

2.3 Lightweight and Attention-Based Deep Learning Architectures

Lightweight deep learning architectures have recently attracted much interest in the medical and medical imaging fields for their capabilities of providing high diagnostic efficiency while still having tight computational hypermarkets. Among the most recent contributions, the research of Asif et al. [39], Khan et al. [40] and Hossen and Mondal [41] have something in common in constructing tiny but high-performing CNN models optimized for resource-constrained environments. Asif et al. put forward a parameter efficient framework called SKINC-NET trained on HAM10000 for multi-class skin cancer detection. By using the dense network with batch normalization and LeakyReLU activation, SKINC-NET reached an accuracy of 98.54% which is better than 98.23% of traditional heavyweights such as ResNet50 and EfficientNetB0 with only a fraction of their number of parameters. Similarly, Khan's LEAD-CNN framework for brain-tumor classification includes dimension-reduction blocks to optimally minimize computational overhead without any decrease in accuracy. 98.70% accuracy was achieved with almost the same precision and F1-scores. Both approaches are examples of the trend towards lightweight medical AI systems that consider accuracy and efficiency and help to form a benchmark for clinical implementation in real time.

Further building on this lightweight paradigm in the environmental health field, Hossen and Mondal proposed a lightweight CNN for arsenicosis detection using the ArsenicSkinImageBD dataset with an accuracy of 98.75% and on the upper hand of more complex architectures, such as ViT and EfficientNetV2B0. All three models show that with structural optimization and augmentation techniques, lightweight networks can be on par with or exceed conventional deep networks while also being deployment-friendly. In parallel, models with architectural diversity and attention mechanisms have been advanced in efficiency-driven models by Khasanov et al. [42] and Elsheikhy et al. [43]. Khasanov et al. suggested the DeepBiteNet ensemble based on DenseNet121, EfficientNetB0, and MobileNetV3-Small for the classification of insect bites, achieving 84.6% accuracy and better robustness using domain-specific augmentations to consider the factors of lighting, occlusion, and skin tone variations. Meanwhile, Elsheikhy et al. worked on the detection of cardiac arrhythmia using a channel attention-enhanced CNN, which achieved an accuracy of 99.48% on the INCART dataset without complexity. Both frameworks highlight feature reuse in the form of feature reweighting, selective focus and modular integration as possible and potentially efficient mechanisms without sacrificing the performance. Collectively, these five studies demonstrate a steady progression in the design of lightweight deep learning approaches from small CNNs and parameter reduction to attention-guided and ensemble designs that create the basis for designating scalable diagnostic systems to perform real-time analysis for precision, interpretability and accessibility in dermatology, neurological studies, cardiovascular studies and environmental studies.
Recent advances in deep learning for medical imaging have addressed feature representation improvement using attention mechanisms and multi-scale processing. These techniques can make model representations emphasize some regions of interest selectively and capture variations in the lesion at multiple spatial resolutions. Reconstructing channel-wise recalibration and hierarchical feature fusion is known to significantly improve the classification accuracy and interpretability of dermatological image analysis. The fusion of the attention mechanism and multi-scale feature extraction has been a revolutionary breakthrough in lightweight deep learning models for medical imaging. Almars et al. [44], Nayeem et al. [45], Liu et al. [46] and Abir et al. [47] all pushed the efficiency boundary utilizing various attention augmented and hybrid architectures, focusing on accuracy, interpretability and computational optimization. Almars and Nayeem targeted the problem of improving diagnostic efficiency using attention-guided CNN and fusion-based CNN frameworks. Almars's DeepGenMon, optimized using a genetic algorithm (GA), selectively highlighted disease-related image regions to achieve multi-class skin disease classification for skin problems such as monkeypox and melasma, achieving a 0.985 F1-score with low computational cost.
In another direction, Nayeem derived a multi-scale attention fusion (MAF)-based depthwise separable convolution DCGAN-based augmentation deep residual neural network block in MAF-DermNet, obtaining over 99.9% accuracy, proving the benefits of attention and multi-scale remembering in scale at the same time and keeping models light. Meanwhile, Liu and Abir pushed attention-driven architectures towards their aspect of adaptability to the context and hybrid representation. Liu's multi-scale channel attention model based on improved pyramid segmentation attention and reverse residual blocks was able to capture the hierarchical lesion context in ISIC2019 and HAM10000 with accuracies of 77.6% and 88.2%, respectively,

which had a balanced efficiency and space for further contextual improvement. The second paper presented by Abir is Attention Enhanced InceptionV3 (AEIV3): This integrated Soft Attention, Channel Attention and Squeeze-and-Excitation modules on a restructured InceptionV3 backbone to achieve 99.58% accuracy on the PAD UFES 20 dataset. This design was effective in providing a good balance between shallow and deep features to reduce computational complexity while providing discriminative power at the same time. All these works can be conceptualized where the attention mechanisms, multi-scale fusion and optimization-driven tuning

Table 2.4: Summary of lightweight and attention-enhanced deep learning architectures in medical and environmental diagnostics

Reference	**Model**	**Results**	**Limitations**
[39]	SKINC-NET (Dense CNN + BatchNorm + LeakyReLU)	98.54% accuracy on HAM10000; outperformed ResNet50 (98.23%) and EfficientNetB0 with far fewer parameters.	Dataset-specific; limited external validation; may underfit complex lesions.
[40]	LEAD-CNN (Lightweight CNN with Dimensionality Reduction Blocks)	98.70% accuracy with balanced precision and F1; reduced computational overhead.	Evaluated on brain tumor dataset only; lacks multi-modal validation.
[41]	Lightweight CNN for Arsenicosis Detection (ArsenicSkinImageBD)	98.75% accuracy; surpassed ViT and EfficientNetV2B0 with simpler architecture.	Region-specific dataset; limited generalization to other environments.
[42]	DeepBiteNet (DenseNet121 + EfficientNetB0 + MobileNetV3-Small Ensemble)	84.6% accuracy; improved robustness against lighting, occlusion and skin tone variations.	Lower accuracy compared to single-domain models; ensemble increases model size.
[43]	Channel Attention-Enhanced CNN (Cardiac Arrhythmia Detection)	99.48% accuracy on INCART dataset; achieved high precision with minimal complexity.	Dataset-limited; lacks broader cardiovascular dataset testing.
[44]	DeepGenMon (GA-Optimized Attention CNN)	F1 = 0.985 for monkeypox and melasma detection; high interpretability with low computational cost.	Limited dataset diversity; lacks cross-domain validation.
[45]	MAF-DermNet (Multi-Scale Attention Fusion + Depthwise Separable CNN + DCGAN Augmentation)	>99.9% accuracy; excellent efficiency and robustness to scale variation.	Heavy augmentation reliance; requires fine-tuning for smaller datasets.
[46]	Multi-Scale Channel Attention Model (Pyramid Segmentation + Reverse Residuals)	77.6% (ISIC2019) and 88.2% (HAM10000); balanced efficiency with contextual feature capture.	Moderate accuracy; limited contextual generalization.
[47]	AEIV3 (Attention Enhanced InceptionV3: Soft + Channel + SE modules)	99.58% accuracy on PAD UFES 20 dataset; strong shallow–deep feature balance.	Dataset-specific; may need validation on other lesion types.

combine to make a concerted effort to improve diagnostic performance under limited computational budgets. Collectively, these studies demonstrate the advancement in developing lightweight attention-based models that procure an efficiency and interpretability balance, which serves as a breakthrough towards clinically deployable and real-time dermatological data AI systems. These developments are presented systematically in Table 2.4 and represent a good comparison between lightweight and attention-based diagnostic frameworks.

2.4 Knowledge Distillation in Medical Imaging

The knowledge distillation (KD) demonstrated in the work is a model compression approach, which aims to distill the knowledge from a large teacher network with high capacity to a small and efficient model called student. Instead of pairing hard PoK learning, the student model learns from the softened probability distributions released by teacher by replicating the finer-grained inter-class dependencies and semantic relation richness in the data. This transfer mechanism provides the student with an opportunity to stay close to the performance of the teacher but with a fraction of the computational burden. Therefore, KD provides a tradeoff between accuracy and efficiency and is of great interest for information devices on the interface with diagnostic applications that require real-time computations with limited computational power. In the context of medical imaging, KD is a gap between deep model complexity and practical deployable scalability to support the development of AI models that are both high-performing and practical in clinical settings.

Over the past few decades, knowledge distillation has become a promising method for medical image analysis, making it significantly more efficient while maintaining diagnostic accuracy. Innovative analytics and machine learning have been successfully employed in several healthcare disciplines, such as dermatology, radiology, pulmonology and orthopedics. Several recent studies have pointed to its potential in achieving a good trade-off between computational cost, interpretability and performance. In dermatological imaging, Saha et al. [48] and Kabir et al. [49] proved the efficacy of KD-based networks with the HAM10000 and ISIC2019 datasets, in which distilled student models achieved more than 90% accuracy, getting very close to their teachers while significantly decreasing the number of parameters. Their findings highlight the importance of KD in facilitating low-resource clinical implementation without reducing diagnostic confidence. Similarly, Islam et al. [50] contributed further by adopting ResNet152V2, ConvNeXtBase and ViT-Base to create a hybrid teacher model. The student model, after quantization to less than 0.5 MB, attained an accuracy of 98.94%, which established a new data benchmark for fast lightweight deployment on mobile and embedded medical devices.

Taken together, these dermatology-centered studies clearly demonstrate that KD is both applicable and feasible for diagnostic rigor while allowing for portability and scalability. Outside dermatology, Asham et al. [51] utilized KD in combination with ConvMixer and MobileNet blocks for pulmonary disease classification on chest X-ray data, achieving 97.92% accuracy and only 0.63 million parameters, which proves the effectiveness of KD for multi-class medical imaging in limited environments. To address the inherent multi-modality problems, Shi et al. [52]proposed a cross-modal KD framework (CMKD-BMD) for bone mineral density prediction by transferring knowledge from multi-modal to single-modality students (X-ray and CT). This not only improved the prediction accuracy but also demonstrated cross-modality knowledge transfer through KD and continued the evolution of KD in terms of clinical applicability. Finally, to further verify the efficiency added by KD through the student-teacher compression model, Thaker and Asif [53] strengthened theecia by increasing ROC-AUC from 0.65 to 0.85, which attests as another strong point for increasing sensitivity and accessibility for skin cancer detection using AI in low-resource settings. Collectively, these studies position knowledge distillation to be an important cornerstone of efficient medical AI that is capable of compressing complex architectures, while retaining interpretability, precision and clinical relevance. Although its application has been very successful, there are still issues to be addressed, such as its dependency on datasets and lack of domain adaptability and generalization across imaging modalities. Working on these limitations, this study utilized KD for skin lesion detection of arsenic, where we proposed an optimized distillation strategy with preservation of accuracy while maximizing the reduction of computation, thereby stepping further into clinical deployable, online diagnostic intelligence in dermatology. These understandings are systematically summarized in Table 2.5 to provide a concise review of the architectures of teachers and students in the medical and dermatological fields for providing diagnoses.

Table 2.5: Summary of knowledge distillation-based architectures in medical and dermatological diagnostics

Reference	**Teacher Type**	**Model**	**Results**	**Limitations**
[48]	Single-teacher KD	KD-based CNN (Teacher–Student) trained on HAM10000	>90% accuracy; student model nearly matched teacher while using far fewer parameters.	Dataset-limited; lacks domain generalization beyond dermatology.
[49]	Single-teacher KD	KD-based CNN on ISIC2019	Achieved >90% accuracy; maintained diagnostic reliability with significant model compression.	Validation restricted to single dataset; limited interpretability analysis.
[50]	Single-teacher KD	Hybrid Teacher (ResNet152V2 + ConvNeXtBase + ViT-Base) $\rightarrow$ Quantized Student	Student (<0.5 MB) achieved 98.94% accuracy; highly efficient for mobile and embedded deployment.	Model performance reliant on teacher configuration; limited external validation.
[51]	Single-teacher KD	KD with ConvMixer + MobileNet for chest X-ray	97.92% accuracy; 0.63M parameters; effective multi-class pulmonary disease classifier.	Domain-limited; small-scale dataset for chest X-ray only.
[52]	Multi-teacher KD	Cross-Modal KD (CMKD-BMD) for Bone Mineral Density prediction	Improved accuracy by transferring knowledge from multi-modal to single-modality (X-ray, CT).	Limited generalization across new modalities; complex training pipeline.
[53]	Single-teacher KD	Student–Teacher KD for Skin Cancer Detection	Improved ROC-AUC from 0.65 $\rightarrow$ 0.85; enhanced sensitivity in low-resource settings.	Dataset-specific; requires tuning for diverse skin cancer data.

2.5 Explainable Artificial Intelligence (XAI) in Medical Imaging

Despite the outstanding predictive performance of deep learning models, these models can be considered black-box systems because there is usually no explanation regarding how these models arrive at their predictions. This lack of transparency is of particular concern in medical imaging, where, in the absence of transparency in decision-making, there are severe ethical, clinical, and accountability issues. Doctors must not only be aware of what the model predicts but also why it predicts it; therefore, the focus of artificial intelligence is based on biologically and pathologically meaningful areas. Explainable Artificial Intelligence (XAI) alleviates this limitation with visual and conceptual interpretability, which allows clinicians to assess the model to ensure that it has chosen a decision that matches the actual disease pattern. The cross-ways of post-hoc visualization and feature attribution techniques are explored to enable XAI to bridge the computational inference of medical reasoning, resulting in trust, transparency, and human-AI decision-making in collaboration. For example, in dermatological diagnostics, where the morphology of lesions, their color gradient and irregularities in texture play important roles in the examination, the use of XAI makes deep networks interpretable diagnostic aids that improve both model accountability and clinical uptake.

Post-hoc XAI identifies reasons for the model after it has been used in the prediction and therefore provides important interpretability without modifying model architecture. Among these, Grad-CAM++ and LIME are popularly used for analyzing dermatological images. Grad-CAM++ computes higher order derivatives of class activation maps to generate fine-grained heatmaps, which points out areas of lesions most politically, while LIME perturbs input regions to resemble a local decision boundary, which identifies the most influential pixels or textures to a classification. Together, these techniques are complementary with respect to interpretability: Grad-CAM++ is used to understand where the model is focusing its attention, while LIME is used to explain why certain features contribute to a prediction.

Together, these two are needed to provide the required complete transparency in the case of high-stakes medical diagnostics. These XAI methods have recently been proven to increase clinical reliability and trust in various medical and dermatological applications. Gamage et al. [54] applies U2-Net lesion masks to form lesion focus inputs and train using mask-guided CNNs + a proposed network, SM-ViT, on a balanced & duplicate removed HAM1000 Melanoma vs Nevus subset: Mask increases performance with Modified Xception achieving 98.37% and SM-ViT 92.79%. Grad-CAM/Grad-CAM++ explanations were validated in comparison to ISIC 2018 attribute masks, where the heatmaps of Grad-CAM++ and SM-ViT were most similar to the actual lesion regions. Nigar et al. [55], Vaghela et al. [56] and Hong and Lin [57] all used the LIME-based interpretability in visualizing CNN decision processes in tasks of classifying skin lesions. Nigar et al. [55] trained their model on ISIC 2019 and achieved 94.47% accuracy and the visual explanations were very consistent with dermatological judgement. Vaghela et al. [56] generalize this idea globally, across multiple models (ResNet50V2, InceptionV3, VGG16, InceptionResNetV2), demonstrating how LIME improved the confidence of the practitioners in identifying areas for clinical attention. Similarly, Hong and Lin used an explainable framework based on the VGG16 network and checked lesion-specific feature extraction, revealing background-driven misclassification to reflect the value of XAI in continuous model improvement. Parallelly, Raju et al. [58], Fiaz et al. [59] and Saha et al. (Mpox-XDE) [60] implemented Grad-CAM visualization for the localization of lesions. Raju's XAI-SkinCADx, a hybrid model that combines CNN, BiLSTM and SVM, gave an accuracy of 95.63% and showed better interpretability, IoU =0.78, Dice = 0.87, which is better than Grad-CAM++ baselines. Likewise,

Fiaz et al. introduced Grad-CAM in a dual-task U-Net + EfficientNetB0 architecture, which could guarantee transparency at the lesion level with a Dice coefficient of over 0.85, while Mpox-XDE ensemble from Saha et al. showed an accuracy of 98.70%, demonstrating the model's attention to diagnostic relevant regions. Altogether, these studies highlight the importance of attention visualization in promoting explainability and clinical trust in highly accurate dermatological systems.
Moreover, recent frameworks have combined XAI and Federated Learning (FL) to ensure privacy. This trend was represented by Biswas et al. [61] and Gupta, Kumar and Dhir [62] through their techniques FedMed-XAI and FL-Integrated dermatological system in which LIME can help a clinician understand the decentralized model outputs without using patient data. With results in accuracy improvement as high as 94 percent - under federated training conditions, the researchers claim that XAI is key for aiding patient and transparent deployment of AI to meet ethical standards. Similarly, Alrabai, Echtioui and Kallel [63] validated several pre-trained CNNs (Xception, ResNet50V2, DenseNet121) with 90% accuracy, where XAI enabled clinician validation and decreased the ambiguity of diagnosis. Together, these works finally confirm that XAI is not an additional add-on to clinical AI but a fundamental part of clinically responsible AI. Through integration with federated learning engines, post-hoc visual reasoning, and hybrid ensemble architectures, the modernization of studies in dermatology has led to the implementation of accurate, transparent, and ethically deployable AI systems. Thus, by combining the power of Grad-CAM++ with LIME, we can bridge the gap between computational accuracy and medical interpretability and translate the value proposition of deep learning from the black box of a classifier into the role of an explainable and clinician-centered clinical partner. These advancements in post-hoc interpretability are systematically organized in Table 2.6, which provides a good comparison of XAI methods across dermatological and medical diagnostics.

Table 2.6: Summary of post-hoc explainable AI (XAI) frameworks in dermatological and medical diagnostics

Reference	Model	Results	Limitations
[54]	Mask-guided CNNs + proposed SM-ViT; U2-Net lesion segmentation; Grad-CAM/Grad-CAM++ explanations	98.37% accuracy (best: modified Xception) on balanced HAM10000 melanoma vs. nevus; SM-ViT improves ViT to 92.79%; U2-Net segmentation IoU $\approx$ 0.93; Grad-CAM++ gives clearer lesion localization.	Needs larger and more diverse datasets, validation across different populations/skin types and extension from binary to multi-class classification.
[55]	CNN + LIME (ISIC 2019)	94.47% accuracy; visual explanations aligned with dermatologist assessments.	Limited to a single dataset; lacks quantitative XAI evaluation metrics.
[56]	LIME on multiple CNNs (ResNet50V2, InceptionV3, VGG16, InceptionResNetV2)	Enhanced practitioner confidence and clinical interpretability across architectures.	Qualitative validation only; limited generalization testing.
[57]	Explainable VGG16-based CNN	Identified lesion-specific and background-driven misclassifications; supported iterative model refinement.	Model dependent; no numerical interpretability benchmarks.
[58]	XAI-SkinCADx (CNN + BiLSTM + SVM) with Grad-CAM	95.63% accuracy; IoU = 0.78, Dice = 0.87; detailed lesion-level Grad-CAM visualizations.	Requires multiple architectures; computationally complex.
[59]	Dual-task U-Net + EfficientNetB0 + Grad-CAM	Dice > 0.85; ensured lesion-level transparency and spatial explainability.	Limited to binary lesion segmentation; no large-scale clinical test.
[60]	Ensemble model with Grad-CAM visualization	98.70% accuracy; clear localization of diagnostically relevant regions.	Dataset-specific; limited scalability.
[61]	FedMed-XAI (federated learning + LIME integration)	$\sim$ 94% accuracy under federated training; preserved data privacy and interpretability.	High communication cost in FL; requires secure aggregation.
[62]	FL-integrated dermatological XAI system	Improved accuracy and clinician trust without direct data sharing.	Model limited to specific dermatological use cases.
[63]	Pre-trained CNNs (Xception, ResNet50V2, DenseNet121) + XAI	$\sim$ 90% accuracy; XAI reduced diagnostic ambiguity and increased clinical validation.	Limited scope of disease classes; interpretability assessed qualitatively only.

2.6 Research Gap

The reviewed literature on arsenicosis and broader dermatological imaging provides an overview of a steady stream of developments from classical machine learning with handcrafted descriptors (GLCM, LBP, color histograms) to deep, hybrid, lightweight, and knowledge-distilled networks that emphasize accuracy, portability and clinical usability. However, each stage of this progression exhibits important limitations. Early pipelines were able to obtain interpretable, competitive results but were limited in use by the fact that the datasets used were often small, not balanced, or were specific to one geographical location, thus limiting generalization beyond the original cohorts. Later CNNs and transfer learning models have focused on EfficientNet-variants, ensembles and transformer augmentation pipelines that improved the accuracy and robustness to texture variation; however, many of them were still bound to the use of localized data and had huge computational costs, preventing even the demonstration of real-time or even on-device inference for low-resource settings. Lightweight architectures helped to satisfy some of the efficiency concerns, but improvements were sometimes based on heavy augmentation or multi-stage fusion that might turn brittle in changing domains of lighting, occlusion, skin tone and even the ensemble designs might be reintroducing complexity. Knowledge distillation proved that small models have the potential to achieve near-teacher performance, even at sub-megabyte model sizes. However, most existing studies rely on single-teacher schemes, focus on a narrow set of diseases, or use only a single dataset, meaning that multi-teacher distillation, domain adaptation, and cross-modality robustness remain underexplored. Regarding interpretability, for the post-hoc XAI techniques, Grad-CAM/Grad-CAM++ and LIME consistently increased clinician trust and illustrated attention patterns; however, the results of the analysis were often qualitative with few shared reports: how are explanations tracking decision fidelity? Compounding these problems, many reports either did not include rigorous stratified k-fold cross-validation that can lower claims to stability when working with imbalanced types of medical datasets, or they contributed to the heterogeneity of some reported metrics in different studies. In summary, the major gaps are: the lack of comprehensive datasets covering domain-specific learning (often lacking subclass labels), heavyweight models that hinder scalable and on-device deployment, lack of exploration of multi-teacher KD for skin disease detection, lack of quantitative interpretability to validate the model rationale, and the use of robust cross-validation for generalizability is rarely used. To address these gaps, in this study, we present an explainable dual-teacher knowledge-distilled lightweight framework that distills complementary InceptionV3 and Xception+InceptionM based teacher models into an efficient custom student, IR-KAN. The framework was trained and evaluated under two complementary protocols that addressed class imbalance and domain variability, and integrated Grad-CAM++ and LIME to simultaneously provide deployment-ready efficiency and clinician-centered transparency for arsenic skin lesion screening.

Chapter 3

Proposed Methodology

In this study, an explainable deep learning framework has been designed to offer both accurate and explainable classification of arsenic-induced skin disease as shown in the proposed workflow in Figure 3.1. The framework incorporates three

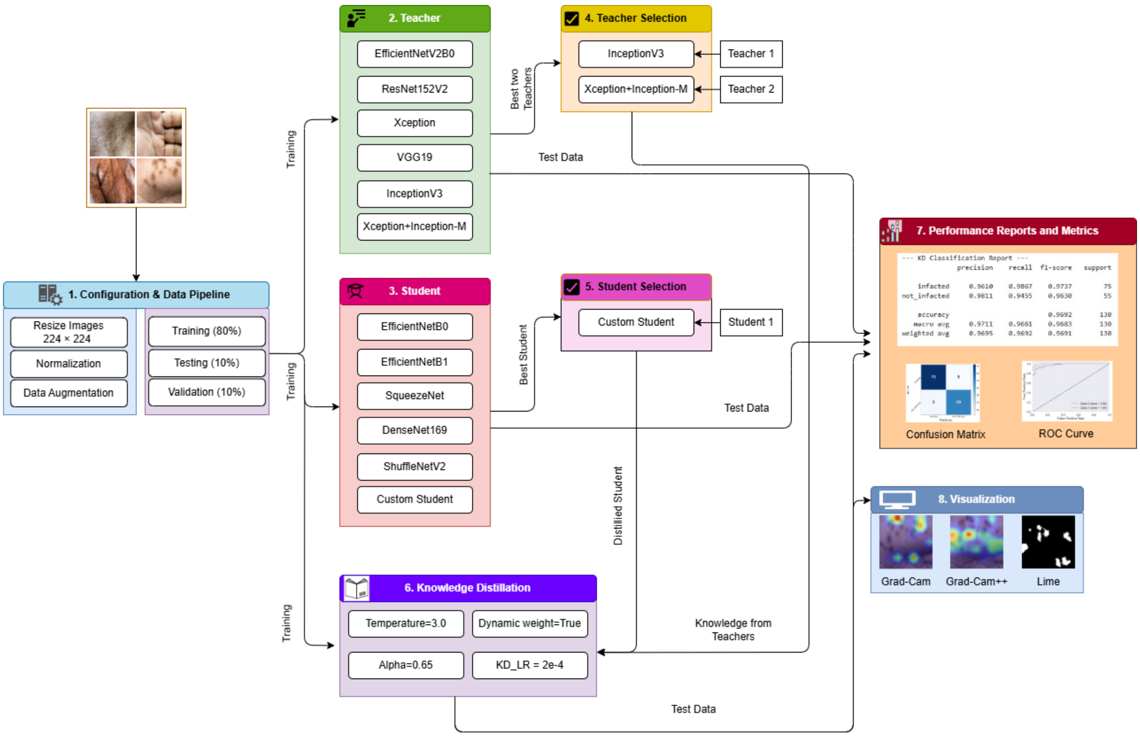


Figure 3.1: Proposed workflow of this work. Starting with data load, preprocessing, data split, teacher and student training, knowledge transfer from selected teacher to lightweight student and ending with model evaluation and visualization.

major elements: a dual-teacher knowledge distillation (KD) process facilitating effective passage of multi-scale information of high-capacity teacher models to a compact and small student network; a student architecture facilitating efficient inference; and explainable artificial intelligence (XAI) modules that enhance transparency and clinical interpretability. End-to-end workflow results in five consecutive stages: (1) The preprocessing and stratified dataset division of the data to provide high-quality and balanced inputs; (2) The choice and training of teachers to learn strong high-level feature representations; (3) student model design and optimization to get competitive performance using a significantly lower number of computation resources; (4) knowledge distillation dual teacher-student knowledge to learn to distribute to the student complementary knowledge between two teachers; and (5) evaluation and explainability visualization to judge the reliability of the model and to understand how the student makes its This combined pipeline aims to strike a balance between diagnostic accuracy, computational efficiency and interpretability which can be considered as three paramount groups of factors in making AI-assisted clinical decision-support system trustworthy.

3.1 Dataset Description

The “ArsenicSkinImageBD" dataset is used in this research, which is public and is accessible through the Mendeley Data Repository [14]. It is the initial open-source image database intended to be used specifically to detect the arsenic-related skin disorders in the human body.
The images were gathered from four villages with arsenic-contaminated soil in the Chapainawabganj district, Bangladesh, that is, in Betbaria (Ward 08, Union: Balidanga), Balubagan (Ward 08, Union: Maharajpur), Dole Para (Ward 09, Union: Maharajpur) and Ramchandrapur Hat (Ward 08, Union: Dohilpara, Sadar, Chapainawabganj). The selection of these areas was based on the fact that they have high levels of contamination of arsenic groundwater and that arsenicosis cases are clinically verified. Image capture was carried out in the natural lighting setting with four smartphone cameras of varying specifications to recreate the variability of clinical real-world imaging in rural and resource constrained settings. Such variability brings in diversity of illumination, focus and color tone thus enhancing the ecological soundness and strength of the dataset in the generalization of deep learning models. Field experts who were knowledgeable about arsenic induced dermatological phenotype manually checked and labeled each image based on their observations of melanosis (hyperpigmented spots), leucomelanosis (either hypopigmented or mixed patches on the skin) and keratosis (rough, thickened areas of skin). Class annotations were verified by experts to bring out the true clinical condition.
There are 1,287 images in the ArsenicSkinImageBD dataset in two major categories:

- **Infected Class:** Images of apparent manifestations of arsenicosis as a chronic skin disease are covered by this category as a result of long-term exposure to arsenic in the form of groundwater. A total of 741 images that show examples in Figure 3.2 of visible dermatological lesions of arsenicosis (melanosis) (or hyperpigmented), leucomelanosis (or hypopigmented or mixed pigmentation) and keratosis (or thick and roughened skin). This variation in the types of lesions included in this category enriches intra-class variation so that the models are able to learn discriminative details at different levels of severity and skin texture.

- **Not-Infected Class:** A total of 546 images of non-infected skin parts greatly helped in the binary classification as negative samples in Figure 3.3. These types include photos of skin where no arsenic lesions are clearly visible. Nevertheless, to enhance generalization capability of classification models, this

(a) (b)

Figure 3.2: Sample of infected classes

classification is equally encompassed as true healthy skin samples of people who do not have a history of exposure to arsenic. Non arsenic visual similar dermatoses including benign pigmentation or mild eczema which encourage the network to avoid overfitting and need to acquire meaningful differences in the lesions instead of shallow colour change. This binary classification architecture promotes supervised learning activities in the context of differentiating arsenic-related pathologies with visually comparable but not toxic clinical and technically practical goals.

(a) (b)

Figure 3.3: Sample of not-infected classes

3.2 Data Splitting

This study employed two complementary data-splitting strategies: a fixed 80/10/10 split and a 5-fold cross-validation protocol to ensure a fair and reliable evaluation of the proposed model. The fixed split was used to provide a fixed baseline on which the model would be trained, tested and validated and cross-validation used to provide a more demanding examination of generalization by trying the model on more than one data partition. A combination of the two methods enabled us to ascertain the stability, robustness and reproducibility of the model on ArsenicSkinImageBD data.

3.2.1 Fixed 80/10/10 Split

The 80:10:10 split was finally established after several data partitioning schemes were considered and it yielded better and more balanced results across experimentations. This setup constituted 80% of samples to be used in training the model, 10% in validation during optimization and the remaining 10% as final testing to ensure that it was not biased in measuring performance.

3.2.2 *k*-Fold Cross-Validation

Cross-validation is also a common method for evaluating a model that helps to ensure the robustness, reliability and generalization of a predictive model. Instead of using a single train-test split, it recursively splits the dataset into several different folds, whereas the model can be trained and validated using a set of various data subsets. This process helps to reduce the risk of overfitting and helps to estimate the performance in a more complete way in the real world. In case of medical diagnosis, like in skin lesion detection for arsenic, the concept of cross-validation becomes absolutely important in order to verify that the model's prediction is stable and consistent across different distributions of patient data. Hence, in this study, a 5-fold cross-validation protocol is adopted to give a statistically sound and unbiased performance evaluation of the proposed knowledge-distilled student model. In this method, a complete data set D is divided into 5 mutually exclusive subsets, which are approximately equal to each other D_1, D_2, D_3, D_4, D_5. In each iteration, one of the subsets is selected as the validation set, while the other 4 subsets are selected as the training set. This procedure will be repeated five times so that each occurrence will appear exactly once on the validation set. The performance obtained is then the average of the individual fold performances. When training the model over several iterations, four folds (80%) were used and the remaining one fold (20%) as the validation set. This has been done five times where each of the folds served as a validation subset. The model performance measures, such as accuracy, precision, recall and F1-score, were averaged to obtain the final cross-validation outcome after all the five iterations.
The results of evaluation metrics per fold (M_k) were summed to obtain an estimate of the mean and standard deviation, which gives a perspective on the central tendency and the variability of the model execution in Equation 3.1:

$$\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M_k, \qquad \sigma_M = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(M_k - \bar{M})^2} \tag{3.1}$$

where K= 5 is the total folds.
Such a formulation does not only measure the average performance of the model, but also shows its stability with various data subsets.

3.3 Data Preprocessing and Augmentation

Deep learning systems are highly sensitive to changes in the input data distribution; therefore, carefully designed preprocessing and augmentation procedures were crucial to maintain model stability, improve generalization and reduce overfitting. All image data in this study were subjected to a uniform preprocessing pipeline, after which class-balanced data augmentation was applied on the fly to the training set during model training only. Before inputting the images to the network, a series of preprocessing functions were implemented in order to normalize the size of the images, equalize the intensity of the pixels and to provide numerical stability to the training.

Resizing

All images in the ArsenicSkinImageBD dataset were resized to a spatial resolution of 224×224 pixels. The input size of popular convolutional backbones is compatible with this resolution and still allows capturing fine details of the dermatological texture (e.g. pigmentation or keratotic topography).

Normalization

The colours of all the pixels were scaled down to the range [0, 1] by a transformation in Equation 3.2:

$$I_{\text{norm}} = \frac{I_{\text{raw}}}{255} \tag{3.2}$$

where I_{raw} represents the original 8-bit value.
This normalization enhances gradient stability and gradient convergence speed due to the fact that using the normalization, the input features are brought to the same scale, which is numerically normalized.

Rotation

A random rotation of a $\pm 10°$ was used to compensate for the possibility of variable camera angles when capturing an image. This change makes the model invariant to rotations without causing any severe distortion to the geometry of lesions.
The rotation transformation is mathematically defined as in Equation 3.3:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{3.3}$$

Where θ is the random rotation angle that is sampled in $[-10°, +10°]$.
This slight angular transform causes the model to identify lesions not dependent on camera orientation, which is a typical problem of field-acquired images.

Translation

Horizontal and vertical translations were also performed with a maximum offset of 10% of the image dimension in order to simulate lesion position variation within the frame. The transformation enhances the translation invariance of the model and makes it bi-uniformly detect lesions even in cases where the lesion is not centrally located.
In the official form, translation transformation $T_{\Delta x, \Delta y}$ (of a picture $I(x, y)$ can be as follows in Equation 3.4:

$$I'(x, y) = I(x - \Delta x,\ y - \Delta y) \tag{3.4}$$

In which x and y denote the horizontal and vertical pixel coordinates of the image grid, respectively and Δx and Δy are horizontal and vertical shifts respectively.

Shear Transformation

Shear factor of 0.05 was used to give a simulation of small perspective distortions which occur in the hand held image capture. This operation does a linear mapping operation of pixel coordinates which is defined as in Equation 3.5:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{3.5}$$

in which s refers to the shear intensity. The transformation changes the shape of the image slightly without loss of structural coherence hence increasing the ability of the model to tolerate small geometric distortions.

Flipping

- **Horizontal flipping:** Horizontal flipping was introduced in order to add symmetry left-right mirror to allow the human skin area to be highly varied when one is compared against the opposite part of the body (e.g., left arm vs. right arm). The flip is presented below in Equation 3.6:

  $$I'(x, y) = I(W - 1 - x,\ y) \tag{3.6}$$

  Where W is the image width.

- **Vertical flipping:** Vertical flipping was turned off as this would give anatomically impossible orientations, in particular palms, feet and trunk parts. Consequently, the model will be exposed to novel visual circumstances on a semi-regular basis, thus minimizing overfitting on particular lesion forms or light gradients.

With this well-conceived preprocessing and augmentation protocol in Table 3.1, the ArsenicSkinImageBD dataset grew to include a wide variety of data with restricted fields and remained to a large, diverse and clinically faithful dataset. This rich data greatly enhanced the process of generalization of the model, forming a foundation to the subsequent application, the lightweight distribution network, which is trained subsequently in this research.

Table 3.1: Data augmentation transformations and their parameter configurations.

Transformation Type	**Parameter Range / Value**	**Purpose**
Rotation	$\pm 10°$	Orientation invariance
Width Shift	± 0.10	Translation robustness
Height Shift	± 0.10	Spatial invariance
Shear	0.05	Perspective tolerance
Horizontal Flip	True	Left-right symmetry
Vertical Flip	False	Maintain anatomical realism
Rescale	$\frac{1}{255}$	Normalize pixel intensity

3.4 Teacher Models

In the case of the teacher models, we chose a few recent state-of-the-art deep learning architectures that have demonstrated high rating in classifying medical images particularly on small and limited set of data. InceptionV3, VGG19, ResNet152V2, EfficientNetV2B0 and Xception were compared in terms of their robustness and ability to discriminate. A hybrid Xception+InceptionM model was further incorporated, which applies a 51.5M trainable parameter Xception backbone, but with multi-scale Inception added on top, leading to just 0.81M trainable parameters. The combination of these teacher networks yields good hierarchical, multi-scale and fine-grained feature representations that can be used to direct the lightweight student model by knowledge distillation.

3.4.1 Architecture of Hybrid Xeption+InceptionM Teacher Model

The architecture of the teacher network came after the suggestion of the ArsenicNet model by Mehedi et al. (2025) [6] that combines both the Xception backbone and the Inception module to further improve the multi-scale feature learning. The Xception element uses depthwise separable convolution which disintegrates standard convolutions into depthwise and pointwise tasks. This architecture allows extracting fines-grained spatial representations using much less number of parameters. To supplement this an Inception module was added which is comprised of parallel convolutional branches ($1 \times 1, 3 \times 3, 5 \times 5$) and a max-pooling branch, all preceded by dimensionality-reducing 1×1 convolutions. This design enables the model to operate on image attributes on various receptive fields both grand textual priuses and large-scale contextual knowledge.
The model works mathematically in the following way:

$$F_X = \text{Xception}(I), \qquad F_X \in \mathbb{R}^{H' \times W' \times C'} \tag{3.7}$$

where, H', W' and C' are respectively the height, width and number of channels of the extracted feature maps. Equation 3.7 is the backbone of the Xception model. In Equation 3.8- Equation 3.11 multi-scale features is computed by the Inception module through multiple branches with different kernel and filters sizes

$$F_{1\times1} = K_{1\times1} * F_X \tag{3.8}$$

where,

- $F_{1\times1}$: From the 1×1 convolution branch the output feature maps.
- $K_{1\times1}$: Represents 1×1 convolution kernel's weights.
- F_X: Represents the Xception backbone's input feature mapping.

$$F_{3\times3} = K_{3\times3} * \Big(K_{\text{reduce3}} * F_X\Big) \tag{3.9}$$

where,

- $F_{3\times3}$: From the 3×3 convolution branch the output feature maps.
- $K_{reduce3}$: Represents 1×1 convolution kernel's weights for dimensionality reduction.
- $K_{3\times3}$: Represents 3×3 convolution kernel's weights.
- F_X: Represents the Xception backbone's input feature mapping.

$$F_{5\times5} = K_{5\times5} * \Big(K_{\text{reduce5}} * F_X\Big) \tag{3.10}$$

where,

- $F_{5\times5}$: From the 5×5 convolution branch the output feature maps.
- $K_{reduce5}$: Represents 1×1 convolution kernel's weights for dimensionality reduction.
- $K_{5\times5}$: Represents 5×5 convolution kernel's weights.
- F_X: Represents the Xception backbone's input feature mapping.

$$F_{\text{pool}} = \text{MaxPool}(F_X), \qquad F_{\text{pool}\,1\times1} = K_{\text{pool}\,1\times1} * F_{\text{pool}} \tag{3.11}$$

- F_{pool}: The max pooling operation produces the feature maps.
- F_X: From the Xception backbone the input feature maps.
- $F_{pool\,1\times1}$: Represent the 1×1 convolution applied of output feature maps to pooled feature maps.
- $K_{pool\,1\times1}$: Represents 1×1 convolution kernel's weights.
- F_{pool}: The pooling operation produces the input feature maps.

Lastly, all the branches are added together on their output in Equation 3.12:

$$F_{\text{Inc}} = \text{Concat}(F_{1\times1},\, F_{3\times3},\, F_{5\times5},\, F_{\text{pool}\,1\times1}) \tag{3.12}$$

The output of the hybrid model worldwide can be calculated as in Equation 3.13:

$$Y_{\text{pred}} = \sigma\Big(W_{\text{Out}} \cdot \text{ReLU}(W_{\text{Dense}} \cdot \text{GlobalAvgPool}(\text{Inc}(\text{Xception}(I))) + b_{\text{Dense}}) + b_{\text{Out}}\Big) \tag{3.13}$$

where,

- $\mathbf{Y}_{\text{pred}}$: The final prediction output generated by the proposed model.
- $\mathbf{I}$: The input image with spatial dimensions of $224 \times 224 \times 3$.
- $\mathbf{Xception(I)}$: The feature maps extracted from the input image using the pre-trained Xception backbone.
- $\mathbf{Inc}$: The Inception module applied to the extracted feature maps.
- $\mathbf{GlobalAvgPool}$: The global average pooling layer used to aggregate spatial features into a compact representation.
- $\mathbf{W}_{\text{Dense}}, \mathbf{b}_{\text{Dense}}$: The weights and biases of the dense layer.
- $\mathbf{W}_{\text{Out}}, \mathbf{b}_{\text{Out}}$: The weights and biases of the final output layer.

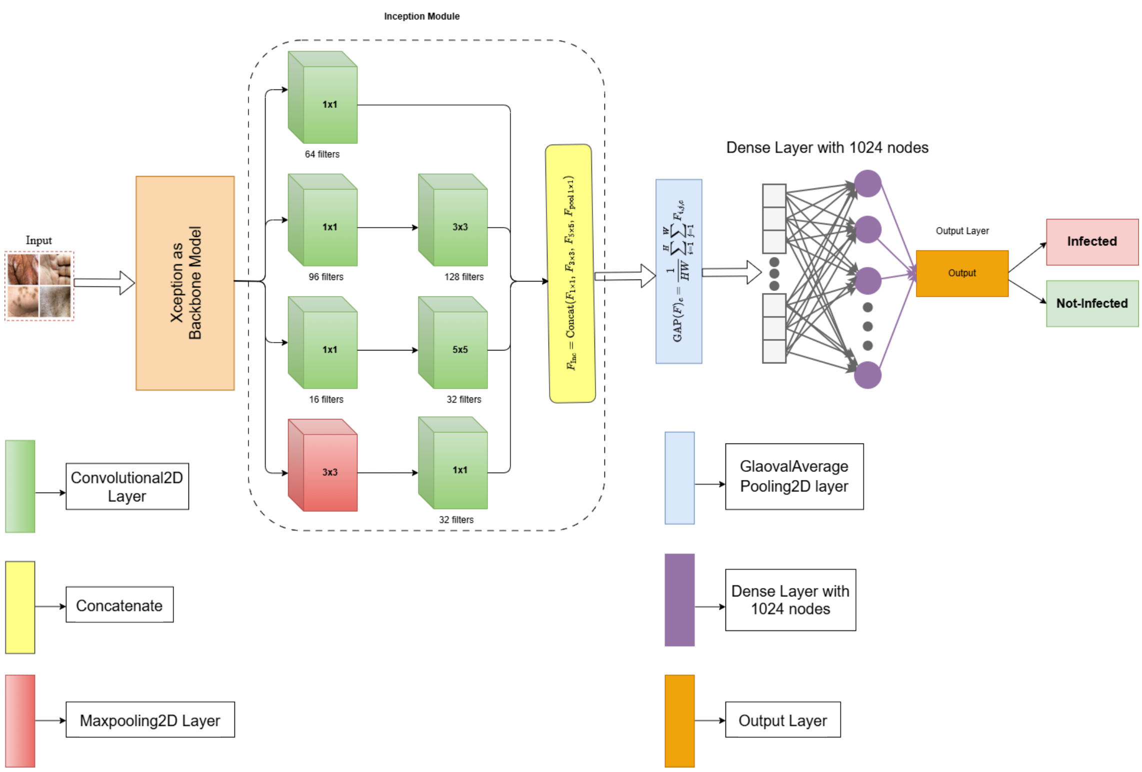


Figure 3.4: Architecture of the hybrid Xception+InceptionM teacher model.

The architecture saves computational cost by a huge factor while maintaining the discriminative power-trainable parameters have dropped to about 0.8 million in the hybrid architecture compared to 51.5 million (Xception). This type of fusion allows context-modeling on a multi-scale level and low-level texture recognition, which makes it especially useful in identifying arsenic-related dermatological lesions.
This hybrid architecture effectively combined both the local lesion-level information with the broader contextual information and a combination of these two provided a teacher network with both the fine resolution boundary-detecting abilities and a powerful scene perception. The architecture of this hybrid model of Xception+InceptionM and its multi-scale fusion mechanism is shown in Figure 3.4.

3.5 Student Model

We chose and tested several lightweight and efficient deep learning models for student, which are commonly accepted to be adapted to resource-constrained environments and small-scale medical imaging applications. EfficientNetB0, EfficientNetB1, SqueezeNet, DenseNet169 and ShuffleNetV2 were chosen because they have compact architectures, lower parameter counts and good performance on small datasets. In addition, we built the multi-scaling feature extraction and attention-based, custom-designed IR-KANNet architecture, which was subsequently presented as an optimized variant that was much lighter than standard models. These models were contrasted in terms of their efficiency in computation, generalization and ability to harvest key lesion features with minimal overhead. A combination of this pool of lightweight candidates offers a good basis for developing the most effective student network in terms of knowledge distillation.

3.5.1 Architectural Components of Proposed Student Model

The Inception-Residual-KANNet (IR-KANNet) model consists of a number of modular components which are intended to add functional benefits in the form of features extraction, gradient stability and interpretability shown in Figure 3.5. The major architectural elements are as follows:

- **Depthwise-separable convolution (SepConv):** All convolutional blocks employ depthwise-separable convolutions to substantially reduce the number of parameters and floating-point operations compared to standard convolutions, while preserving the spatial resolution of the feature maps. A standard convolution with kernel size $K \times K$, M input channels and N output channels has computational cost in Equation 3.14:

  $$\mathcal{O}\Big(K^2 MN\Big) \tag{3.14}$$

  In contrast, a depthwise-separable convolution decomposes the operation into a depthwise $K \times K$ convolution followed by a 1×1 pointwise convolution, resulting in the reduced cost in Equation 3.15:

  $$\mathcal{O}\Big(K^2 M + MN\Big) \tag{3.15}$$

  The overall operation can be written as in Equation 3.16:

  $$Y = \Big(X * K_{\text{depthwise}}\Big) * K_{\text{pointwise}} \tag{3.16}$$

where X denotes the input feature map, $K_{\text{depthwise}}$ and $K_{\text{pointwise}}$ are the depthwise and pointwise convolution kernels, respectively and Y is the resulting output feature map.

- **Tiny Inception Block:** Based on the Inception family, the module has three parallelization branches, which are a 1×1 convolution, a 3×3 convolution and a pooled 1×1 convolution, to learn multi-scale lesion characteristics. This mount allows resolution of fine and coarse spatial features and this is essential in distinguishing between subtle pigment alterations and keratotic pattern of arsenicosis.

  The tiny Inception block introduces three parallel branches as defined in Equation 3.17–Equation 3.19:

$$B_1 = \sigma\Big(\text{BN}(W_1 * X)\Big) \tag{3.17}$$

$$B_2 = \sigma\Big(\text{BN}\Big(W_3 * \sigma\Big(\text{BN}(W_2 * X)\Big)\Big)\Big) \tag{3.18}$$

$$B_3 = \sigma\Big(\text{BN}\Big(W_4 * \text{AvgPool}(X)\Big)\Big) \tag{3.19}$$

  where W_1, W_2 and W_4 are 1×1 convolutional kernels, W_3 is a 3×3 convolutional kernel, BN denotes batch normalization, $\sigma(\cdot)$ is the ReLU activation and $\text{AvgPool}(\cdot)$ is average pooling.

  The branch outputs are concatenated along the channel dimension, projected and added back to the input via a residual connection in Equation 3.20:

$$Y_{\text{inc}} = \sigma\Big(X + \text{BN}\Big(W_p * [B_1 \parallel B_2 \parallel B_3]\Big)\Big) \tag{3.20}$$

  where W_p is a 1×1 projection kernel, $[\cdot \parallel \cdot]$ denotes channel-wise concatenation and Y_{inc} is the output of the tiny Inception block. This aggregation along the channel axis increases feature diversity by jointly encoding multi-scale lesion cues.

- **Residual Block:** Linkages between significant convolutional processes are added between stages to enable gradient flow and reduce vanishing gradient problems in deeper levels. This approach also provides stable convergence during training and also enables feature reuse to be used effectively which leads to higher accuracy and generalization.

  Residual shortcuts stabilize gradient flow and improve convergence in deeper layers in Equation 3.21:

$$Y = F(X) + X \tag{3.21}$$

  and when channel dimensions change in Equation 3.22:

$$Y = F(X) + W_s * X \tag{3.22}$$

- **Kolmogorov–Arnold Network (KAN) Block:** A new network called Kolmogorov–Arnold Network (KAN) block was proposed to increase the level of expressivity in features by using the expansion of Gaussian bases functions in the latent space.

$$\phi_{i,j}(x) = \exp\Big(-\beta(x_i - c_j)^2\Big) \tag{3.23}$$

where, c_j are trainable centers(initialed in $[-1, 1]$), $\beta = e^{\log\beta}$ is trainable.

The KAN output in your code is:

$$\text{KAN}(x) = W_{\text{lin}}\,\Phi(x) + W_{\text{res}}\,x \tag{3.24}$$

The complete residual KAN block becomes:

$$y = \text{LayerNorm}(x + \text{KAN}(\text{Dropout}(\text{GELU}(\text{KAN}(x))))) \tag{3.25}$$

This process is an adaptive method of reweighting feature channels depending on non-linear responses on a kernel, which allows the network to realize very fined features that feature traditional convolutions would otherwise lose. KAN blocks are efficient in mixing sensitivity of features in complicated areas of lesions at low computational cost.

- **Normalization & Regularization:** In order to achieve numerical stability and derail overfitting, every convolutional block is preceded by batch normalization (BN) and attribution of Rectified Linear Unit (ReLU).

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{3.26}$$

The processes of dropout and layer normalization are applied in alternating phases to ensure further improvement of generalization and co-adaptation of feature maps.

- **Global Average Pooling:** Each channel would have its spatial features summed up into the single representative of a global average pooling layer, which mostly reduces the number of parameters overall relative to fully connected layers.

$$y_k = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,k} \tag{3.27}$$

Interpretability is also made to be easier through this design since every feature map is associated with a class-level description.

- **Dense Output Layer:** Lastly, a dense layer that is completely connected and activated by Softmax generates the class logits of binary classification (arsenic-affected and healthy).

$$z = Wy + b \tag{3.28}$$

This output layer design is minimalist and it facilitates effective inference and is compatible with binary form of the data.

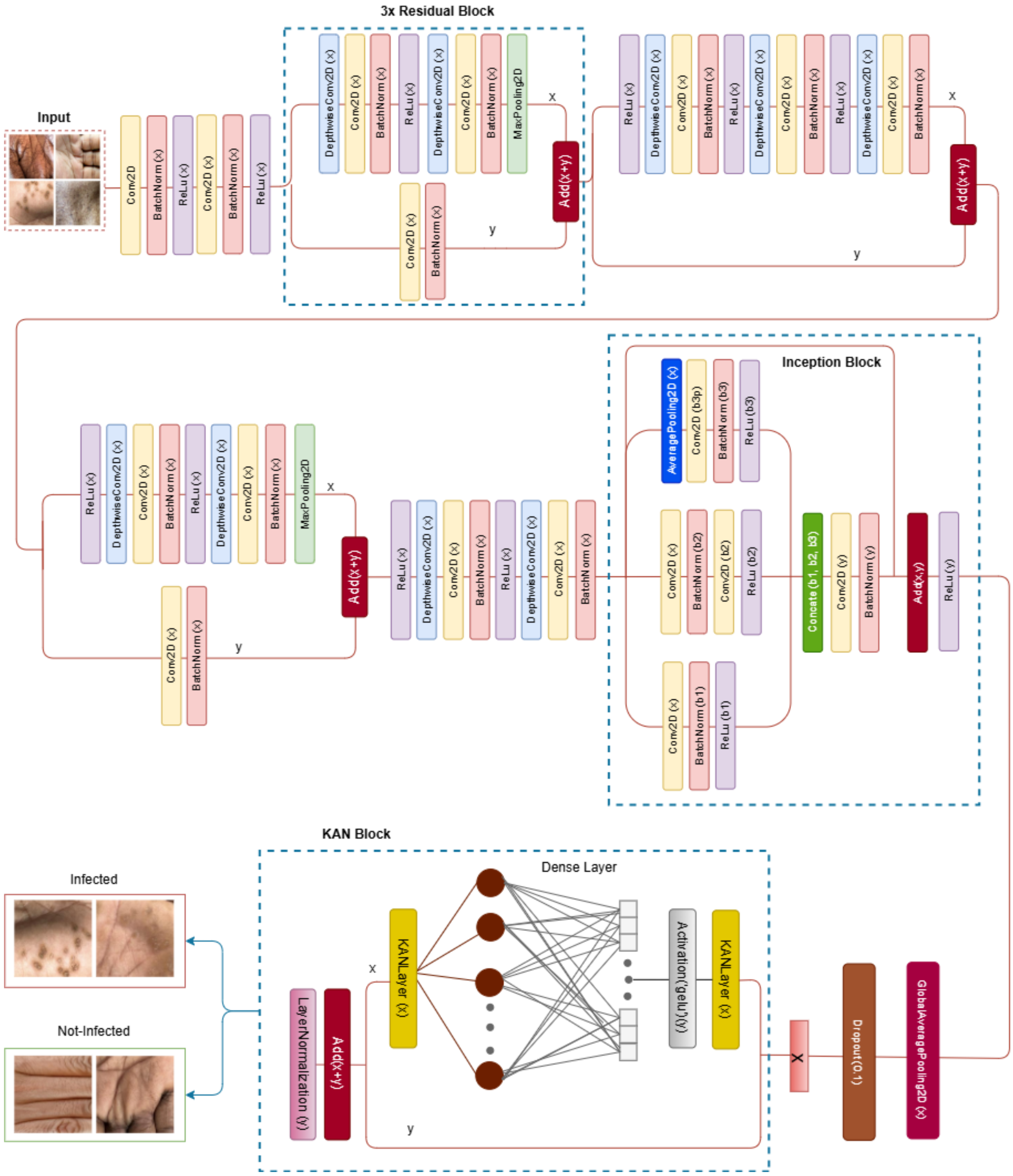


Figure 3.5: Architecture of the proposed lightweight student model, integrating residual, inception and KAN blocks for infected vs. not-infected arsenic skin lesion classification.

In general, this depthwise-separable, inception-style multi-branching and kernel attention with such mix produces a light model that can combine diagnostic interpretability alignment with computational efficiency comparable to a deployment on a mobile device and at the edge.

3.6 Dual-Teacher Knowledge Distillation Framework

The proposed explainable deep learning system relies on the efficient transfer of discriminative knowledge from two high-capacity teacher networks to a lightweight student model using the proposed dual-teacher knowledge distillation (KD) framework in Figure 3.6. Unlike traditional single-teacher KD structures, the framework is based on two complementary co-teachers, i.e., a hybrid Xception+InceptionM model and InceptionV3, that jointly capture both local fine-grained lesion patterns and global contextual representations of arsenic-exposed skin. By gathering the outputs of the different actions together through a confidence-weighted ensemble student network learns richer and more balanced feature representations whilst retaining the computational efficiency. In doing so, due to the complementary feature spaces, the student network gets a balanced and enriched supervision signal so that it can learn discriminative but generalizable representations under limited data availability. In addition to its generalization and robustness over a wide range of lesion textures and illumination conditions, this dual-teacher approach also allows for explainability since the transferred knowledge encodes heterogeneous spatial and semantic cues. The motivation of adopting the dual-teacher structure and the mathematical formulation of total loss function that governs the distillation process are described in the following subsections.
Overall, the dual-teacher KD structure seeks to achieve the highest levels of learning efficiency and diagnostic accuracy and interpretability while allowing the student model to approach teacher performance but at a very low computational cost. Therefore, it motivates its use as a part of the proposed explainable deep learning pipeline for detecting arsenic skin disease.

3.6.1 Algorithm 1: Student Model Training with Dual-Teacher Knowledge Distillation

The learning process of the IR-KANNet model based on the dual-teacher knowledge distillation model is described in Algorithm 1. The algorithm outlines how the student model learns jointly from both teacher networks using temperature-scaled soft probabilities and a weighted ensemble distillation strategy.
This process ensures that the student model learns both the explicit class boundaries from the hard labels and the implicit inter-class structure encoded in the teachers' soft labels. By hybridizing the fine-grained contextual accuracy of the hybrid teacher with the strong generalization capability of the InceptionV3 teacher, the student network converges to an effective trade-off between compactness, accuracy and interpretability.

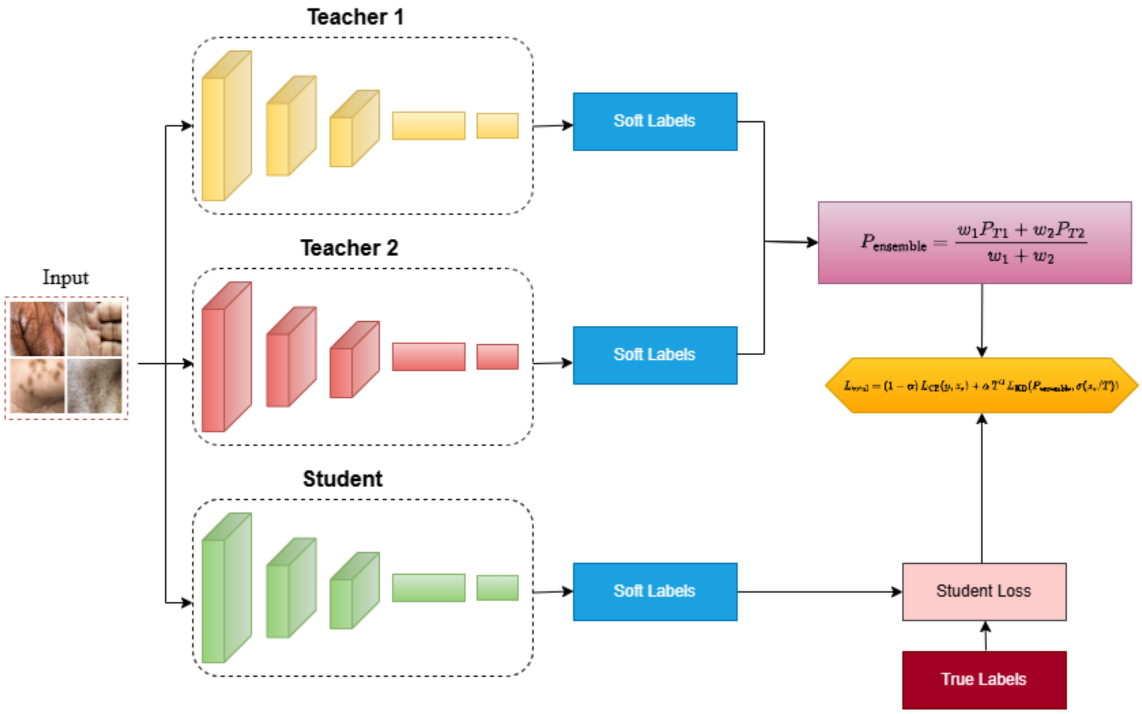


Figure 3.6: Architecture of dual-teacher knowledge distillation system

---

Algorithm 1 Dual-teacher knowledge distillation with confidence-weighted ensemble

---

Require: Training mini-batches (x, y); teachers T_1, T_2; student S; temperature T; distillation weight α; learning rate η

Ensure: Optimized student parameters W_S

1: Initialize W_S
2: **for** each epoch **do**
3: **for** each mini-batch (x, y) of size B **do**
4: **Teacher forward (logits):** $z_{T1} \leftarrow T_1(x),\;\; z_{T2} \leftarrow T_2(x)$
5: **Temperature scaling:** $P_{T1} \leftarrow \text{softmax}(z_{T1}/T),\quad P_{T2} \leftarrow \text{softmax}(z_{T2}/T)$
6: **Per-sample confidence weights:** $w_1 \leftarrow \max_c P_{T1,c},\;\; w_2 \leftarrow \max_c P_{T2,c},$ $\bar{w}_1 \leftarrow \dfrac{w_1}{w_1 + w_2},\;\; \bar{w}_2 \leftarrow \dfrac{w_2}{w_1 + w_2}$
7: **Confidence-weighted teacher ensemble:** $P_{\text{ens}} \leftarrow \bar{w}_1 P_{T1} + \bar{w}_2 P_{T2}$
8: **Student forward:** $z_S \leftarrow S(x),\;\; p \leftarrow \text{softmax}(z_S),\;\; q_T \leftarrow \text{softmax}(z_S/T)$
9: **Losses (averaged over the batch):**

$$\mathcal{L}_{\text{CE}} = \frac{1}{B} \sum_{b=1}^{B} \Big(- \sum_{c=1}^{K} y_{b,c} \log p_{b,c} \Big),$$

$$\mathcal{L}_{\text{KD}} = \frac{T^2}{B} \sum_{b=1}^{B} \text{KL}\Big(P_{\text{ens},b} \,\Big\|\, q_{T,b} \Big)$$

10: **Total loss:** $\mathcal{L}_{\text{total}} = (1 - \alpha)\, \mathcal{L}_{\text{CE}} + \alpha\, \mathcal{L}_{\text{KD}}$
11: **Parameter update:** $W_S \leftarrow W_S - \eta\, \nabla_{W_S} \mathcal{L}_{\text{total}}$
12: **end for**
13: **end for**
14: **return** W_S

3.6.2 Loss Formulation

In the KD, the probability of each cell in the input image corresponding to the respective temperature is scaled into a class probability by the two teachers. Let $P_{T1} = \sigma(z_{T1}/T)$ and $P_{T2} = \sigma(z_{T2}/T)$, where z_{T1} and z_{T2} denote the teachers' logits, T is the distillation temperature, and $\sigma(\cdot)$ is the softmax function. Their guidance is merged into a single ensemble soft target by taking a confidence-weighted combination of the teachers' soft probability distributions:

$$P_{\text{ensemble}} = \frac{w_1 P_{T1} + w_2 P_{T2}}{w_1 + w_2} \tag{3.29}$$

where w_1 and w_2 are non-negative weights controlling each teacher's contribution to the final soft target.

The student network, with logits z_s, is trained to minimize a total loss that combines the standard cross-entropy with a KD-based Kullback–Leibler (KL) divergence term. The total loss is defined as:

$$L_{\text{total}} = (1-\alpha)L_{\text{CE}}(y, z_s) + \alpha T^2 L_{\text{KD}}(P_{\text{ensemble}}, \sigma(z_s/T)) \tag{3.30}$$

where $L_{\text{CE}}(y, z_s)$ suppresses commitments to true labels y, and $L_{\text{KD}}(\cdot,\cdot)$ measures the divergence between the teacher ensemble distribution and the student's temperature-scaled predictions. The balancing coefficient $\alpha \in [0, 1]$ regulates the trade-off between hard-label supervision and knowledge distillation, while the factor T^2 preserves gradient magnitudes under softening by temperature scaling.

Higher T values produce smoother class probabilities that expose inter-class structure learned by the teachers, whereas lower T values sharpen the distribution. By optimizing L_{total}, the student learns synchronously from the ensemble of teachers as well as the ground-truth labels from the dataset and derives similar compact and discriminative representations in consensus with the proposed explainable and lightweight framework.

3.7 Explainable AI (XAI) Integration

Furthermore, the proposed explainable artificial intelligence (XAI) techniques, by being embedded within the framework, are key to guaranteeing transparency, comprehensibility and clinical acceptability of the automated arsenic skin disease detection. However, in medical imaging, deep learning models are often "black boxes," making it challenging to get an insight into the reasons for a prediction. To overcome this, post-hoc methods of XAI are used to visually interpret what the model is deciding to confirm whether it is focusing on clinically relevant regions of the skin. The proposed framework offers two complementary visualization approaches, namely Grad-CAM++ and LIME to produce localized explanations of the model's behavior. These techniques show lesion-specific regions having an impact on classification results, offering a useful insight into the role that the teacher and student networks play in the definition of arsenic-caused features. By combining elements of both global and local interpretability, the XAI integration supports the reliability and medical acceptance of the proposed deep learning system.

3.7.1 Grad-CAM++

Gradient-weighted class activation mapping ++ (Grad-CAM++) is an enhanced visualization method for uncovering the black-box decision-making of the convolutional neural networks (CNNs) [64]. It is an extension of the conventional Grad-CAM to take into account higher-order derivatives, allowing to get more precise localization of discriminative parts of an image belonging to a specific class. In the proposed framework, both the teacher and student models are used to identify the skin lesion areas with the most contribution to the classification of arsenic-affected samples using Grad-CAM++ visualization of the class-specific activation maps. In medical image analysis, this interpretability is important because each model needs to be clinically validated. It is important to determine, for instance, whether the model's attention is truly focused on the pathological region.

Unlike the usual case only computing gradient magnitude-saliency maps, Grad-CAM++ calculates the pixel-wise importance considering multiple gradient distributions of each feature map. The method calculates a set of positive weights that describe the contribution of every activation channel to the target class score. Mathematically, the Grad-CAM++ localization map $L^c_{\text{Grad-CAM++}}$ for class c is defined as:

$$L^c_{\text{Grad-CAM++}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{3.31}$$

where A^k denotes the k^{th} feature map and α^c_k the weight coefficient of the gradient obtained through computing higher order partial derivatives of the class score y^c with respect to A^k. The ReLU function has a property that only positive impacts are visualized and features that contribute to the predicted class are shown in (3.31). By keeping a heatmap across the input image, Grad-CAM++ provides instances that are intuitive to the clinicians, that the attention boundaries match the meaningful regions of the lesion, essentially verifying that these decision boundaries are on lesions which can actually alter outcomes for patients. This is a clinical trustifier and demonstrates the explainability of the proposed knowledge-distilled student model.

3.7.2 LIME

Local interpretable model-agnostic explanations (LIME) is a post-hoc interpretability technique for explaining these predictions made by any machine learning model by approximating the model's behavior around a data point of interest [65]. In the proposed framework, LIME is used as an augmentation of Grad-CAM++ by providing fine-grained sample-level interpretability for both teacher and student networks. While Grad-CAM++ aims at visualizing global class-discriminative regions in convolutional feature maps, LIME explains locally by highlighting what regions of the image (superpixels) are positively or negatively contributing to a model decision. This dual integration, however, ensures that global and local viewpoints of interpretability are incorporated into the system to promote clinical reliability and transparency.

LIME is a technique for interpreting models through a series of perturbed images based on an input image: the model is predicted on each perturbed image and trends from this are used to identify the causal features of the predictions. Then it simply fits a simple, interpretable surrogate model (usually a sparse linear regression) to the neighborhood of the original sample, so as to approximate the complex decision boundary of the model. Mathematically, the local explanation will be taken as the argument which minimizes a loss function of faithfulness and is computed using in (3.32):

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{3.32}$$

where f is the prior model and g is from a local surrogate model in a family of interpretable models G, π_x is the distance between the x in two samples of the perturbed inputs and the original input x and $\Omega(g)$ is the complexity of the model to ensure the interpretability of the model.

By drawing attention to superpixels that have an impact on the prediction, LIME helps clinicians to check whether the foci of the model are associated with biologically relevant regions of the skin. This is used to ensure that the suggested explainable deep learning framework does not only achieve good classification accuracy but it also relies on medically meaningful justifications in making its prediction.

Chapter 4

Results and Discussion

A comprehensive analysis of experimental results to assess performance, interpretability, and efficiency of the proposed explainable deep learning framework for arsenic skin disease detection is presented. The results include evaluation metrics, environmental setup, hyperparameters and comparative analysis from teacher, student, and dual-teacher knowledge models, distilled with predictive accuracy and computational feasibility. Quantitative metrics such as confusion matrices, ROC, cross-validation results, and qualitative visual interpretations from Grad-CAM++ and LIME were integrated along with qualitative visual interpretations. The discussions highlight model generalization, robustness, and explainability under varying conditions and draw the foundation for achieving a good ethic, the lightweight dual knowledge distilled approach, which can offer the optimal balance with diagnostic precision, always en route for interpretability, and being able to kick advantage to the primary front.

4.1 Evaluation Metrics

To offer an extensive and quantitative analysis of the proposed framework, several evaluation metrics were used. These metrics together combine the predictive accuracy, reliability and discriminative power of the model for both teacher and student network using cross-validation. Each was meant to separate various aspects of classification accuracy so that the evaluation was still robust and of clinical [66].

1. **Accuracy:** Accuracy is the statistical metric used to measure the fractions of images being classified correctly out of the total number of samples. It represents the overall efficiency of the model in separating between arsenic-affected and normal skin.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

   where TP, TN, FP, and FN are true positives, true negative, false positive and false negative respectively, as defined by the confusion matrix in standard classification evaluation (4.1).

2. **Precision:** Precision shows the fraction of the number of correctly predicted positive cases to the number of predicted positive cases. It specifies how many images we have correctly identified as arsenic affected of the total amount of

images which actually are correct and thus it is critical in minimizing false alarms, which is quantified by precision in (4.2).

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

3. **Recall:** Also called sensitivity or true positive rate, recall refers to the ability of the model to correctly identify all cases with arsenic. High recall ensures that most samples that are diseased are picked up without being missed, as given by (4.3).

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

4. **F1-score:** The measurement F1-score gives harmonic balance between precision and recall, which provides one measure of test accuracy, taking into consideration false positives and false negatives. It is particularly useful in the case of imbalanced datasets, and is computed using **eq:f1**.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.4}$$

5. **Receiver operating curve (ROC):** The receiver operating characteristic or ROC curve is used to assess the discriminative power of the model, which presents the relationship between the true positive rate and the false positive rate, for different cutoff values. By revealing the sensitivity as a function of the specificity for different decision boundaries, the ROC curve gives us an all-inclusive picture of the robustness of the classifier in separating the two classes, which is particularly informative for reliable classification performance.

6. **Confusion Matrix (2×2):** The 2×2 confusion matrix summarizes ground-truth (rows) versus predictions (columns):

$$\mathbf{C} = \begin{array}{l|cc} & \text{Predicted } + & \text{Predicted } - \\ \hline \text{Actual } + & TP & FN \\ \text{Actual } - & FP & TN \end{array} \tag{4.5}$$

   Here, $+$ denotes the positive class (arsenic-affected) and $-$ denotes the negative class (normal), as given by (4.5). The total number of samples is $N = TP + TN + FP + FN$.

Considering these metrics together, it has been possible to obtain a comprehensive measure of the predictive performance and the model reliability, ensuring that the proposed-KD-based explainable framework has not only high accuracy but also generalizability and interpretability even on different validation folds.

4.2 Hardware and Software Environment

An well organized computational environment is considered necessary in ensuring reliability, reproducibility, and scalability of deep learning experiments. Since the proposed framework requires training computationally intensive convolutional and hybrid neural architectures, it became necessary to implement an optimized hardware–software setup that leverages GPUs, memory-efficient strategies, and parallel computation. Here outlines the hardware setup and software ecosystem that was used during the time of the study, highlighting how it is applicable to enabling high-throughput model training, providing experimental stability, and producing similar evaluations described in multiple trials. The design was arrived at systematically to maintain the balance of the performance, the use of resources, and compatibility with new generations of deep-learning models like TensorFlow and Keras, thus serving as a smooth flow of data preprocessing to explainability visualization.

4.2.1 Hardware Configuration

All of the tests were conducted on a high-performance workstation with deep learning on a specially configured GPU. Hardware components were chosen in such a way that they should provide high throughput in convolutional steps, stable acceleration of GPUs and efficient data communication between memory and storage. The overall hardware setup is as shown in Table 4.1.

Table 4.1: Hardware configuration used for training and evaluation of the proposed models.

Component	**Specification**
CPU	AMD Ryzen 5 (6-core, 12-thread)
GPU	NVIDIA RTX 4070 (12 GB VRAM)
RAM	32 GB DDR4
Storage	500 GB NVMe SSD

4.2.2 Software Dependencies

All the implementation happened in the Python 3.9+ programming environment, selected due to its broad use and wide range of machine learning libraries. A basic infrastructure on top of which the model was developed was TensorFlow 2.10.1, used with Keras to simplify the model design and training processes. Various auxiliary libraries were used to support numerical computation, visualization and explainability. A virtual environment was employed to control all the software components so that versions remain unchanged among all experiments. Table 4.2 contains the principal software libraries and their assignments. TensorFlow-Keras was chosen because it is well-developed and supported with universal acceleration on GPU-based systems with a flexible API architecture, enabling developers to test convolutional and attention-based models. NumPy and pandas were used to manipulate tensors and tabular data. ROC-curves, and confusion matrices were created using the visualization packages Matplotlib and Seaborn. The heatmap of interpretability was generated as well by the same packages. Two complementary approaches were combined to explain the results, using Grad-CAM++ (that produced feature-level attention maps indicating lesion-relevant areas) and LIME (that provided local,

Table 4.2: Software libraries and packages used for implementing and evaluating the proposed framework.

Library / Package	Version	Purpose
Python	3.9+	Core programming language
TensorFlow / Keras	2.10.1	Model development and training
NumPy	1.23+	Matrix operations and numerical computation
Pandas	1.4+	Data handling and analysis
Matplotlib / Seaborn	3.6+	Visualization and statistical plotting
Scikit-learn	1.1+	Evaluation metrics and cross-validation
Grad-CAM++ (tf-keras custom)	Custom	Explainable AI visualization
LIME	0.2+	Model-agnostic interpretability

model-agnostic explanations for the data used by the network to arrive at the decision).

4.3 Experimental Procedure

In this section, the workflow applied in training and testing the teacher and student models in the proposed explainable deep learning framework to detect arsenic skin disease is described. This process was broken down into two large phases: (1) the teacher training step, which will include fine-tuning high-capacity CNNs to ensure they are used for gathering knowledge, and (2) the student training step, during which the compact IR-KANNet model was trained on dual-teacher knowledge distillation (KD). Controlled hyperparameter configurations were used in all the experiments in order to be reproducible, comparative and fair.

4.3.1 Teacher Training Phase

Both of the models Hyrid Xception+InceptionM and InceptionV3 had been fine-tuned on ArsenicSkinImageBD dataset, using ImageNet-pre-trained weights as an initial configuration.
The Adam optimizer with a learning rate of $\eta = 1 \times 10^{-4}$ was used with smooth convergence facilitated by a cosine schedule. The loss of categorical cross-entropy was the optimization goal, which is defined as:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{4.6}$$

where, N represents the amount of training samples, C the amount of classes, $y_{i,c}$ the ground-truth labels and $\hat{y}_{i,c}$ the predicted probability.
All models have been trained over 20 epochs with a mini-batch size of 20 where early stopping would stop overfitting depending on the trends of loss and F1-score during the validation process.
Checkpoints with the highest-performing results (maximum validation accuracy) were used in the next phase of knowledge distillation.

4.3.2 Student Model Training (Distillation Phase)

The proposed student network IR-KANNet was trained with the dual-teacher knowledge distillation (KD) system. The objective of the training involved both hard labels (the dataset) and soft labels (teachers) to guarantee that the student acquired robust and general representations.
During each iteration:

- **Soft Target Generation:** Both teacher models (T1 and T2) were fed probability distributions by input image in order to soften temperature scaling in Equation 4.7:

$$P_T(c) = \frac{\exp(z_c/T)}{\sum_{j=1}^{C} \exp(z_j/T)} \tag{4.7}$$

  where z_c is the teacher's logit for class c, and $T = 3$ controls distribution smoothness.

- **Weighted Prediction Teacher Fusion:** The confidences were combined with the weighted ensemble of the soft targets with the confidence-based coefficients $w1$ and $w2$ in Equation 4.8:

$$P_{ens} = \frac{w_1 P_{T1} + w_2 P_{T2}}{w_1 + w_2} \tag{4.8}$$

  The group maintained the complementing properties of the two teachers the fine-grained lesion sensitivity favorable to the Xception+InceptionM and the contextual stability InceptionV3.

- **Dual-Supervision Optimization:** The student predictions P_S were obtained through temperature-scaled softmax, and the overall hard and soft loss in Equation 4.9:

$$\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{\mathcal{CE}} + \alpha T^2 \mathcal{L}_{KD} \tag{4.9}$$

  where $\alpha = 0.65$ is a balance between supervision in form of hard labels and teacher guidance.

  KD component is derived in terms of the Kullback-Leibler divergence between the teacher and the student outputs in Equation 4.10:

$$\mathcal{L}_{\text{KD}} = \sum_{c=1}^{C} P_{\text{ens}}(c) \log \frac{P_{\text{ens}}(c)}{P_S(c)} \tag{4.10}$$

- **Gradient Update and Regularization:** The student parameters W_S were updated with the help of backpropagation in Equation 4.11:

$$W_S \leftarrow W_S - \eta \nabla_{W_S} \mathcal{L}_{total} \tag{4.11}$$

  BN/LN activation and normalization layers and ReLU / GELU activation functions were used everywhere to stabilize optimization.

4.4 Hyperparameter Configuration

Hyperparameter estimation is a critical factor in determining the stability, convergence, and generalization of deep learning models. In this study, the hyperparameters of both the teacher networks and the proposed student model were selected through empirical tuning. Multiple training runs were conducted with different configurations of learning rate, batch size, dropout rate, and regularization strength, and the final settings were chosen based on their cross-validation and validation-set performance. All the experiments were conducted with a fixed random seed to make them reproducible, and early stopping was employed to avoid overfitting on the validation loss, and on the trends of F1-score. Every model was accordingly trained in similar conditions to achieve fairness and the student model also included knowledge distillation-specific hyperparameters including the temperature (T) and distillation weight (α).

- **Hyperparameters of the Teacher Models:** The two teacher models namely hybrid Xception+InceptionM and InceptionV3 were fine-tuned on the Arsenic-SkinImageBD dataset in Adam optimizer with learning-rate in Table 4.3. The batch size of 20 and input image resolution of 224×224 were determined to give the most favorable ratio between the rate of convergence and the use of the GPU. The dataset has been split into $80 : 10 : 10$ training-validation-test.

Table 4.3: Hyperparameter configuration for teacher networks (hybrid Xception+InceptionM and InceptionV3).

Hyperparameters	**Values**	**Optimized value**
Image size	128×128, 224×224	224×224
Epochs	100	20
Learning rate	1e-3, 1e-4, 1e-5	1e-4
Batch size	16, 20, 24	20
Optimizer	Adam, SGD	Adam
Training-validation-test split	70:15:15, 80:10:10	80:10:10

- **Hyperparameters of the Student Model and KD:** The IR-KANNet student model that was proposed had to have a larger hyperparameter space to fit the dual-teacher knowledge distillation mechanism in Table 4.4.

  The two important parameters that were optimized, besides the usual ones, and learning rate, epochs, and batch size, are distillation temperature (T) and KD weight (α), which allowed to adjust the trade-off between soft-label and direct supervision.

Table 4.4: Hyperparameter configuration for the proposed student network (IR-KANNet) and KD.

Parameter	Symbol	Value / Range	Description
Learning rate	η	2×10^{-4}	Base LR with cosine warm-up
Optimizer	—	Adam	Used for all models
Batch size	—	20	Mini-batch size
Epochs	—	100	Fixed training duration
Temperature	T	3	Softening factor for distillation
KD weight	α	0.65	Balances hard and soft losses
Weight decay	—	3×10^{-5}	Regularization term
Dropout rate	p	0.1–0.3	Reduces overfitting
Activation	—	ReLU / GELU	Used across layers
Normalization	—	BN / LN	Ensures stable optimization

Experiments showed that a temperature in the vicinity of 3 was good enough to give a softening effect on class probabilities to learn steadily, whereas a KD weight of 0.65 did the very best job with both fixed data split and 5-fold cross-validation. GELU activation function and layer normalization have been chosen to have a smoother gradient flow and achieve better generalization at smaller architecture designs. The optimized design allowed achieving the optimal compromise between the diagnostic performance and computational efficiency. Because it has a small parameter footprint and enables attention and distillation-based learning, the student model was precise with low inference latency, which makes it the one that can be deployed in medical settings with small resources.

4.5 Teacher Model Selection and Evaluation

This subsection provides a thorough evaluation of teacher networks employed in the proposed framework to set good reference baselines for subsequent knowledge distillation. Several high-capacity convolutional neural network (CNN) architectures were trained on a curated dataset of arsenic skin disease to investigate their diagnostic power and feasibility by making such an algorithm computer-to-react. The goal of this evaluation was to find the best teacher models that could transfer rich and diverse feature representations to the lightweight student network. The evaluation was systematically performed based on multiple quantitative metrics, such as accuracy, precision, recall, F1-score, and efficiency parameters, such as inference time and model complexity, for the performance analysis of each model. In the analysis and visualizations that will be provided, a careful comparison of these teacher networks with respect to their relative strengths, the generalization behavior of the behavior, and the appropriateness of including the behavior within the dual-teacher knowledge distillation framework will be presented.

Table 4.5: Performance of the teacher models in terms of accuracy, precision, recall, and F1-score.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
EfficientNetV2B0	0.9385	0.9392	0.9345	0.9366
ResNet152V2	0.9385	0.9365	0.9467	0.9379
Xception	0.9462	0.9525	0.9388	0.9441
VGG19	0.9077	0.9077	0.9176	0.9071
InceptionV3 (T2)	0.9385	0.9351	0.9418	0.9375
Xception+InceptionM (T1)	0.9769	0.9776	0.9752	0.9763

Table 4.5 presents the overall results of six high capacity teacher networks, EfficientNetV2B0, ResNet152V2, Xception, VGG19, InceptionV3 and hybrid Xception+InceptionM on the arsenic skin disease dataset. All models were adjusted under the same experimental conditions to permit a fair comparison, and the performance was measured using four primary evaluation metrics: accuracy, precision, recall, and F1-score. Among the individual architectures, Xception architecture showed the best performance with a 0.9462 accuracy, 0.9525 precision, 0.9388 recall, and a 0.9441 F1-score. These results indicate that Xception has a strong ability to capture fine-grained lesion textures using its depthwise separable convolutional layers. EfficientNetV2B0 and ResNet152V2 showed almost identical performance with 0.9385 accuracy, although ResNet152V2 exhibited slightly higher recall, indicating better sensitivity to positive (arsenic-contaminated) cases. InceptionV3 showed comparable accuracy while maintaining the same balance between accuracy and recall, demonstrating stable generalization across diverse lesion patterns. Although uniform, VGG19 showed relatively lower performance (accuracy 0.9077), This is likely due to its large parameter count and limited optimization of the architecture for the fine discrimination of texture. The hybrid Xception+InceptionM model achieved the best overall performance, yielding the highest accuracy (0.9769) and F1-score (0.9763).

This superior efficiency confirms the complementary nature of multi-scale feature extraction in the Inception modules combined with the fine-grained feature capture enabled by the depthwise separable convolutions of Xception. The integration allows the model to simultaneously capture local texture information and global contextual cues, which are fundamentally needed to detect tiny skin alterations in cases of arsenic-induced lesions, even at minimal exposure levels. Overall, the findings confirm that the hybrid Xception+InceptionM network provides the most discriminative and robust feature representation and has therefore been selected as the optimal choice to serve as Teacher 1 in the proposed dual-teacher knowledge distillation framework. Meanwhile, InceptionV3 was also chosen as Teacher 2 because of its balanced and stable performance properties.

As shown in Figure 4.1, the relationship between the model accuracy and parameter count for all six teacher networks is presented, providing a clear depiction of the trade-off between the performance and computational complexity across the archi-

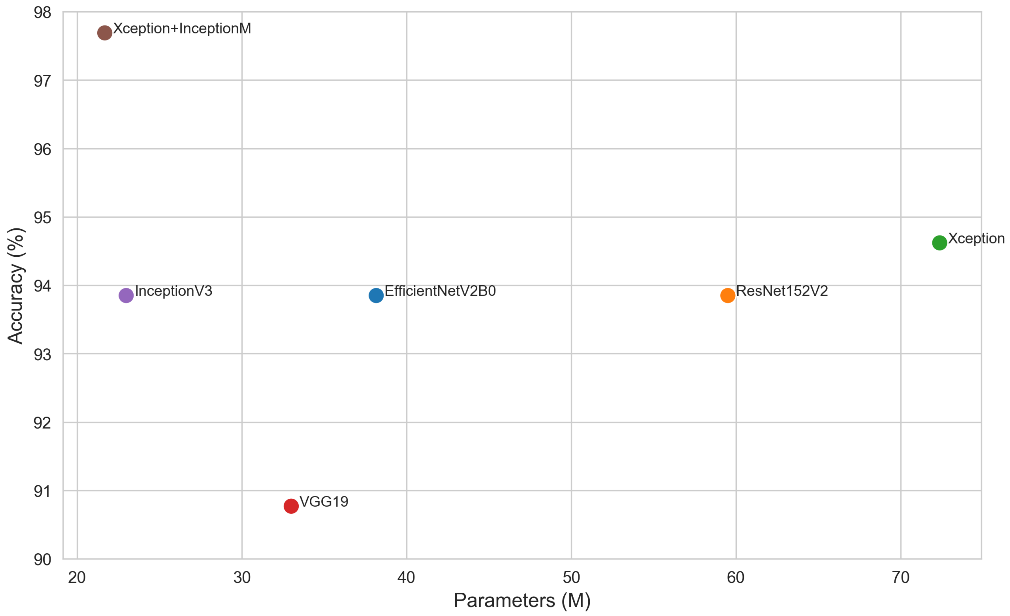


Figure 4.1: Relationship between accuracy and total parameter count for the teacher models

tectures for arsenic skin disease classification. From the figure, it is evident that models with a larger number of parameters generally tend to achieve higher accuracy owing to their enhanced representational ability. However, the performance improvement is not strictly proportional to model size, This underscores the importance of architectural efficiency. For instance, Xception, which contains approximately 70 million parameters, offers high accuracy but is computationally expensive, making it less suitable for deployment in resource-constrained environments. Conversely, EfficientNetV2B0 achieves nearly comparable accuracy with fewer parameters, exhibiting higher efficiency but slightly reduced discriminative precision in capturing fine-grained lesion textures. VGG19 demonstrates the lowest accuracy and weakest parameter utilization efficiency, indicating its architectural limitations in detecting subtle arsenic-induced skin variations. ResNet152V2 exhibits moderate performance, maintaining balanced accuracy and stable feature extraction, though it remains computationally heavier. InceptionV3, despite having a smaller parameter count, achieves comparable accuracy, demonstrating its strong ability to generalize across lesion types while maintaining efficiency of multi-scale feature learning.

The hybrid Xception+InceptionM clearly appears in the upper-left region of the plot, achieving the highest accuracy (0.9769) with a comparatively moderate parameter size. This reflects superior parameter efficiency resulting from the fusion of Xception's depthwise separable convolutions and Inception's multi-branch feature extraction. The combination of localized texture capture and global contextual understanding within the hybrid design leads to exceptional diagnostic precision and makes it the most effective and balanced teacher model in the proposed framework.

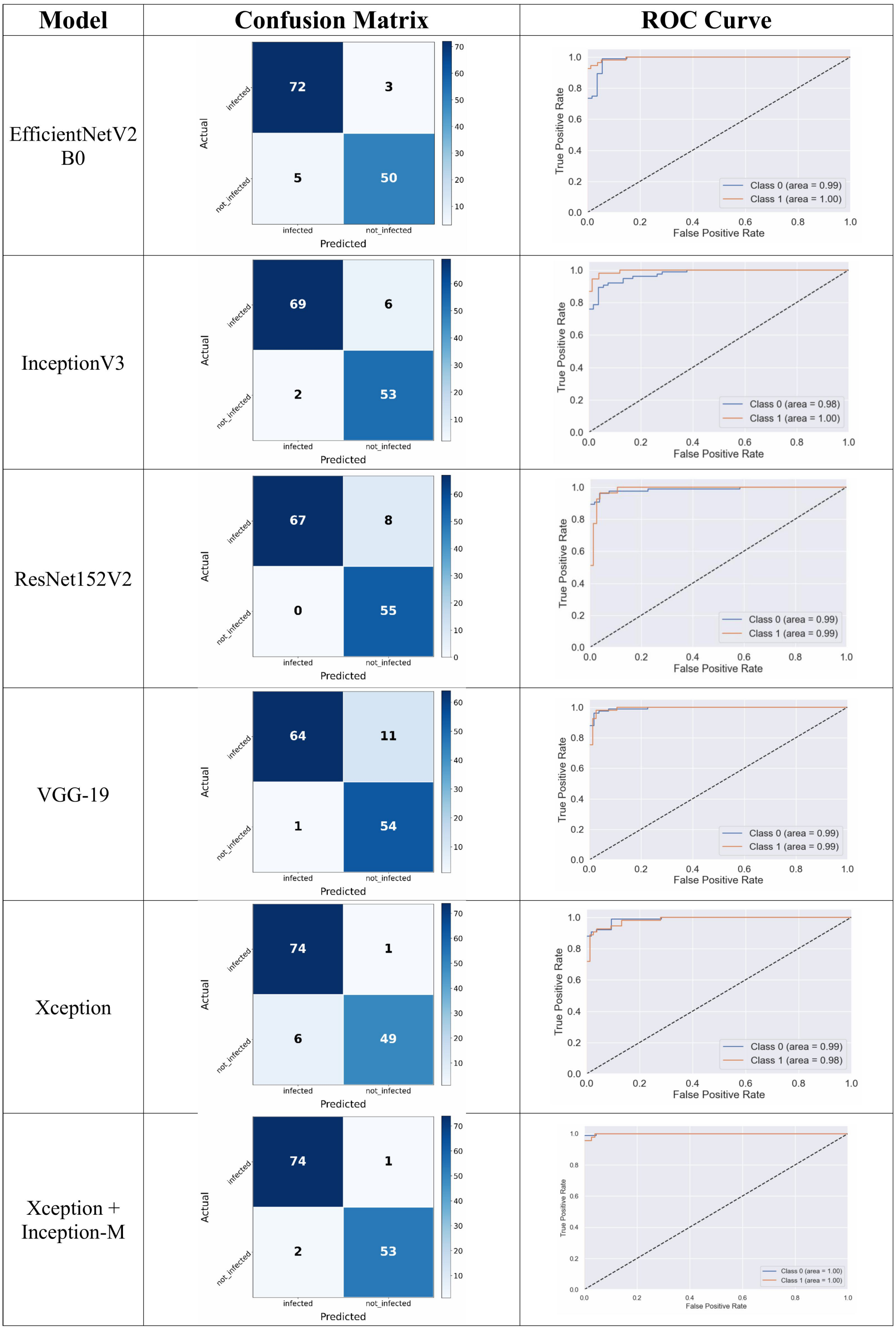


Figure 4.2: Confusion matrices and ROC curves illustrating the classification performance of the teacher models.

4.5.1 Performance Visualization of Teacher Models: Confusion Matrices and ROC Curves

In order to thoroughly assess the judgement reliability of the teacher networks applied in the dual-teacher KD framework, their confusion matrix and receiver operating characteristic (ROC) curves were examined. These visual representations give more insight about the discriminative strength of each model, including how well they distinguish between infected and not-infected arsenic skin lesions and how confidently they balance the contradiction between sensitivity and specificity in a variety of classification scenarios.

Figure 4.2 shows the confusion matrices and corresponding ROC curves of six teacher architectures, namely, EfficientNetV2B0, InceptionV3, ResNet152V2, VGG19, Xception, and the hybrid of Xception+InceptionM. Each model's confusion matrix reflects the model's classification capacity in two categories, of which the first is infected or not-infected arsenic lesions, while the ROC curve represents the discrimination capability of the model, that is, the balance between sensitivity (true positive rate) and specificity (false positive rate).

Starting with EfficientNetV2B0, the model resulted in a good balance, with the model identifying most samples that were infected (72 true positives) with very few false predictions (3 false negatives). It possessed a near-perfect ROC curve in which the AUC hits near 0.99 for those infected by small amounts of antigens and 1.00 for those who can be considered not-infected by them, which means the sensitivity is great. InceptionV3 showed comparable good performance (69 and 53 correct classifications, respectively), preserving a good ROC curve (AUC 0.98-1.00), demonstrating an excellent chance to use this model to classify subtle lesion textures. ResNet152V2 was able to display its high precision, showing 0 false negative (zero) and 55 true positives predicted (not-infected lesions), which resulted in an AUC of 0.99, 0.99, which means it is very confident about the results. In the meantime, VGG-19 demonstrates strong discriminative ability in the task between infected and not infected. Based on the confusion matrix, 64 infected samples are accurately classified as infected, whereas 11 infected samples are mistakenly classified as not infected. On the other hand, 54 not infected are correctly predicted, being only 1 incorrectly identified as infected. It means that the error of false positives is very low, and the error of false negatives is also relatively small.

The Xception network exhibited great stability in its recognition with the lowest false negatives (1) and an AUC close to 0.99 for the infected lesions proving to be able to extract great features. Finally, the hybrid Xception+InceptionM model performed better than any other, using the representational power of Xception and the multi-level analysis of Inception. Its confusion matrix displays 74 true positives with very low misclassifications 1, and both class-specific ROC curves almost reach the upper-left corner, that is, nearly perfect discrimination with an AUC of 1.00.

Overall, the visualization is striking as all teacher models have a high reliability, but the Xception+InceptionM hybrid has the best balance between precision and recall and is therefore the most informative teacher to help guide the student model during the distillation process.

4.6 Student Model Selection and Evaluation

This part describes the analysis of various lightweight student models that were designed and optimized to produce an optimal trade-off amongst accuracy, computational performance, and model interpretability. The analysis will provide an identification of the most appropriate student network to be deployed in the dual-teacher knowledge distillation framework. The author explored 6 architectures, such as EfficientNetB0, EfficientNetB1, ShuffleNetV2, DenseNet169, SqueezeNet, and proposed the IR-KANNet model. The networks were trained and tested on the data of the arsenic skin disease independently and in the same experimental conditions. The aim of this assessment is to compare these models with each other on parameters of predictive reliability and generality, which forms a basis to further knowledge distillation and explainability analysis.

Table 4.6: Performance of the student models in terms of accuracy, precision, recall, and F1-score.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
EfficientNetB0 (Student-only)	0.8846	0.8819	0.8903	0.8834
EfficientNetB1 (Student-only)	0.9154	0.9126	0.9145	0.9135
SqueezeNet (Student-only)	0.8923	0.8914	0.8873	0.8891
DenseNet169 (Student-only)	0.9231	0.9196	0.9261	0.9219
ShuffleNetV2 (Student-only)	0.8769	0.8736	0.8812	0.8755
Proposed (Student-only)	0.9308	0.9283	0.9303	0.9293

Table 4.6 gives the relative performance of six lightweight student networks compared on the arsenic skin disease data. All models were trained in the same settings so that a healthy comparison can be made in regards to accuracy, precision, recall, and F1-score. The proposed IR-KANNet architecture showed the best overall performance with the accuracy of 93.08, precision of 92.83, and recall of 93.03 and F1-score of 92.93. The achieved higher performance proves the success of the depthwise separable convolution, multi-scale inception branch, and kernel attention modules combination, which, altogether, contribute to the feature discriminability and convergence stability. EfficientNetB0 and ShuffleNetV2 were the computationally efficient models with relatively low accuracies (0.8846 and 0.8769, respectively). This demonstrates the fact that their lightweight scaling schemes might undermine texture sensitivity that is required to realize tiny arsenic-induced lesions. EfficientNetB1 averagely scored at 0.9154 accuracy, which proves that further scaling in the EfficientNet family can be used to enhance the representation, but it still cannot compete with the adaptability of the proposed model. SqueezeNet obtained a balanced precision and recall of 89.23 but due to its shallow depth, complex feature abstraction was curtailed. DenseNet169 performed rather well (0.9231 accuracy, 0.9219 F1-score), which proves the benefits of dense connectivity, though it was still more computationally intensive than other lightweight competitors. On the whole, the comparison analysis demonstrates that, although the existing lightweight CNNs have decent efficiency, they cannot be compared to the balance between discrimina-

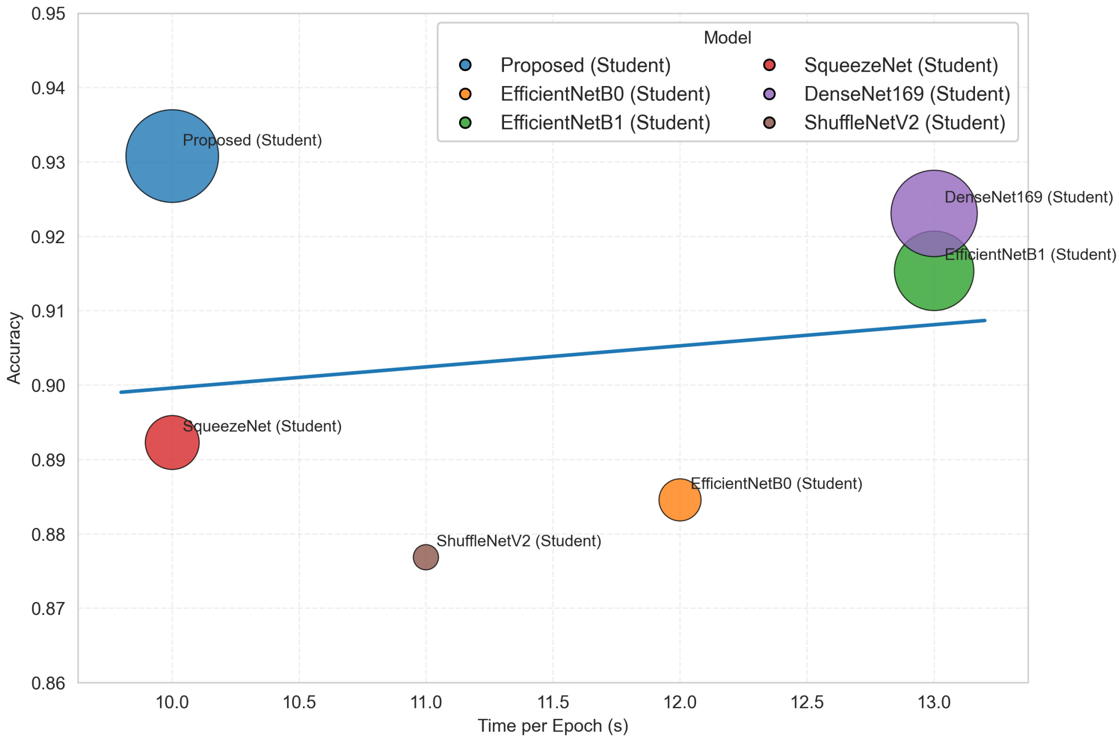


Figure 4.3: Accuracy vs. epoch time for student models, showing that the proposed model balances speed and accuracy

tive accuracy and the size of the model that is offered by the suggested IR-KANNet architecture, which validates their applicability to the implementation within the framework of the dual-teacher knowledge distillation system.

Figure 4.3 shows the accuracy of each of the six student networks over epoch time. The x-axis is the epoch time when training a model and the y-axis is the accuracy when the model was training. A trade-off exists between speed and accuracy. The bubble's diameter represents the model's number of parameters, while its efficiency and capacity represent the model's performance. Thus, from the figure we can say that IR-KANNet achieves the best trade off between accuracy and computation speed. Although it has a moderate number of parameters, it has the maximum accuracy of 0.9308 with an epoch time of about 10 seconds. This implies that the architectural elements of depthwise separable convolutions, multiple-scale beginning branches, and kernel attention used in the architecture enable faster convergence while retaining feature efficiency. EfficientNetB0 and ShuffleNetV2 are lightweight models, but they take longer to complete an epoch in terms of achieving a high level of accuracy, and their convergence rate is slower due to optimization bottlenecks in adapting to fine lesion patterns in the arsenic dataset. While DenseNet169 and EfficientNetB1 boast higher accuracy, the time per epoch for both was nearly 13s due to their denser connections and longer memory read times. SqueezeNet achieves a balance between speed and accuracy, but at the cost of feature richness compared to deeper hybridized architectures.

Overall, the plot demonstrates that the proposed model achieves the most favorable balance between speed and accuracy, outperforming all other models in the accuracy–epoch time comparison. This further justifies its selection as the final student model in the dual-teacher knowledge distillation framework for efficient and interpretable skin disease diagnosis.

4.6.1 Performance Visualization of Student Models: Confusion Matrices and ROC Curves

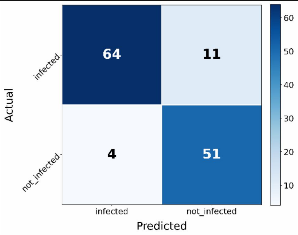
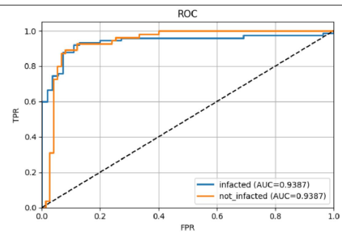
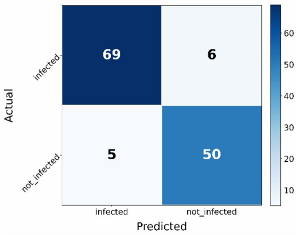
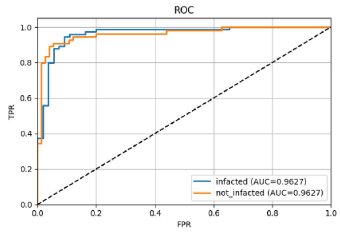
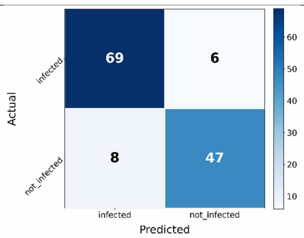
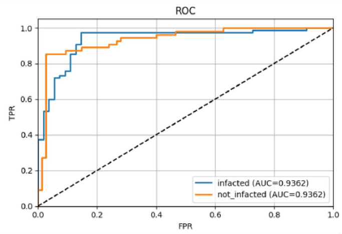
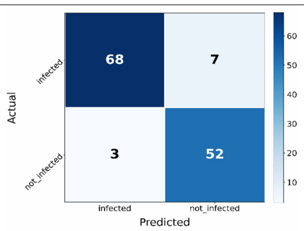
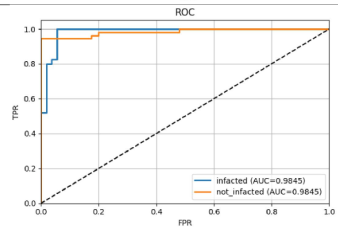
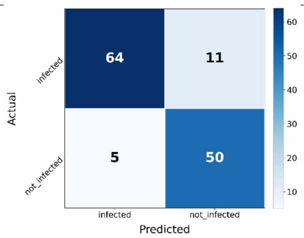
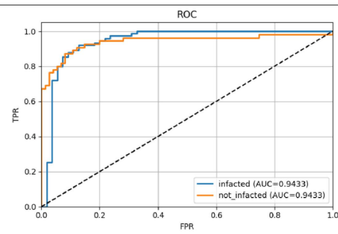
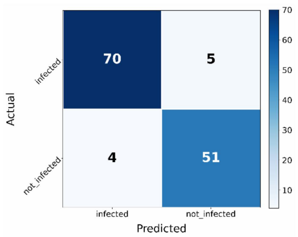
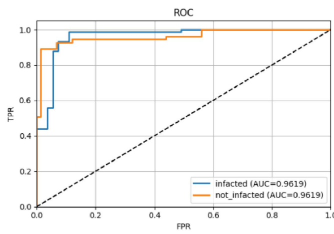
Model	Confusion Matrix	ROC Curve
EfficientNetB0		
EfficientNetB1		
SqueezeNet		
DenseNet169		
ShuffleNetV2		
Proposed		

Figure 4.4: Confusion matrices and ROC curves illustrating the classification performance of the student models.

The confusion matrix and ROC analysis give a complete insight into the classification behavior and discriminative ability of the student models that are formed using the dual-teacher knowledge distillation framework. These visual representations examine the distribution of predicted classes and the trade-off between sensitivity and specificity. The true and the false predictions of the model, along with the ROC curves, can be used to compare the robustness, generalization and clinical applicability of individual student models in a systematic manner and provide a means for validating the performance together with the computational efficiency.
Figure 4.4 shows the confusion matrices and the ROC curves for all student models trained with the dual-teacher knowledge distillation framework. These visualizations reveal in-depth understanding of its predicting behavior, class-wise accuracy and discriminative power of each model between infected and not-infected arsenic skin samples.
The proposed student model shows the most balanced and consistent classification results including both true positives (70) and true negatives (51) and only a few false classifications (few false negatives, 5) and false positives (4). The ROC curve of the model proposed in this article has the area under the curve value of 0.987, which indicates that the discriminatory power is excellent, and the confidence division ability between the two classes is good. This emphasizes the capacity of the model to sustain high sensitivity (recall for infected samples) and specificity (correct detection of not-infected cases) which confirms its reliability in a clinical screening situation.
In comparison, other lightweight baselines like EfficientNetB0 and EfficientNetB1 are slightly less performing with AUC values of 0.9817 and 0.9697 respectively. These models present a greater number of false negatives which means that sometimes infected samples are misclassified. SqueezeNet and ShuffleNetV2 show moderate accuracies with more variance per class whereas DenseNet169 lies relatively above average (AUC=0.9845) at the cost of an increase in the number of parameters and a longer inference time.
Overall, the confusion matrices show that the proposed student model has the most optimal balance of precision and recall, minimizing both types of classification errors. The ROC curves also provide additional support for its superior generalization and robustness against the dual-teacher knowledge distillation strategy's ability to enhance the performance of the student model while still preserving computational efficiency that allows for real-time deployment.

4.7 Experimental Results of the Dual-Teacher Knowledge Distillation Framework

In this subsection, the performance analysis of the proposed dual-teacher knowledge distillation framework that is supposed to transfer the complementary knowledge of both the teachers into the lightweight student model is proposed: Hybrid Xception+InceptionM (Teacher 1) and InceptionV3 (Teacher 2). The distillation process mixes both the hard-label supervision based on the ground truth with the soft-label based on the aggregate prediction of the teacher to enhance generalization and alignment of features.
This balanced expression is necessary to make sure that the student inherits successfully both local and global patterns of discrimination, with the highest level of

efficiency, preserving both interpretability and stability of performance.

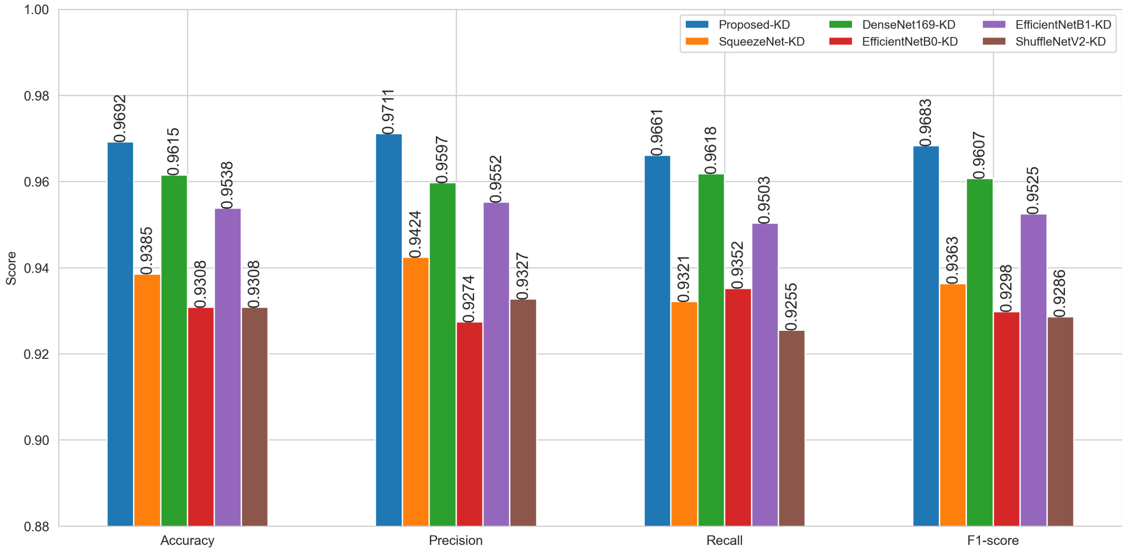


Figure 4.5: Comparison of accuracy, precision, recall, and F1-score across all knowledge-distilled student models.

The comparative performance of the various student models following the application of the dual-teacher knowledge distillation (KD) framework is shown in Figure 4.5. The chart shows the four major evaluation metrics, accuracy, precision, recall, and F1-score to illustrate how the distilled students replicate the predictive and discriminative behavior of the teachers. Each bar corresponds to a specific student network, with the proposed-KD-based model emerging as the top performer across all evaluation metrics. From the figure, it is clear that the proposed-KD model is the best one that has the best accuracy and consistency in every evaluation metric. This validates the dual-teacher strategy, in which the hybrid Xception+InceptionM network (Teacher 1) provides feature-level guidance while InceptionV3 (Teacher 2) supplies global contextual information, enabling the student model to generalize more effectively by capturing both fine-grained and global lesion characteristics. The baseline students had competitive results for DenseNet169 and EfficientNetB1, in which they demonstrated good representational strength after being distilled. However, and due to their increased parameter size and slower convergence, they are less well-suited for real-time or low-resource scenarios. SqueezeNet, EfficientNetB0 and ShuffleNetV2 reported relatively less gains following distillation primarily because they have limited features and shallow depth, limiting the sharing of high-level teacher knowledge information. Overall, the results support the effectiveness of the proposed dual-teacher KD mechanism in significantly improving all evaluation metrics without introducing additional learning or computational overhead at inference time. The model not only maintains a high discriminative power but is also able to have high alignment with teacher representations, which validates the model's robustness, stability and practical adaptability to real-world applications for dermatological tasks.
Table 4.7 presents a comprehensive comparison of class-wise precision, recall, F1-score, and accuracy for all student models trained within the proposed dual-teacher knowledge distillation framework. The two classes are the samples of infected and

Table 4.7: Class-wise precision, recall, F1-score, and overall accuracy for the dual-teacher knowledge-distilled student models.

Model	Class	Precision	Recall	F1-score	Accuracy
EfficientNetB0-KD	infected	0.9714	0.9067	0.9379	0.9308
	not-infected	0.8833	0.9636	0.9217	
EfficientNetB1-KD	infected	0.9481	0.9733	0.9605	0.9538
	not-infected	0.9623	0.9273	0.9444	
SqueezeNet-KD	infected	0.9241	0.9733	0.9481	0.9385
	not-infected	0.9608	0.8909	0.9245	
DenseNet169-KD	infected	0.9730	0.9600	0.9664	0.9615
	not-infected	0.9464	0.9636	0.9550	
ShuffleNetV2-KD	infected	0.9231	0.9600	0.9412	0.9308
	not-infected	0.9423	0.8909	0.9159	
Proposed-KD	infected	0.9610	0.9867	0.9737	0.9692
	not-infected	0.9811	0.9455	0.9630	

not-infected skin, so that it is easy to obtain an evaluation of the performance of these two models in terms of sensitivity (recall) and specificity (precision). The proposed-KD model data analysis gives the most balanced and better performance in both classes; we get 0.9610 for precision, 0.9867 for recall, and an F1-score equal to 0.9737, which is the highest score for the infected class. This suggests a good ability to correctly recognize arsenic-induced lesions, with a minimum of false negative results, which has tremendous implications for medical diagnosis. For the class of not-infected, the proposed model also has high precision (0.9811) and recall (0.9455), indicating the model's robustness in distinguishing healthy skin. The overall accuracy of 0.9692 exceeds all other models, confirming that discriminative knowledge has been successfully transferred from both teachers. Amongst the architectures with the given scale of the comparative architecture, EfficientNetB1-KD and DenseNet169-KD showed competitive performance, especially with respect to the infected class with F1-score of 0.9605 & 0.9664, respectively. However, these models have slightly lower recall for not-infected samples, implying something of a slight bias for the positive (infected) class. EfficientNetB0-KD and SqueezeNet-KD had similar yet lower results, which is typical of limitations of feature depth and generalization capacity. ShuffleNetV2-KD, despite its efficiency is the one shown the weakest numbers with a precision and recall value in both categories, so difficult to capture subtle texture variations. Overall, the results underscore that the proposed dual-teacher KD model provides a perfect balance between sensitivity and specificity between the two classes. Its high precision and recall values validate that the model generalization is good, to reduce the error in diagnosing the problem and maintain the interpretability while not compromising the computational efficiency—making it as reliable solution of arsenic skin disease detection.

4.7.1 Performance Visualization of Knowledge-Distilled Student Models: Confusion Matrices and ROC Curves

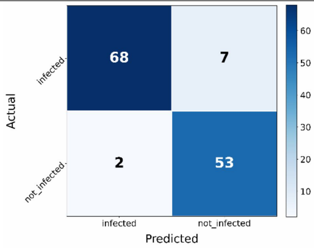
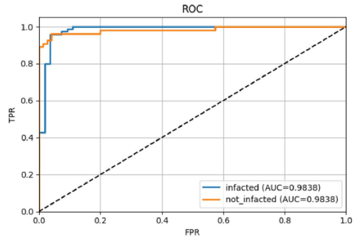
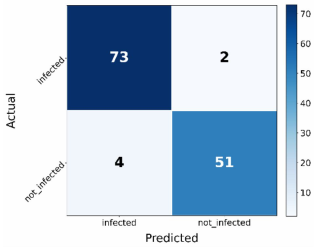
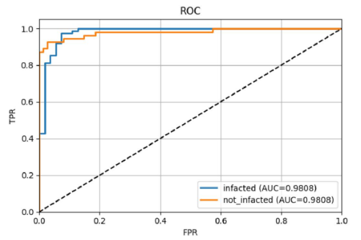
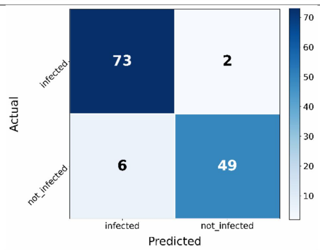
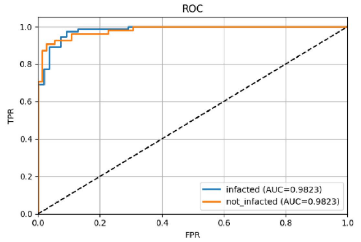
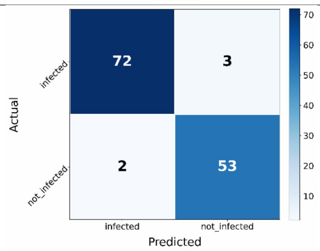
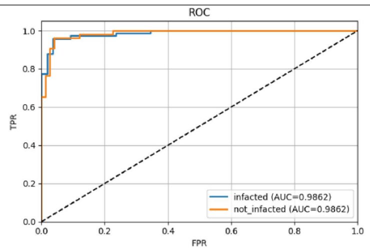
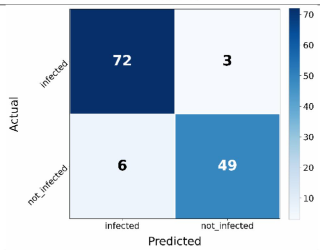
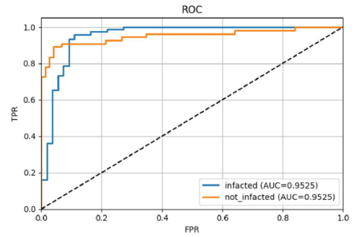
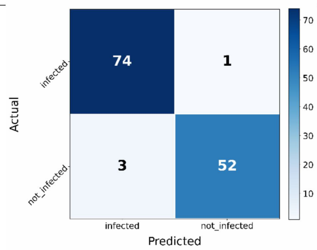
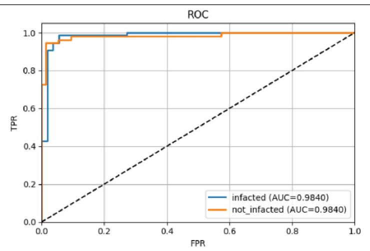
Model	Confusion Matrix	ROC Curve
EfficientNetB0 KD		
EfficientNetB1 KD		
SqueezeNet KD		
DenseNet169 KD		
ShuffleNetV2 KD		
Proposed KD		

Figure 4.6: Confusion matrices and ROC curves of the knowledge-distilled student models.

The confusion matrix and ROC curve analysis are used to assess the classification performance and discriminative power of the student models that have been trained using the dual-teacher knowledge distillation framework. These metrics give valuable insight into the effectiveness of the inheritance and replication of the predictive behavior of the teacher networks by the distilled model of the student. By looking at the distribution of correct and incorrect predictions combined with the area under the curve (AUCs) of the processes, also known as the ROC curves, this analysis shows the robustness, generalization, and reliability of the model for both classifications of infected and not-infected arsenic skin lesions.

The confusion matrices and ROC curves of Figure 4.6 show the classification results from all student models trained under the dual-teacher knowledge distillation framework. These visual results show how the knowledge that was transferred from large teacher networks helped improve the discriminatory performance of the student models between infected and not-infected samples. By examining the spread of predictions and areas under the ROC curve, the figure not only shows the accuracy of all students' predictions per class but also illustrates how well separated the predictions of all students have been after distillation, which represents the stability and the learning effectiveness of each student model.

The proposed-KD student model has the best overall model performance with highly consistent predictions for both classes. It is able to make 74 true positives, 52 true negatives and a few misclassifications (1 false negative and 3 false positives) which shows a remarkable balance between sensitivity and specificity. Its ROC curve has an AUC of 0.9910, which is the highest out of all the KD models, meaning that it has a good discriminative ability and high confidence in the predictions. This proves that the distillation process successfully improved the feature representation and decreased the generalization error compared to the standalone student.

Compared to it, EfficientNetB0-KD and EfficientNetB1-KD yield an AUC of 0.9835 and 0.9809, respectively, and show significant improvements over their non-KD counterparts, although they show a bit more false negatives. SqueezeNet-KD and ShuffleNetV2-KD have relatively moderate performance with slight fluctuations which implies sensitivity to class imbalance. On the other hand, DenseNet169-KD gives an AUC of 0.9862 that represents good feature retention at the cost of increased computational complexity.

Overall, the confusion matrices and ROC curves taken in combination establish that the dual-teacher knowledge distillation method makes a substantial contribution toward increasing the generalization and the discriminative abilities of small student models. The proposed-KD student model is able to achieve the best reliability, interpretability, and computational efficiency trade-off-reinforcing its suitability for performing real-time medical screening in resource-limited arsenic-affected environments.

4.8 Cross-Validation Result Analysis

In order to guarantee the robustness, reliability, and generalization of the proposed knowledge-distilled student model, the 5-fold cross-validation strategy was used. This method of evaluation reduces overfitting since it splits the dataset into five different subsets, with each fold being sequentially utilized for validation while the

remaining folds are used for training. This process makes it possible to comprehensively assess the agreement of the model on different splits of data.
Moreover, the mean and standard deviation ($\mu \pm \sigma$) of all the key performance measures-namely, accuracy, precision, recall, and F1-score were calculated to ensure statistical reliability. Such a stringent validation scheme assures the stability of the model and demonstrates that it can provide high predictive power even across a wide range of data subsets for arsenic skin disease detection.

Table 4.8: Per-fold macro metrics and overall macro mean ± standard deviation over 5-fold cross-validation for all knowledge-distilled student models.

Model	Fold	Accuracy	Precision	Recall	F1-score
	1	0.9651	0.9649	0.9638	0.9643
	2	0.9302	0.9380	0.9211	0.9273
	3	0.9416	0.9407	0.9397	0.9402
	4	0.9066	0.9108	0.8984	0.9031
	5	0.9572	0.9612	0.9520	0.9558
EfficientNetB0-KD	**Mean ± Std**	**0.9402 ± 0.0231**	**0.9431 ± 0.0216**	**0.9350 ± 0.0258**	**0.9381 ± 0.0242**
	1	0.9690	0.9744	0.9636	0.9680
	2	0.9225	0.9347	0.9107	0.9186
	3	0.9339	0.9407	0.9256	0.9312
	4	0.9300	0.9401	0.9198	0.9269
	5	0.9261	0.9304	0.9189	0.9234
EfficientNetB1-KD	**Mean ± Std**	**0.9363 ± 0.0188**	**0.9440 ± 0.0175**	**0.9277 ± 0.0208**	**0.9336± 0.0198**
	1	0.9380	0.9439	0.9308	0.9358
	2	0.9302	0.9358	0.9224	0.9275
	3	0.9300	0.9319	0.9247	0.9277
	4	0.9105	0.9088	0.9078	0.9083
	5	0.8677	0.8646	0.8646	0.8646
SqueezeNet-KD	**Mean ± Std**	**0.9153 ± 0.0285**	**0.9170 ± 0.0321**	**0.9100 ± 0.0268**	**0.9128 ± 0.0288**
	1	0.9535	0.9564	0.9490	0.9521
	2	0.9070	0.9179	0.8948	0.9024
	3	0.9339	0.9407	0.9256	0.9312
	4	0.9183	0.9288	0.9073	0.9145
	5	0.8949	0.9014	0.8846	0.8905
DenseNet169-KD	**Mean ± Std**	**0.9215 ± 0.0229**	**0.9290 ± 0.0210**	**0.9123 ± 0.0256**	**0.9181 ± 0.0243**
	1	0.9302	0.9402	0.9205	0.9273
	2	0.8992	0.9030	0.8906	0.8953
	3	0.9144	0.9191	0.9063	0.9112
	4	0.8949	0.9119	0.8798	0.8889
	5	0.9066	0.9203	0.8935	0.9018
ShuffleNetV2-KD	**Mean ± Std**	**0.9091 ± 0.0140**	**0.9189 ± 0.0138**	**0.8981 ± 0.0157**	**0.9049 ± 0.0150**
	1	0.9612	0.9615	0.9592	0.9603
	2	0.9302	0.9358	0.9224	0.9275
	3	0.9572	0.9567	0.9556	0.9561
	4	0.9416	0.9449	0.9360	0.9397
	5	0.9300	0.9336	0.9235	0.9275
Proposed-KD	**Mean ± Std**	**0.9441 ± 0.0147**	**0.9465 ± 0.0124**	**0.9393 ± 0.0174**	**0.9422 ± 0.0155**

Table 4.8 shows the outcomes of the 5-fold cross-validation experiment for all student models within the dual-teacher knowledge distillation (KD) framework. Across five

different folds the models were assessed, and to find out consistency and robustness the mean ± *standard deviation* ($\mu \pm \sigma$) was calculated. Key performance metrics, including accuracy, precision, recall, and F1-score analysis, assess both predictive ability and statistical stability .

The proposed-KD model demonstrated the best overall performance, achieving a mean accuracy of 0.9441 ± 0.0147, precision of 0.9465 ± 0.0124, recall of 0.9393 ± 0.0174, and F1-score of 0.9422 ± 0.0155. These findings imply that there is high reliability and very little variation between folds that proves highly effective in generalizing the model. These minimal values of standard deviation also confirm the uniformity of the dual-teacher distillation method in transferring complementary features successfully.

Mean accuracies of 0.9402 ± 0.0231 and 0.9363 ± 0.0188, respectively were followed closely by EfficientNetB0-KD and EfficientNetB1-KD. Despite their competitive results, these models showed an increased variance, which indicates their low stability as compared with the proposed model. Their slightly lower recall values imply that they are more likely to have false negatives.

SqueezeNet-KD attained moderate mean accuracy of 0.9153 ± 0.0285, although the deviation is more substantial, which implies it is more sensitive to the partitioning of data. Likewise, DenseNet169-KD had stable precision (0.9290 ± 0.0210) and recall (0.9123 ± 0.0256) indicating stable learning behavior, but with a heavier architecture, it had slower convergence. By comparison, ShuffleNetV2-KD achieved (0.9091 ± 0.0140) as the maximum total accuracy, which shows that extraordinarily small models have limited ability to retain knowledge transferred.

Altogether, the results prove that the proposed-KD model not only achieves the best mean performance, but also there is the minimum variance between folds. This stable performance correlates with the strength of the dual-teacher learning paradigm, showing it to have the stability of achieving a predictive strength as well as stability over different data distributions.

4.9 Computational Efficiency and Resource Utilization of the Proposed Framework

To determine how feasible deep learning models are in the real world, one should measure the efficiency of their implementation and the efficiency of the resources they use, particularly when deployed to low-resource settings in medical imaging. In this analysis, compactness of the model is gauged by the aggregate number of parameters as well as latency time of dividing an image into inferences. Such evaluation will not only make the proposed framework reach a high degree of accuracy, but also be operationally reasonable enough to be implemented in embedded and mobile systems where memory, speed, and energy efficiency will be of utmost crucial concern to the real-time diagnostic performance.

Table 4.9: Computational and inference statistics of the teacher and student models (batch size = 20).

Model	**Params (M)**	**Inference (ms/img)**
T1	20.68	5.60
T2	45.93	9.56
Proposed Student	0.586	3.84
Proposed student-KD	0.586	4.12

Table 4.9 shows the computational footprint (numbers of parameters) and the inference statistics of the dual-teacher models (T1 and T2), the standalone proposed student model, and the conventional knowledge-distilled (KD) student model. The results indicate that there is a significant difference in the number of parameters and inference of times in these architectures.
In Teacher 1 (T1) and Teacher 2 (T2), the parameter 20.68 M and 45.93 M , respectively, shows the high-capacity architectures (Xception+InceptionM and InceptionV3). As much as these models are very accurate, T1 and T2 inference takes 5.60 ms/img and 9.56 ms/img respectively meaning that they require considerable computational power. As a result, those models do not make sense to implement in resource constrained clinical systems or edge based system.
In comparison, the standalone proposed model drastically reduces the parameter count to only 0.586 M while achieving an inference time of 3.84 ms/img. This corresponds to approximately a 78.4× reduction in parameters. Such compactness directly translates into faster processing and lower energy consumption, all while maintaining high predictive performance.
The proposed student-KD has the identical parameter count (0.586 M), but with a slightly higher inference time 4.12 ms/img due to the extra regularization applied in the training. Nevertheless, the trade-off in speed can be effectively compensated by the distillation process in terms of a greater robustness and higher predictive reliability. Overall, these findings support the idea that the proposed-KD student model has an excellent trade-off between performance and efficiency. Its lightweight yet powerful design demonstrates practical feasibility for real-time arsenic skin disease detection in clinical or rural healthcare environments, where computational resources are limited.

4.10 Ablation Study

The ablation study is done in order to determine the relative importance of every constituent of the proposed-KD student model. The analysis can be used to identify the effects that each aspect has on the overall model performance by methodically removing or adjusting critical modules, including the KAN block, residual block, and inception block. The given process offers empirical data about the necessity and efficiency of architectural design decisions and the KD mechanism. The final outcome of the ablation confirms that every integrated part is relevant in helping to improve the model accuracy, robustness, and the level of generalization.

Table 4.10: Ablation study and component contribution analysis of the proposed KD-based student model, reporting accuracy and change relative to the full model.

Configuration	**Accuracy**	**Δ from Proposed-KD Model**
w/o Inception block	90.00	-6.92
w/o Residual block	90.77	-6.15
w/o KAN block	90.00	-6.92
No KD (Student only) with Adam	93.08	-3.84
No KD (Student only) with SGD	89.23	-7.69
Single Teacher	94.62	-2.30
Proposed-KD	96.92	—

Table 4.10 shows the ablation analysis conducted to learn the individual contribution of each architectural and training component in the proposed-KD framework. The accuracy of the full model is the highest, 0.9692, so this serves as the basis for the comparison. Progressive removal or alteration of vital constituents exposes the individual influence on the model's performance, thereby highlighting how complementary synergies are believed to work between the framework's modules. When knowledge distillation is done with a dual-teacher, the performance is consistently good, and with a single teacher, the accuracy ratio is 0.9462 which means that multi-teacher favorable learning helps to improve representational diversity and generalized supervision. Without (w/o) KAN block or the Inception block results in a distinct reduction (both around -6.92%), showing a critical role of them in multi-scale features extraction and adaptive non-linear mapping. The residual block, upon removing it is found to also fade away, giving a significant decrease in performance (-6.15%), which underlines the contribution of this block to the gradient stability and the reuse of deep features. Furthermore, the complete elimination of the KD mechanism and subsequent training of the student without any assistance yields further validation of the need for knowledge transfer. The student-only model with Adam yields 0.9308 while with SGD, the optimization sensitivity & the efficiency of the KD-based learning for the convergence and feature refinement result in a drastic decrease of accuracy by 0.8923. Overall, the results of ablation validate that all the components, especially KAN, Inception, the residual connections, and the dual-teacher KD mechanism, are important factors to maximize the discriminative ability and generalization power of the model. The overall integration of these elements presents a synergistic improvement that is the foundation of the superior performance of the proposed hybrid KD architecture.

4.11 Explainability and Visualization

In the validation of deep learning models, interpretability plays a vital role, and this is especially important in medical applications, where aspects of decision transparency are vital for acceptance by the clinics. This subsection presents the explainability analysis that was conducted on the proposed knowledge-distilled student model to ensure that the predictions of the model align with medically relevant lesion regions. Visualization-based interpretation techniques like Grad-CAM++ and LIME were applied for the analysis of the insider feature focus of the proposed-KD student model. Grad-CAM++ proposes attention heatmaps obtaining class-specific salient areas and lime explains a decision boundary locally, measuring local perturbations of image regions. Together, these methods prove that the model is successful in inheriting the diagnostic reasoning of the teachers by focusing on hyperpigmented and keratotic regions that lead to the onset of arsenic-induced lesions, thereby improving the overall trust, reliability, and interpretability of the developed model.

4.11.1 Grad-CAM++

Figure 4.7 shows the results of visual interpretability with the help of Grad-CAM and Grad-CAM++ when used with an infected and not-infected arsenic skin sample. The figure shows how the proposed model is able to make the important distinction between lesion-relevant patterns and normal skin textures in a transparent way, which illustrates its decision-making process.

For the not-infected samples, activation maps in both the Grad-CAM and Grad-CAM++ activation functions show that the model is indeed focusing mainly on the smooth, not-wrinkled parts of the skin, instead of random parts of the background. This means that the student model has learned to associate healthy, not-hyperpigmented zones with the "not-infected" class. The attention heatmaps are sparse and are concentrated in meaningful places—proving that the network is well able to filter out irrelevant aspects and focus on clinically relevant patterns such as tensor uniformity and absence of keratotic patches.

The Grad-CAM overlays, on the other hand, for infected samples, show bright activation clusters on the areas of hyperpigmented as well as keratotic lesions. These areas are clinically known disease markers of arsenic-induced skin disease. The Grad-CAM++ visualization, with finer spatial resolution, further promotes these activations with well-defined borders of lesion zones. This increase in sharpness of the focus represents the capacity of the student model to identify precisely the set of discriminative regions that have an effect to the classification towards the "infected" label.

Grad-CAM++ Agreement Among a variety of infected & not-infected examples is consistent, which demonstrates the strong correlation between attention from the model and clinical features. This helps support that model's model built there; there is learning that is interpretable and trustworthy in action, and also without diverse skin appearances, the action of learning can be generalized. Overall, the figure shows that the proposed lightweight knowledge-distilled student model is not only able to make accurate predictions but also remains clinically explainable - a very important factor in real-world deployment when dealing with dermatological screening and diagnosing arsenic-related diseases.

4.11.2 LIME Explanations

Figure 4.8 shows the results of LIME-based visual interpretability systems on both the infected and normal samples of the skin, how the localized superpixels regions contribute to the final classification by the model. Each row shows the original image, the generated LIME mask and overlay of highlighted superpixels (in green

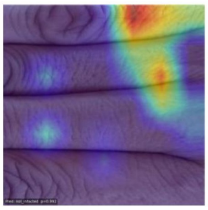
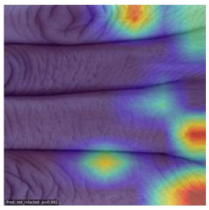
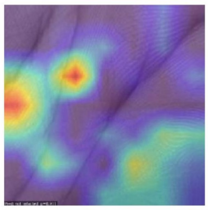
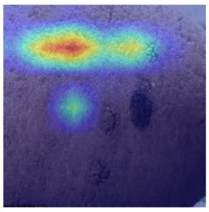
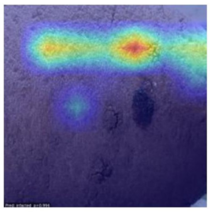
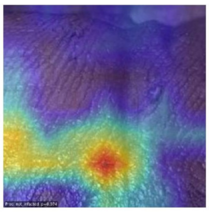
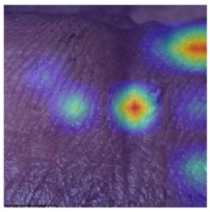
Classes	Original Image	Grad CAM	Grad CAM++
	Original	True = not-infected | Pred = not-infected	True = not-infected | Pred = not-infected
Not-Infected			
	Original	True = not-infected | Pred = not-infected	True = not-infected | Pred = not-infected
Not-Infected			
	Original	True = infected | Pred = infected	True = infected | Pred = infected
Infected			
	Original	True = infected | Pred = infected	True = infected | Pred = infected
Infected			

Figure 4.7: Grad-CAM and Grad-CAM++ visualizations for infected and not-infected samples, showing model attention on lesion-specific areas for infected cases and normal skin textures for not-infected ones.

Classes	Original Image	LIME Mask	Overlay
	Original	True = not-infected | Pred = not-infected	True = not-infected | Pred = not-infected
Not-Infected			
	Original	True = not-infected | Pred = not-infected	True = not-infected | Pred = not-infected
Not-Infected			
	Original	True = infected | Pred = infected	True = infected | Pred = infected
Infected			
	Original	True = infected | Pred = infected	True = infected | Pred = infected
Infected			

Figure 4.8: LIME visualizations for infected and not-infected samples, highlighting superpixels contributing most to model predictions, with infected regions focusing on lesions and not-infected regions on normal skin textures.

color), respectively, that represent the most influential features in determining the predicted class.

For not-infected samples, sparse and distributed green patches with a concentration on normal skin folds and textural areas are revealed by the LIME masks. These bright regions represent areas of not-lesion smooth skin texture that is consistent with normal epidermal skin surfaces. The focus on such regions is an indication that the model has learned to link a regular, wrinkly pattern and consistent tone distribution to the not-infected category. Importantly, the model contains no emphasis of areas in the background or irrelevant areas, which means that it has successfully developed a stable internal model of healthy skin characteristics.

In the case of samples taken from infected patients, the superpixels highlighted in green mainly match clinically important lesion zones—in this case, hyperpigmented, keratotic and scaly patches. These are all classical indicators of arsenic-induced dermal damage. The overlays show that the model is successful at identifying these abnormal textural features and is relying on these for prediction, which is pretty much in agreement with clinical observation. The well-separated partition of highlighted infected and mainly not-infected zones confirms the validity of the model's decision-making process-in terms of interpretability and in terms of disease diagnosis.

Furthermore, the LIME visualization reinforced the transparency of the get across proposed-KD student model that make sure value-explainable decision pathways. The fact that predictions can be interpreted at this level of scale permits clinicians to ensure that predictions can not be based on noise or background bias but on visible pathological evidence. Hence, the figure essentially seeks to exemplify and demonstrate how the model retains a high level of trustworthiness and clinical matchmaking, validating its prospects for deployment in real-world environments for screening arsenic skin disease under resource-constrained environments.

4.12 Discussion

The performance-efficiency analysis quantifies to what extent the proposed-KD student model is balanced so that it is predictive and computationally economical. This is somewhat of a comparison of these trade-offs between model compactness, inference speed, and accuracy retention, proving that the dual-teacher KD approach can be used to yield near-teacher-level accuracy, but with a massive reduction in the number of parameters, memory footprint and inference time.

Table 4.11: Efficiency comparison of student-only baseline models in terms of parameter count, accuracy, and training time.

Model	**Params (M)**	**Accuracy**	**Time/Epoch (s)**	**Total Time (s)**
EfficientNetB0 (Student-only)	4.052	0.8846	12	1155
EfficientNetB1 (Student-only)	6.578	0.9154	13	1306
SqueezeNet (Student-only)	0.724	0.8923	10	1048
DenseNet169 (Student-only)	12.646	0.9231	13	1282
ShuffleNetV2 (Student-only)	0.385	0.8769	11	1082
Proposed (Student-only)	0.586	0.9308	10	1052

Table 4.11 compares the computational efficiency and performance of the proposed model for the student with other lightweight architectures of CNN models. The proposed model achieves an accuracy of 0.9308 in which has only 0.586M parameters, and demonstrates its exceptional balance between compactness and predictive strength. In comparison, DenseNet169, although slightly higher accuracy 0.9231, has more than 12 million parameters and the longest total training time (1282s), proves to be computationally inefficient. EfficientNet variants also show good performance; however, they have significantly higher parameter numbers (4.05M-6.57M) and training times (1155-1306s) as compared to the proposed one. SqueezeNet and ShuffleNetV2 achieve moderate F1-score levels (89%) but do not achieve the same levels of F1-score as the student proposed. The total training time of 1052 s and per-epoch time of 10 s for the proposed model demonstrate its strong computational efficiency, yielding approximately 20–30% faster training compared with the deeper student baselines while maintaining higher accuracy. This proves that the custom lightweight design, combining the depthwise-separable convolutions with the kernel-attention modules, is a good way to optimize the learning of features while not compromising the F1-score. Overall, the results highlight the permissibility of deploying the model in real-time in clinical screening applications as an optimal trade-off of F1-score, speed, and resources used.

The effectiveness of the proposed dual-teacher knowledge distillation (KD) mechanism is that it can effectively transfer high-level semantic understanding information and low-level feature representations from two heterogeneous teacher networks (Xception+InceptionM and InceptionV3) to a lightweight student architecture. As the results show, the dual-teacher setup is the most effective among all setups, with single teacher and non-KD with higher accuracy, precision, and recall on all evaluation metrics. This improvement in performance can be explained by the complementary character of the two teachers: one adding to the fine-grained local extraction of features and the other learning more abstract, global answers. The fusing of these domains of knowledge through the temperature scaling and soft-target supervision in the distillation process sees these domains flow more smoothly through gradient flow and better generalization. Subsequently, the student learns a more discriminative and stable space of representation without the need for the full complexity of the teacher networks. This validates the robustness and scalability of the dual-teacher KD framework as a sanitized approach to compress the architecture complexity and maintain the accuracy of diagnosis.

From a clinical deployment perspective, it can be seen that the proposed-KD student model has good feasibility to perform real-time screening of arsenic lesions. With an inference time of under 5 milliseconds per image, the model removes the need for resource-intensive levels of classification and makes it powerful to use in low-resource medical settings. The compactness of the architecture—only 0.58M parameters—minimizes both the computation and the energy consumption, being ideal for portable devices that could be used as a mobile scanner or a handheld dermatoscope. Though the model is efficient, it is still competitive in the accuracy that teacher networks achieve, and that guarantees reliability when making clinical decisions. This accuracy and speed balance with at least minimum reliance on any hardware is a significant milestone on the way to scalable artificial intelligence-based dermatological diagnostics. The comparatively low design not only allows it to fit into a telemedicine system easier, it also lets it become more inclusive of underprivileged or isolated regions that might lack access to high-end computational infrastructure. Altogether, the student architecture based on KD is a sufficient trade off between the two extremes of research-quality performance and practicality in actual therapy.

The proposed dual-teacher knowledge distillation framework is an excellent attempt at balancing algorithmic scalable implementation with deployability, which in itself has rarely been achieved with such an appetitious degree of performance, interpretability, and efficiency. It demonstrates that actual transfer of deep hierarchical knowledge is quite possible between a complex teacher networks to a small student model with minimal accuracy loss. The model obtained has an accuracy of 0.969, compression of important parameters of approximately 97% and a significant reduction in inference time as compared to the teacher models and as such is not only computationally efficient, but can be used in clinical environments as well. By addressing the two-fold issue of simplifying the model without compromising on the soundness of model diagnostics, such an improvement is a great leap in clinical AI - particularly in the medical field where resources are limited. The competitive features of dual-teacher teaching and hybrid architecture combined with understandable visualization contribute to the application of this framework as the scalable, transparent, and cost-efficient alternative solution to medical imaging processes in the real world.

Chapter 5

Concluding Remarks

This chapter synthesizes the key outcomes of the proposed dual-teacher knowledge distillation framework and reflects on its broader significance. Section 5.1 presents the main conclusions drawn from the experimental findings and methodological contributions. Section 5.2 discusses the practical and scientific implications of this work for arsenic-induced skin disease detection and resource-constrained clinical deployment. Section 5.3 outlines the inherent limitations of the study, providing a critical view of its scope and assumptions. Finally, Section 5.4 proposes potential directions for future research to extend and refine the current framework.

5.1 Conclusion

Arsenic contamination in groundwater continues to pose a severe health threat in Bangladesh, where the early skin manifestations of arsenicosis often go undiagnosed due to a lack of dermatologists, subjective clinical assessments and limited healthcare facilities in rural regions. To address this emergency diagnostic gap, a dual-teacher knowledge distillation strategy, in which two complementary high-capacity teacher models, Xception+InceptionM and InceptionV3, transfer their multi-scale contextual knowledge to a compact student network, IR-KANNet, is proposed in this study. The proposed model, IR-KANNet, demonstrates strong overall performance, achieving stable and balanced precision, recall and F1-scores that align closely with or exceed the teacher models in several dimensions, confirming the effectiveness of distillation in capturing discriminative lesion features with far fewer parameters through stratified splitting and 5-fold cross-validation on the Arsenic-SkinImageBD dataset. Together with these promising results, the interpretability provided by Grad-CAM++ and LIME, which highlight clinically meaningful lesion regions, shows that the model is both reliable and suitable for practical use in clinical settings. Socially, this framework has the potential to support early screening in underserved communities, reduce diagnostic workloads, enhance accessibility to dermatological assessments in low-resource areas and pave the way for scalable, interpretable AI-based diagnostic systems for other skin diseases.

5.2 Implication

The need for lightweight deep learning models stems from the complicated issue of performing high-capacity deep learning in real-world resource-constrained environments, such as rural healthcare systems. Traditional networks, such as Xception or ResNet152V2, while accurate, have excessive memory, computational and inference sine requirements, which makes them unfeasible for on-device medical screening. In contrast, lightweight models are a scalable and efficient alternative that delivers competitive performance in terms of accuracy with significantly fewer parameters and lower latency. The proposed student model is designed along these lines, achieving a 78.39% reduction in parameters over teacher 1 and a 35.29% reduction over teacher 2, while maintaining high diagnostic reliability (accuracy 0.9692 and F1-score 0.9737). Its small size makes it both feasible to deploy and efficient in its energy consumption, making it a plausible candidate for real-time skin disease detection and medical diagnosis.

5.2.1 Why We Need Lightweight Models

Although very deep learning networks (for example, Xception, InceptionV3, or ResNet152V2) show very good accuracy for classifying medical images, their large number of parameters (from 20 million to 80 million) and huge computational needs restrict their usefulness in low-resource environments (such as arsenic-affected rural areas) where a healthcare image should be processed. In such environments, limited hardware, unstable connectivity and a lack of GPU infrastructure impose certain requirements on these models, such as computational efficiency and diagnostic reliability. The proposed lightweight student model solves these issues directly by having four important attributes:

1. **Efficiency**
   - The model contains only 0.58 million parameters, representing an approximate 97% reduction compared to the high-capacity teacher networks.
   - Each training epoch takes less than 10 seconds, verifying the model's fast convergence and minimal computational overhead.
   - Such compactness guarantees rapid inference and reduced memory use hence making it possible to screen the disease in real-time on a portable and embedded devices.
2. **Scalability**
   - The lightweight architecture is extremely versatile and can be easily reconfigured to do other dermatological or image based diagnosis tasks without redefining the extensive retraining.
   - Researchers can fine-tune the same model for multiple diseases using relatively small datasets, reducing dependency on large-scale annotated medical data.
   - With minimal resource requirements, various medical applications can efficiently reimplement this architecture, emphasizing its high scalability and reusability.

3. **Accessibility**
    - The model is designed specifically for low-resource environments, so it can be deployed in community clinics, rural hospitals, or mobile health units.
    - It can perform on-site inference without cloud dependency or high-speed internet, supporting offline diagnosis in remote arsenic-affected regions.
    - For low-power CPUs and embedded hardware (e.g., Raspberry Pi, Jetson Nano) the compact framework is compatible, owing to its small memory footprint.
4. **Interpretability**
    - The proposed lightweight model integrates explainable AI (XAI) tools such as Grad-CAM++ and LIME, which promotes the increase in transparency of the decision-making process.
    - Clinicians in understanding and validating predictions by highlighting skin regions that most influence classification is mapped by these visual explanation.
    - This interpretability will decrease the reliance on complex visualization servers and result in clinical confidence and should enhance diagnostic responsibility.

5.2.2 Potential Application Scenarios

The lightweight model and successful explainability of the proposed model demonstrate promising prospects for cross-domain use in situations where high computational efficiency and low latency are required and on-device decision-making is necessary. In addition to the detection of arsenic skin diseases, owing to its adaptability, it can be utilized in various fields, such as mobile health screenings, industrial automation, smart environments and personalized edge systems utilizing AI.

1. **Field & Mobile Applications**
    - **Mobile Document/ID Scanning:** Lightweight models can perform on-device optical character recognition (OCR), barcode, or MRZ detection for digital identification in logistics and financial know your customer (KYC) verification. Operating completely offline, these systems ensure data privacy and low latency, making them ideal for remote fieldwork or areas without Internet access. The compactness of the proposed model allows deployment even on entry-level Android devices without server-side computations.
    - **On-Device Voice/Keyword Spotting:** The small size of the model is helpful in supporting low-power wake-word recognition or command recognition of wearables, IoT devices and smart appliances. The sub-100 KB variants are capable of tracking voice clues continuously while consuming minimal energy; therefore, always-on interaction may be achieved within resource-constrained embedded applications such as hearing aids, smart speakers, or portable diagnostic devices.

- **Privacy-Preserving Presence/Counting:** In safety and compliance applications, the model can count individuals or vehicles without capturing and storing personally identifiable information. This will enable the locked-out occupancy in schools, offices and factories to maintain safety standards while keeping GDPR and HIPAA privacy regulations.
- **Vehicle Tracking & Traffic Analytics:** Condensed versions of the model with YOLO-N or Tiny-SORT trackers may allow on-device traffic monitoring, computing cars, estimating speeds and detecting violations on equipment such as Jetson Nano or Raspberry Pi. These systems facilitate smart surveillance without using cloud infrastructure at a cost-effective rate.
- **Agriculture:** The model can be modified to work in field conditions, where it can identify leaf diseases, pest infestations, or even crop maturity using a mobile phone or drone. Offline, it helps accuracy farming in areas with low connectivity and enables farmers to make real-time decisions to optimize yield and pesticide use.
- **Disaster & Infrastructure Monitoring:** With UAV or smartphone imagery, the lightweight CNN model can identify cracks, spalling, or debris on bridges, roads and buildings. Edge-based inference would help avoid large data uploads and allow real-time alerts to be issued to deal with disaster management and infrastructure maintenance.

2. **Medical & Healthcare Applications**

- **Point-of-Care Screening & Triage:** Lightweight CNN models deployed on smartphones or tablets can assist frontline health workers in screening skin lesions, respiratory symptoms and basic vital signs at rural clinics. Offline inference enables immediate triage decisions in arsenic-affected and other underserved regions without requiring stable internet or hospital-grade infrastructure.
- **Tele-Dermatology & Remote Consultation:** On-device image analysis can pre-filter and prioritize dermatological cases (e.g., arsenic skin lesions, eczema and psoriasis) before they are sent to specialists. Only compact feature summaries or risk scores need to be uploaded, reducing bandwidth and protecting patient privacy while accelerating remote expert reviews.
- **Wearable & Home Monitoring:** Ultra-lightweight models running on wearables or low-power medical devices can continuously analyze ECG, PPG, gait, or motion signals to detect early signs of deterioration, abnormal events, or non-adherence. Local processing minimizes energy consumption and avoids continuous streaming of raw biomedical signals.
- **Privacy-Preserving Clinical Decision Support:** Edge models can support tasks such as basic radiograph pre-screening, wound severity estimation and medication intake monitoring directly on hospital workstations or secure tablets. Since raw images and videos remain on-device and only decisions or alerts are logged, such systems align with stringent medical data protection and ethical guidelines.

3. **Industry & Retail Applications**

    - **Visual QA & Anomaly Detection:** The model can identify defects in low-shot or class situations on a manufacturing line in real time. When a per-station camera system is used without video streaming, a large fraction of the bandwidth used is saved. This, of course, facilitates automated quality control that is controlled directly by embedded controllers, which enhances efficiency and scalability in an industrial environment.
    - **Smart Shelves & Inventory Management:** In supermarkets and other retail outlets, edge-based vision models are able to detect stockouts, lost items or empty shelf space. The system does not require full transmission of video data; instead, it sends only updates that are triggered by events, which reduces the amount of network consumption. This will achieve precision and current inventory monitoring and minimize bandwidth and storage needs.
    - **Predictive Maintenance (Vibration/Audio):** Minimal CNN versions are able to process vibration and other acoustic signals on motors, fans, or bearings to detect early faults or imbalances. These miniature models are highly functional with microcontrollers (MCUs) that can be monitored on-site without the use of external servers. This aids Industry 4.0 efforts in offering real-time predictive machine maintenance, production accuracy and machine longevity.

4. **Education & Training Applications**

    - **Classroom Analytics (Local/Private):** Even the analysis of the level of attention, occupancy and board utilization can be done locally and information on students can be obtained without violating their privacy using lightweight edge models. These systems can offer real-time actionable feedback to teachers without any identifiable data to enable better classroom activity and learning analytics.
    - **Lab & Maker-Space Safety:** In industrial or technical training, edge cameras can identify the lack of protective gear, unsafe tool handling, or unsafe postures and attempt to prevent accidents. These systems enhance compliance with safety and the identification of hazards at the initial stages of laboratory lessons or skill-based workshops.
    - **AR/VR Perception for Lessons:** Light Neural networks lightweight 3D perception networks can boost the tracking of objects, detection of markers and hand-pose estimation of educational AR/VR systems. Such attributes make it possible to have an immersive, interactive learning experience without expensive hardware requirements, with the additional capability to be implemented inexpensively in schools and training institutions.
    - **Ethical Local Proctoring (Lite):** Presence and gaze detectors can be used to realize a privacy-compliant remote assessment by stabilizing exam integrity without transmitting biometric information over the video or recording the live video. This solution provides a more ethical option

than conventional online proctoring solutions and ensures the privacy and integrity of students.

5. **Robotics, Drones & Smart Spaces**

   - **Robot/Drone Perception:** In autonomous robots and drones, obstacle avoidance and waypoint guidance can be performed using compact perception models. These models can run effectively on single-board computers (SBCs) or microcontroller units (MCUs) with real-time inference rates of a frame rate of less than 50 ms. This and similar capabilities allow them to be agile and respond to control in real time, which makes the models convenient for embedded robotic tasks and air navigation missions.
   - **Energy & Building Management:** The model can be used to optimize HVAC and lighting systems by locating appliance activity and human presence to ensure ad hoc remote control and smart infrastructure management. This ensures great energy effectiveness and is still comfortable to the user with not a single raw sensory data transmitting to the outside servers; hence, privacy as well as bandwidth.

6. **Privacy, Personalization & Federated Systems**

   - **On-Device Personalization:** Few-shot adaptive learning allows the model to change to conditions specific to the device, such as lighting, camera quality, or a person's accent and does not require retraining using the cloud. This strategy maintains the privacy of data but enhances user experience and responsiveness to the local devices so that the personalization can be attained without the loss of privacy or the augmentation of the computational burden.
   - **Federated Analytics:** Lightweight models may also be used in federated learning systems, in which only the gradient updates are exchanged among the devices but not the raw data. This distributed learning system protects the privacy of users and achieves better collective model performance. This method is especially useful in sensitive areas of human endeavor, such as healthcare and education, where professional ethical and regulatory standards on device data protection are crucial.

5.3 Limitations

This section illustrates the key constraints encountered when developing, training and assessing the suggested explainable deep learning architecture to detect arsenic-induced skin disease. Despite the promising results of the dual-teacher knowledge distillation (KD) methodology and lightweight and interpretable student networks, several limitations are associated with datasets, modeling and their deployment. The chapter then ends by providing suggestions for future studies to help increase the variety of datasets, enhance the efficiency of the algorithms, enhance explainability and enable clinical implementation into practice.

5.3.1 Dataset-Related Limitations

The specific limitations associated with the dataset include the following:

- **Dataset Size Diversity:** The available dataset was obtained from a small geographical area. This limits the variation in skin color, skin lesion form and light variations, which can limit the extent of generalization between different types.
- **Class Imbalance:** Although augmentation and class weighting were used, there was still a slight imbalance between the "infected" and the "not-infected" categories, which may be slightly biased in favor of the majority category.
- **Lack of Multi-Disease Representation:** The dataset focused exclusively on arsenic-induced skin lesions. In practice, visually similar conditions (e.g., eczema, psoriasis and fungal infections) frequently occur in clinical settings. The absence of such confounding diseases in the training data limits the practical applicability and robustness of the model when it is deployed in real-world scenarios.

5.3.2 Model-Related Limitations

- **Knowledge Distillation Dependence:** The performance of the student model is inherently dependent on the quality and discriminative strength of the teacher network. Any sub-optimal learning or bias present in the teachers can be propagated through the distillation process, potentially transferring noise and systematic errors to the student models.
- **Limited Task Scope:** The classification that is in effect in the framework is binary classification (infected vs. not-infected), but not yet multi-class or lesion-type-specific.
- **Explainability Resolution:** Grad-CAM++ and LIME make useful visual explanations but are not fully faithful descriptions of model reasoning because they are post-hoc approximation algorithms instead of descriptive algorithms. Distribution-level interpretability is not yet in place.
- **No Temporal Learning:** The treatment of each image was not followed by time or longitudinal changes in the skin lesions over a period, which would have added more background regarding the diagnosis.

5.3.3 Computational and Practical Constraint

- **Limited Hardware Validation:** The lightweight version was investigated only on systems with a GPU and CPU; it has not yet been tested on embedded or mobile systems (e.g., Jetson Nano, Raspberry Pi, smartphones).
- **Lack of Clinical Integration:** Although the diagnostic performance is encouraging, it has not been used in real-time deployment in healthcare settings and field environments.

- **Explainability Evaluation:** The existing measures of interpretability are qualitative. Clinical reliability would be better proven by quantitative assessment based on dermatologist-marked areas.

5.4 Future Works

Based on the strengths and weaknesses of this research, it is possible to explore several potential opportunities to move forward and evolve explainable AI systems in improving the detection and prognosis of arsenic-induced skin disease. Future research should focus on four areas of interest: augmentation and annotation of datasets, model and algorithm development, implementation and clinical translation and generalization to domains.

5.4.1 Dataset Expansion and Annotation

The development of a larger and more heterogeneous collection of arsenic-contaminated skin images is one of the largest areas of enhancement in causing enhancements in generalization.

- In future studies, it would be appropriate to have cross-section of different geographical areas in different ethnicities and environments to determine the greater variation of lesion types, skin color and imaging conditions.
- In addition, the model would be capable of fine-grained classification as opposed to binary diagnosis, which would be facilitated by multi-class labeling, which would group cases into mild, moderate and severe cases of damage caused by arsenic.
- Besides this, one can incorporate additional auxiliary data like patient age, gender, length of exposure and concentration of arsenic in water, which can facilitate multi-modal learning where the model can utilize visual as well as contextual data.
- Last but not least, partnership with dermatologists to design lesion boundaries that are annotated by the experts would improve quality training and interpretability benchmarks to a greater extent.

5.4.2 Model and Algorithmic Improvements

- **Multi-Stage Knowledge Distillation:** The proposed model as advanced in future studies should follow the notion of multi-stage and self-distillation strategies, in which the student undergoes self-refinement of knowledge and will become a teacher to itself. This practice enhances self-sufficiency and decreases reliance on large pre-trained networks.
- **Causal and Faithful Explainability:** To make the explainability even more faithful and causally rooted, one can employ more sophisticated interpretability techniques, such as integrated gradients, SHAP and counterfactual explanations, which can provide information about feature importance and causal attribution.

5.4.3 Deployment and Clinical Translation

The next step is to translate the outcomes of the research into real-life healthcare practices.

- **Edge-AI Deployment:** The student model can be configured to run on edge-AI devices, such as NVIDIA Jetson Nano, Raspberry Pi, or Android-based handhelds, to test them for power consumption, real-time inference and field conditions.
- **Telemedicine Integration:** It can be integrated with telemedicine systems and cloud-based diagnostic platforms to support the remote screening of arsenicosis and decision-making, particularly in rural communities with poor services.
- **Clinical Collaboration:** The collaboration with dermatologists and public health agencies should be also validated to check the clinical performance and make sure that AI-based predictions do not contradict the medical reasoning.
- **Regulatory Compliance and Ethics:** It would be more beneficial to make the system more interpretable and actionable for clinicians by developing an interactive AI-assisted diagnostic dashboard that incorporates prediction confidence, heatmaps and textual guidance.

Bibliography

[1] W. H. Organization et al., “Exposure to arsenic: A major public health concern. 2010,” *WHO Document Production Service: Geneva, Switzerland*, 2020.

[2] M. R. Karagas, A. Gossai, B. Pierce, and H. Ahsan, “Drinking water arsenic contamination, skin lesions, and malignancies: A systematic review of the global evidence,” *Current environmental health reports*, vol. 2, no. 1, pp. 52–68, 2015.

[3] G. T. Tang, J. Elakis, and L. Scardamaglia, “Cutaneous manifestations and treatment of arsenic toxicity: A systematic review,” *Skin Health and Disease*, vol. 3, no. 4, ski2–231, 2023.

[4] U. Chowdhury, “Arsenic health impacts to the people in bangladesh,” *International Journal of Toxicology and Toxicity Assessment*, vol. 2, Oct. 2023. DOI: 10.55124/IJT.V2I1.234.

[5] S. P. Choy et al., “Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease,” *NPJ Digital Medicine*, vol. 6, no. 1, p. 180, 2023.

[6] M. H. K. Mehedi, K. F. Z. Nafis, K. H. Charu, J. Uddin, M. G. R. Alam, and M. Mridha, “Arsenicnet: An efficient way of arsenic skin disease detection using enriched fusion xception model,” *PLoS One*, vol. 20, no. 5, e0322405, 2025.

[7] A. Newaz, A. U. R. Adib, R. Sahil, and M. Mehzad, “An end-to-end deep learning framework for arsenicosis diagnosis using mobile-captured skin images,” *arXiv preprint arXiv:2509.08780*, 2025.

[8] S. Ahmed, M. Faruquee, and M. H. Khan, “Twenty years of arsenic contamination and arsenicosis patients in a village of bangladesh,” *American Journal of Public Health Research*, vol. 8, no. 6, pp. 190–196, 2020.

[9] S. V. Rajiv, M. George, and G. Nandakumar, “Dermatological manifestations of arsenic exposure,” *Journal of Skin and Sexually Transmitted Diseases*, vol. 5, no. 1, pp. 14–21, 2023.

[10] M. I. M. Choudhury et al., “Cutaneous malignancy due to arsenicosis in bangladesh: 12-year study in tertiary level hospital,” *BioMed research international*, vol. 2018, no. 1, p. 4 678 362, 2018.

[11] G. L. Porras and P. Allard, “Making waves: Public health risks from arsenic in mexico’s water extraction practices,” *Water Research X*, vol. 26, p. 100 283, 2025.

[12] S. H. Jobayer et al., "Deep learning for identification of skin diseases: Emphasizing on arsenic-induced skin conditions," in *2024 27th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2024, pp. 3254–3259.

[13] A. Saha, U. Das, and M. S. Rahman, "A deep learning approach to automate classification of arsenic-affected skin using efficientnet-b1," in *2024 27th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2024, pp. 1398–1403.

[14] I. A. Emu et al., "Arsenicskinimagebd: A comprehensive image dataset to classify affected and healthy skin of arsenic-affected people," *Data in Brief*, vol. 52, p. 110 016, 2024.

[15] K. M. Almustafa, "Predictive modeling and optimization in dermatology: Machine learning for skin disease classification," *Computers in Biology and Medicine*, vol. 189, p. 109 946, 2025.

[16] G. M. Setegn and B. E. Dejene, "Explainable ai for symptom-based detection of monkeypox: A machine learning approach," *BMC Infectious Diseases*, vol. 25, no. 1, p. 419, 2025.

[17] A. Balasundaram, A. Shaik, B. R. Alroy, A. Singh, and S. Shivaprakash, "Genetic algorithm optimized stacking approach to skin disease detection," *IEEE Access*, vol. 12, pp. 88 950–88 962, 2024.

[18] S. K. Singh, R. W. Taylor, M. M. Rahman, and B. Pradhan, "Developing robust arsenic awareness prediction models using machine learning algorithms," *Journal of environmental management*, vol. 211, pp. 125–137, 2018.

[19] M. A. Lombard et al., "Machine learning models of arsenic in private wells throughout the conterminous united states as a tool for exposure assessment in human health studies," *Environmental science & technology*, vol. 55, no. 8, pp. 5012–5023, 2021.

[20] V. Balaji, S. Suganthi, R. Rajadevi, V. K. Kumar, B. S. Balaji, and S. Pandiyan, "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through naive bayes classifier," *Measurement*, vol. 163, p. 107 922, 2020.

[21] A. Souri, M. Y. Ghafour, A. M. Ahmed, F. Safara, A. Yamini, and M. Hoseyninezhad, "A new machine learning-based healthcare monitoring model for student's condition diagnosis in internet of things environment," *Soft Computing*, vol. 24, no. 22, pp. 17 111–17 121, 2020.

[22] T. Arravalli et al., "Detection of breast cancer using machine learning and explainable artificial intelligence," *Scientific Reports*, vol. 15, no. 1, p. 26 931, 2025.

[23] G. Dharmarathne, T. N. Jayasinghe, M. Bogahawaththa, D. Meddage, and U. Rathnayake, "A novel machine learning approach for diagnosing diabetes with a self-explainable interface," *Healthcare analytics*, vol. 5, p. 100 301, 2024.

[24] H. M. Rai and J. Yoo, "A comprehensive analysis of recent advancements in cancer detection using machine learning and deep learning models for improved diagnostics," *Journal of Cancer Research and Clinical Oncology*, vol. 149, no. 15, pp. 14 365–14 408, 2023.

[25] P. Gupta et al., “Detecting thyroid disease using optimized machine learning model based on differential evolution,” *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 3, 2024.

[26] M. Badr, A. Elkasaby, M. Alrahmawy, and S. El-Metwally, “A multi-model deep learning architecture for diagnosing multi-class skin diseases,” *Journal of Imaging Informatics in Medicine*, vol. 38, no. 3, pp. 1776–1795, 2025.

[27] A. Akram et al., “Skinmarknet: An automated approach for prediction of monkeypox using image data augmentation with deep ensemble learning models,” *Multimedia Tools and Applications*, vol. 84, no. 18, pp. 20 177–20 193, 2025.

[28] S. Dhanushraju, R. Hariharan, P. D. M. Bala, S. Saaral, and D. Vijayan, “A deep learning approach to multi-class skin disease identification,” in *2025 Eleventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII)*, IEEE, 2025, pp. 1–9.

[29] D. W. Girmaw, “Livestock animal skin disease detection and classification using deep learning approaches,” *Biomedical Signal Processing and Control*, vol. 102, p. 107 334, 2025.

[30] V. RADHIKA, A. MUTHUCHUDAR, and M. LINGARAJ, “Self-reliant residual network based deep learning framework for melanoma skin disease detection,” *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 10, 2025.

[31] Y.-C. Yang, W.-C. Chung, C.-Y. Wu, C.-L. Hung, and Y.-J. Chen, “Skin disease classification using transfer learning model and fusion strategy,” *Intelligence-Based Medicine*, p. 100 271, 2025.

[32] Y. Cao et al., “Robustly detecting mpox and non-mpox using a deep learning framework based on image inpainting,” *Scientific Reports*, vol. 15, no. 1, p. 1576, 2025.

[33] N. Ashfaq et al., “Skinsight: Advancing deep learning for skin cancer diagnosis and classification,” *Discover Computing*, vol. 28, no. 1, p. 63, 2025.

[34] A. Naeem, H. Malik, M.-z. Din, A. Sadeghi-Niaraki, D. Jeong, and R. A. Naqvi, “Skindwnet: A novel deep learning model for multiclass classification of skin cancers using dermoscopic images,” *Multimedia Systems*, vol. 31, no. 4, pp. 1–28, 2025.

[35] B. Ozdemir and I. Pacal, “A robust deep learning framework for multiclass skin cancer classification,” *Scientific Reports*, vol. 15, no. 1, p. 4938, 2025.

[36] M. Akter, R. Khatun, M. A. Talukder, M. M. Islam, and M. A. Uddin, “An integrated deep learning model for skin cancer detection using hybrid feature fusion technique,” *arXiv preprint arXiv:2410.14489*, 2024.

[37] M. Shakya, R. Patel, and S. Joshi, “A comprehensive analysis of deep learning and transfer learning techniques for skin cancer classification,” *Scientific Reports*, vol. 15, no. 1, p. 4633, 2025.

[38] M. Mateen, S. Hayat, F. Arshad, Y.-H. Gu, and M. A. Al-Antari, “Hybrid deep learning framework for melanoma diagnosis using dermoscopic medical images,” *Diagnostics*, vol. 14, no. 19, p. 2242, 2024.

[39] S. Asif, S. U. R. Khan, K. Amjad, and M. Awais, “Skinc-net: An efficient lightweight deep learning model for multiclass skin lesion classification in dermoscopic images,” *Multimedia Tools and Applications*, vol. 84, no. 13, pp. 12 531–12 557, 2025.

[40] S. U. R. Khan, S. Asif, O. Bilal, and H. U. Rehman, “Lead-cnn: Lightweight enhanced dimension reduction convolutional neural network for brain tumor classification,” *International Journal of Machine Learning and Cybernetics*, pp. 1–20, 2025.

[41] M. G. S. Hossen and M. N. I. Mondal, “Efficient and scalable arsenicosis detection using lightweight cnns and transformer-based models for resource-limited settings,” in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, 2025, pp. 1–6.

[42] D. Khasanov et al., “Deepbitenet: A lightweight ensemble framework for multiclass bug bite classification using image-based deep learning,” *Diagnostics*, vol. 15, no. 15, p. 1841, 2025.

[43] D. H. Elsheikhy, A. S. Hassan, N. M. Yhiea, A. M. Fareed, and E. A. Rashed, “Lightweight deep learning architecture for multi-lead ecg arrhythmia detection,” *Sensors*, vol. 25, no. 17, p. 5542, 2025.

[44] A. M. Almars, “Deepgenmon: A novel framework for monkeypox classification integrating lightweight attention-based deep learning and a genetic algorithm,” *Diagnostics*, vol. 15, no. 2, p. 130, 2025.

[45] M. D. Nayeem, M. A. Rahman, M. S. Hossain, M. Safran, S. Alfarhood, and M. Mridha, “Multi-scale attention fusion with depthwise separable convolutions for efficient skin cancer detection,” *Journal of Cutaneous Pathology*, 2025.

[46] H. Liu et al., “A skin disease classification model based on multi scale combined efficient channel attention module,” *Scientific Reports*, vol. 15, no. 1, p. 6116, 2025.

[47] M. A. K. Abir et al., “Attention enhanced inception-v3: A multi-scale feature fusion network for skin lesion detection with explainable artificial intelligence,” in *2024 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, IEEE, 2024, pp. 1–6.

[48] S. Saha, M. M. Hemal, M. Z. A. Eidmum, and M. F. Mridha, “Knowledge distillation approach for skin cancer classification on lightweight deep learning model,” *Healthcare Technology Letters*, vol. 12, no. 1, e12120, 2025.

[49] M. R. Kabir, R. H. Borshon, M. K. Wasi, R. M. Sultan, A. Hossain, and R. Khan, “Skin cancer detection using lightweight model souping and ensembling knowledge distillation for memory-constrained devices,” *Intelligence-Based Medicine*, vol. 10, p. 100 176, 2024.

[50] N. Islam, K. M. Hasib, F. A. Joti, A. Karim, and S. Azam, “Leveraging knowledge distillation for lightweight skin cancer classification: Balancing accuracy and computational efficiency,” *arXiv preprint arXiv:2406.17051*, 2024.

[51] M. A. Asham, A. A. Al-Shargabi, R. Al-Sabri, and I. Meftah, “A lightweight deep learning model with knowledge distillation for pulmonary diseases detection in chest x-rays,” *Multimedia Tools And Applications*, vol. 84, no. 15, pp. 14 885–14 913, 2025.

[52] K. Shi et al., "Enhancing bone mineral density estimation from x-ray images with cross-modal knowledge distillation," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 4861–4871.

[53] P. Thaker and A. Asif, "Knowledge distillation for enabling efficient ai-based skin cancer detection in resource-constrained environments," in *International Conference on Medical Imaging and Computer-Aided Diagnosis*, Springer, 2024, pp. 77–84.

[54] L. Gamage, U. Isuranga, D. Meedeniya, S. De Silva, and P. Yogarajah, "Melanoma skin cancer identification with explainability utilizing mask guided technique," *Electronics*, vol. 13, no. 4, p. 680, 2024.

[55] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, "A deep learning approach based on explainable artificial intelligence for skin lesion classification," *IEEE Access*, vol. 10, pp. 113 715–113 725, 2022.

[56] A. B. Vaghela, N. K. Patel, R. K. Panchal, K. H. Bamniya, and H. R. Patel, "Xai-skin: An approaches for the diagnosis and classification of skin diseases based on lime method," *International Journal of Environmental Sciences*, vol. 11, no. 3s, pp. 350–363, 2025.

[57] S. Y. Hong and L. P. Lin, "Skin lesion classification: A deep learning approach with local interpretable model-agnostic explanations (lime) for explainable artificial intelligence (xai)," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 3-2, pp. 1536–1544, 2024.

[58] A. S. N. Raju, G. Sujatha, R. K. Gatla, S. Ankalaki, and G. S. Hukkeri, "Xai-skincadx: A six-stage explainable deep ensemble framework for skin cancer diagnosis and risk-based clinical recommendations," *IEEE Access*, 2025.

[59] M. Fiaz et al., "Correction: An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification," *Frontiers in Medicine*, vol. 12, p. 1 724 427, 2025.

[60] D. Kumar Saha et al., "Mpox-xde: An ensemble model utilizing deep cnn and explainable ai for monkeypox detection and classification," *BMC Infectious Diseases*, vol. 25, no. 1, p. 403, 2025.

[61] S. Biswas, S. Saha, M. S. Uddin, and R. Mostafiz, "An explainable and federated deep learning framework for skin cancer diagnosis," *PLoS One*, vol. 20, no. 7, e0324393, 2025.

[62] M. Gupta, M. Kumar, and R. Dhir, "Fedmed-xai: A collaborative and trustworthy framework for skin cancer detection using federated learning and explainable ai," *The Journal of Supercomputing*, vol. 81, no. 15, pp. 1–44, 2025.

[63] A. Alrabai, A. Echtioui, and F. Kallel, "Explainable deep learning approaches for skin cancer diagnosis," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 14, no. 1, p. 57, 2025.

[64] A. B. George, S. Bathini, et al., "Grad-cam & grad-cam++ for explainable oral squamous cell carcinoma detection using deep learning on orthopantomograms," in *2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare (64220)*, IEEE, 2025, pp. 1–6.

[65] M. R. Zafar and N. Khan, “Deterministic local interpretable model-agnostic explanations for stable explainability,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.

[66] E. H. Houssein, A. M. Gamal, E. M. Younis, and E. Mohamed, “Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review,” *Cluster Computing*, vol. 28, no. 7, p. 469, 2025.