Multi-Dataset Plantar Thermography Integration for Scalable Deep Learning-Based Early Detection of Diabetic Foot Ulcers

by

Syed Mohammad Jarif Tahmid
22101286
Wais Kafia Chowdhury
22101251
Nowmi Islam
22101581
Zajaul Ehsan Sajid
24341195
Mahmud Hasan Foysal
22101039

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

Department of Computer Science and Engineering
Brac University
Month 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/our 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 does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Students Full Name & Signature:

Syed Mohammad Jarif Tahmid

22101286

Wais Kafia Chowdhury

22101251

Nowmi Islam

22101581

Zajaul Ehsan Sajid

24341195

Mahmud Hasan Foysal

22101039

Approval

The thesis/project titled Multi-Dataset Plantar Thermography Integration for Scalable Deep Learning-Based Early Detection of Diabetic Foot Ulcers submitted by

1. Syed Mohammad Jarif Tahmid (22101286)
2. Wais Kafia Chowdhury (22101251)
3. Zajaul Ehsan Sajid (24341195)
4. Nowmi Islam (22101581)
5. Mahmud Hasan Foysal (22101039)

of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Summer, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Chowdhury Mofizur Rahman

Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Diabetic foot is a severe complication of diabetes, which can lead to major health complications such as ulceration and amputation if left untreated. While deep learning offers promising avenues for diagnosis, existing studies are often constrained by small, homogeneous datasets, high computational costs, and a tendency to overfit to single data sources. To address these limitations, this study develops a robust classification pipeline that integrates multiple plantar thermography datasets to improve model generalizability and accessibility. The pipeline focuses on systematic data integration, preprocessing, and validation, training a suite of lightweight convolutional neural networks (CNNs) with transfer learning on standardized and combined thermograms. Model explainability was verified using Grad-CAM visualizations, which confirmed that predictions were based on physiologically relevant plantar regions rather than background artifacts. The best-performing architecture in this pipeline, GhostNet_100, outperformed previous methods, achieving state-of-the-art results with mean values of 96.0% accuracy, 97.1% precision, 97.9% recall, 97.5% F1-score, and 94.3% specificity on the combined datasets. By unifying diverse datasets and deploying efficient CNNs within a reproducible workflow, this work establishes a scalable, cost-effective pipeline for early diabetic foot screening, demonstrating strong potential for practical deployment in low-resource healthcare environments.

Keywords: Artificial Intelligence, Skin Diseases, Early Detection, Diabetic Foot Ulcer, Convolutional Neural Networks, Grad-CAM, Thermograms, Dermatology, Explainable AI.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Diabetes mellitus is a metabolic disorder which causes hyperglycemia (increased blood sugar) due to insulin deficiency or resistance. It can lead to long-term health issues, affecting nerves, blood vessels, and the immune system. Primarily, two types of diabetes occur in patients: Type 1 Diabetes (T1DM) which causes autoimmune destruction of pancreatic beta-cells and leads to insulin deficiency; and Type 2 Diabetes (T2DM) causes insulin resistance that leads to the dysfunction of beta-cells [1].

Diabetes can lead to multiple diseases such as neuropathy (nerve damage), vasculopathy (blood vessel damage) and ischaemia (reduced blood flow and oxygenation). Severe hyperglycemia causes nerve fibers to get damaged, resulting in the loss of the protective response of the body. Peripheral neuropathy numbs or eliminates pain experienced during injuries, while autonomic neuropathy causes dry skin and cracks, both of which consequently increase the risks of infection as the body fails to perceive and heal injuries [2]. Diabetic vasculopathy causes both macrovascular damage, such as peripheral arterial disease (PAD), which reduces blood supply (ischaemia), and microvascular damage, where thickened capillary membranes reduce oxygen and nutrient delivery. As a result, wounds heal poorly, increasing the risk of ulceration [3]. Additionally, chronic ischaemia caused by PAD causes hypoxia and necrosis (death of tissues). This delays the healing of injuries, allowing minor injuries to progress into ulcers or even gangrene [4].

Diabetic foot refers to the range of health complications of the foot that arise in people with diabetes due to the loss of protective sensation and ischaemia. The primary result of diabetic foot is often recurrent ulceration that is caused by the combination of neuropathy and vasculopathy. 50-75 % of the ulcerations are caused by peripheral neuropathy [5]. Poor blood sugar control leads to deformities in pressure points of the foot, while vasculopathy and neuropathy can cause these deformities to turn into severe injuries [6].

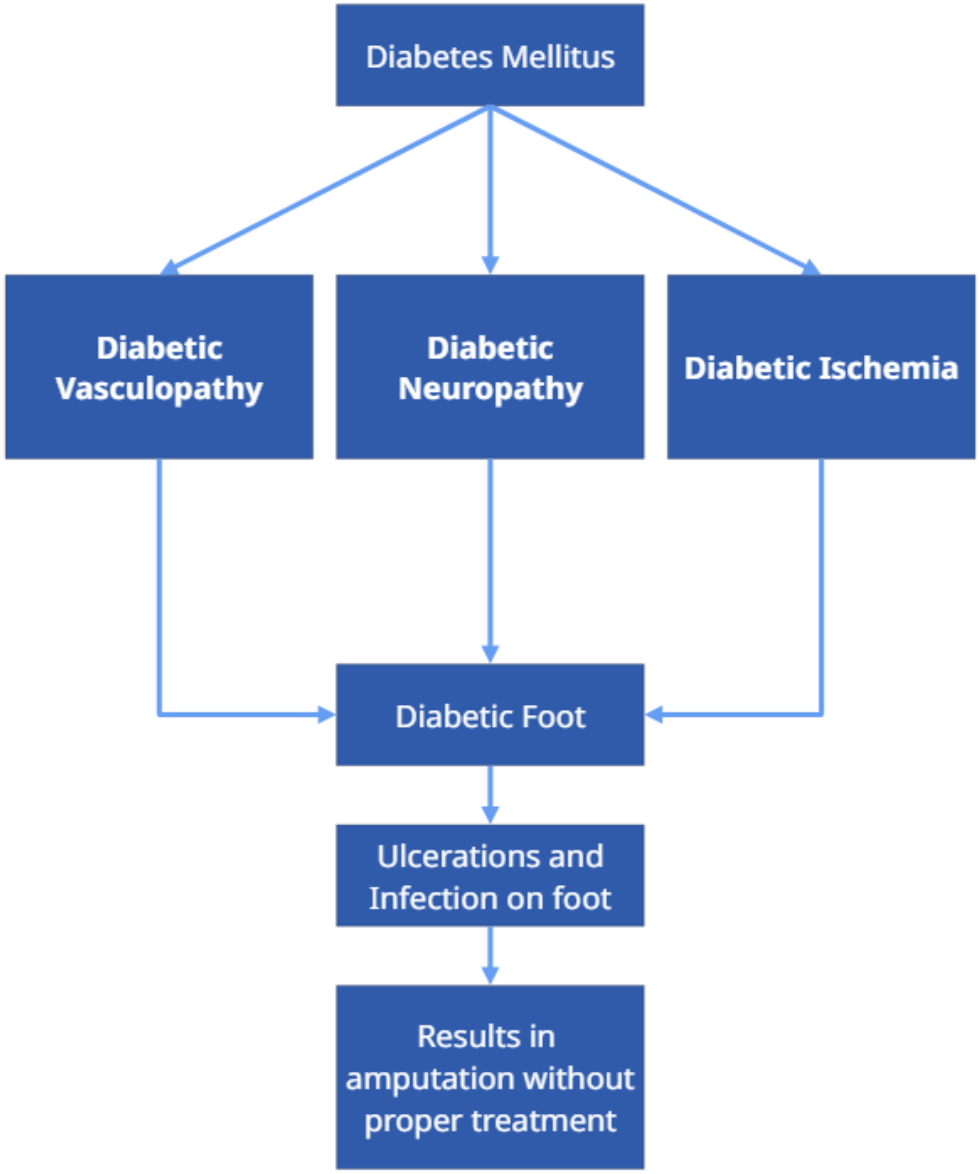

Figure 1.1: Schematic progression of diabetic foot from neuropathy to ulceration and amputation

1.2 Motivation of the Study

As diabetic foot often leads to infections, ulcerations and eventualy, lower-limb amputations [7], early detection of diabetic foot is crucial as it prevents a long list of severe complications. Early detection reduces the risk of ulceration in patients [8], which is particularly important as 85% of diabetes-related amputations start with an ulcer [5]. Notably, 77% of diabetes-related ulcers, and consequently amputations, are preventable within a year if diabetic foot is diagnosed in the initial stages [6], further emphasizing the need for early detection. Early detection of diabetic foot also has the potential to significantly lower health care costs by preventing the disease from reaching deadly stages which often push the burden of frequent hospital visits, costly procedures and medication onto the patient. Plantar thermal imaging (thermography) offers a non-invasive approach to aiding the detection of diabetic foot early by detecting irregular heat patterns formed by the damaging of nerves in the foot [9]. Plantar thermography can detect temperature asymmetries which indicate inflammation or compromised blood flow by analyzing the butterfly pattern, a regular heat pattern typically found in the feet of healthy individuals, as opposed to the irregular patterns found in people with diabetic complications [10]. These irregular patterns are early identifiers of neuropathy and ischaemia and allow timely interventions and prevention of ulcers and infection [9]. AI-based thermography and neural networks have been verified to assist in real-time risk prediction of diabetic foot, preventing severe complications [11]. Although a viable option and a valid aid for the detection of diabetic foot, the current methods of thermography can be quite expensive. Plantar thermography often uses expensive medical equipment, which cost several thousand euros [12]. This poses a significant challenge in accessibility, specially for low-budget clinics in underdeveloped states and rural regions. Thus, taking into account the high costs of purchase, certification, and operation of thermography equipment, our study is motivated by the need to develop an early detection system for diabetic foot complications that is readily available, accessible, and cost-effect

1.3 Problem Statement

A key limitation in existing research of thermogram-classification of diabetic foot is the lack of available datasets for training. The majority of the existing researches utilize the IEEE thermogram dataset [13], which contains plantar thermograms captured by a FLIR E60 camera [14]. The equipment is considerably expensive with a starting price of EUR 6000. Along with the expensive purchase cost, the maintenance and operating costs need to be considered as well.

The IEEE thermogram dataset also provides a small and very limited pool of only 167 individuals, 122 of whom are diagnosed with diabetes and 45 non-diabetic individuals being the control group [13]. Datasets with a limited pool such as this are likely to produce results that lack reliability and generalizability. Hence, studies utilizing this dataset alone have a very high probability of producing overfitted results. These concerns of overfitting are further solidified due to the lack of verification utilities such as Grad-CAM in existing research.

In contrast, this study utilizes the STANDUP database which consists of thermograms captured with cost effective equipment [15] alongside the IEEE thermogram dataset. The STANDUP dataset contains thermal images captured using the FLIR ONE Pro camera, which is a smartphone-compatible camera priced at approximately EUR 350. The dataset contains thermal images of 62 healthy individuals as the control group and 120 diabetic individuals, with two images from each individual. While this dataset has been applied in segmentation algorithms, it is not widely utilized for classification purposes. Utilizing this dataset for classification in parallel with the IEEE thermogram dataset will help us develop models that are more adaptable and scalable in real-world, low-resource scenarios. It builds upon the IEEE dataset hence, the study has a larger and more diverse training data, helping to reduce overfitting issues and help create a more generalizable model.

1.4 Objective

This primary objective of this study is to develop and evaluate deep learning based framework for early-stage diabetic foot detection using plantar thermographic images. Specifically, the study aims to:

- Address the limitations of single-dataset evaluation and of using non-harmonized datasets by integrating and standardizing the IEEE and STANDUP thermogram datasets to improve cross-dataset generalizability.
- Improve robustness across heterogeneous imaging devices through a colormap harmonization pipeline.
- Incorporate model explainability by applying Grad-CAM visualizations to verify that the models focus on physiologically relevant plantar regions.
- Prevent data leakage and reduce overfitting by implementing patient-wise (ID-based) data splitting.
- Benchmark lightweight CNN architectures suitable for real-time inference on cost-effective thermal imaging devices.
- Identify an efficient and deployable deep learning model for early diabetic foot detection in low-resource clinical settings.

1.5 Methodology in Brief

This study will utilize two primary datasets for analysis - The IEEE thermogram dataset and the STANDUP dataset. Both of the datasets will undergo preprocessing which includes normalization, resizing and data augmentation techniques such as flipping. This will be done in order to enhance model performance and generalization.

Convolutional Neural Networks (CNNs) are the core deep learning technique that will be utilized in this study. CNNs are well-established in classification tasks of medical imaging and perform well due to their ability to learn spatial patterns from

features. CNNs have also been used in classification of plantar thermographic data to detect diabetic foot complications and have shown promising results [11].

Since there is a difference in the colormaps of two datasets (the IEEE dataset providing multi-colored thermograms while the STANDUP dataset providing white-hot thermograms), the IEEE dataset was preprocessed with a conversion pipeline to generalize the data. A combination of both datasets was utilized to train and evaluate the models. The data was split into a 80:20 ratio, reserving 80% for training, 20% for testing. From the 80% training split, 20% was be reserved for a validation split. Standard evaluation metrics (accuracy, precision, recall, specificity, AUC and F1-score) was be used to evaluate the results. Cross-validation studies and XAI such as Grad-CAM was also be used to assess the robustness and generalizability of the pipeline to help prevent overfitting issues and produce reliable results.

1.6 Scopes and Challenges

This study focuses on the binary classification of plantar thermograms into healthy and diabetic foot categories. Intermediate conditions such as pre-diabetes or severity of diabetic foot complication are not addressed through this study. The scope of this study is limited to identifying irregularities in the heat patterns in plantar themograms that indicate the presence or formation of diabetic foot without further subclassification of specific complications (e.g neuropathy vs. ischaemia)

A key challenge in this study was the generalizability of the two datasets. The IEEE thermogram dataset provides high-resolution, multicolor thermograms generated by FLIR A300 thermal cameras whereas the STANDUP dataset offers grayscale thermal images captured using a more cost-effective FLIR ONE Pro camera. This difference in image type can make preprocessing, standardization, and model training challenging, especially when combining both datasets. To counter this challenge, we implemented a colormap conversion pipeline to convert the Rainbow High Contrast Colormap of the IEEE dataset into white-hot colormap like the Standup dataset while retaining thermal information.

Another challenge was the limited sample size. Although combining both datasets increases the overall training pool, the data still remains relatively small by deep learning standards, which can cause concerns of overfitting and reduced generalizability. To counter the lack of data, extensive augmentation techniques were applied to the datasets, which doubled the pool of data whilst retaining an uniform class distribution. To counter overfitting concerns, we utilized a carefully crafted data separation pipeline to ensure no data leakage occurs in any splits. We further utilize Grad-CAMs to verify the models focus on plantar regions and not background noise.

The study addresses these challenges and aims to demonstrate the feasibility of using cost-effective thermal imaging and deep learning for early diabetic foot detection in environments with limited resources.

Chapter 2

Literature Review

2.1 Preliminaries

2.1.1 Infrared Thermography (IRT)

Infrared thermography (IRT) captures thermal radiation emitted from an object's surface to determine its temperature. Human skin has an emissivity of 0.98, enabling calibrated thermal cameras to accurately estimate skin temperature. Modern thermal imaging produces two-dimensional thermograms where each pixel encodes temperature information, visualizing temperature distribution across the surface. Skin temperature patterns reveal physiological status and underlying abnormalities. Blood flow regulates body temperature, where ischaemic regions (poor blood flow) become cooler while inflamed areas create hotspots. Peripheral Arterial Disease and Diabetic Neuropathy disrupt normal thermal patterns, making abnormal hotspots or cold areas on plantar thermograms indicative of early diabetic foot complications [10].

2.1.2 IRT and diabetic foot

Skin temperature patterns can help determine physiological status of an individual and can also highlight any underlying disease or abnormalities. Normally the human body maintains a core temperature of approximately 37°C by the use of mechanisms like vasodilation (dilation of blood vessels), vasoconstriction (shrinking of blood vessels) and sweating. Blood flow determines the temperature of the body and ischaemic regions (regions with poor blood flow) become cooler while inflamed or infected regions result in hotspots. Peripheral Arterial Disease and Diabetic Neuropathy alter regular thermal patterns, and hence, in the context of diabetic foot, abnormal hotspots or cold areas on a plantar thermogram can hint at early complications of diabetic foot. [9]

2.1.3 Convolutional Neural Networks (CNN)

Convolutional neural networks (CNNs) are deep learning models that are used to automatically learn stacked spatial patterns in images. Convolutional layers, pooling layers, and fully-connected layers make up a CNN. Convolutional layers apply filters

(kernels) across the image to detect local patterns (edges, textures, etc.), while pooling layers downsample the feature maps to highlight spatial relations. By stacking many such layers and using backpropagation, CNNs automatically learn hierarchical feature representations from raw pixels. This makes CNNs a very promising option to handle the high-dimensional, structured nature of image data in a scalable way.

2.1.4 Challenges of CNNs with IRT

A major challenge when working with thermograms and CNNs is the data heterogeneity issue. Pictures from different hospitals look different due to their dissimilar setup, environment, cameras, patient etc. As a result, models trained on dataset from one hospital or setup may perform poorly or not perform at all on unseen new data (in a dataset from a different environment). Hence to resolve this issue, usage of diverse data and external validation is necessary.

However, despite diverse datasets, overfitting remains another key concern when using CNNs. Overfitting occurs when a model memorizes training data too well and fails on new data, this occurs the most when the pool of data is very low. To deal with overfitting many techniques like data augmentation (creating more training data by rotating, flipping and readjusting existing ones), regularization techniques (like dropout, weight decay and batch normalization for simplifying the model) and hold-out testing (evaluation on unseen data) are used. Grad-CAMs are used in the final evaluation to verify whether the model focuses on background noise or on the regions of interest. By using these strategies, reliable and robust DL models can be made to detect plantar thermograms.

2.2 Publicly Available Datasets

From our findings, there are two publicly available plantar datasets. The first one is from IEEE DataPort [13] (hereafter referred to as the IEEE Dataset), and the second one is the STANDUP Multispectral Plantar Foot Dataset [15] (hereafter referred to as the STANDUP Dataset).

2.3 Review of Existing Research

Multiple studies have been conducted to implement Plantar Thermogram analysis with Deep Learning models. In this section, the studies will be reviewed and summarized to help explain the current status of the research. The methodological limitations and shortcomings are evaluated in order to help advance research rigor

- A study [16] compared traditional Machine Learning (ML) classifiers against CNNs for binary classification (Diabetic vs Control) using the IEEE dataset. The dataset consisted of 244 diabetic and 90 control subjects. The study

worked out 39 features (e.g., mean, temperature, Thermal Change Index, Normalized Temperature Ranges) and thermal parameters (temperature estimation and hotspot estimation). Twenty-eight optimized features were selected out of the 39. Out of ten ML classifiers, AdaBoost with Random Forest feature selection and the top 10 features achieved the highest performance (96.71% accuracy, 97.75% sensitivity, 93.85% specificity, and 96.70% F1-score). For CNNs, six architectures (ResNet18, ResNet50, DenseNetV201, InceptionV3, VGG19, MobileNetV2) were used with transfer learning, with MobileNetV2 achieving 95.81% accuracy, 96.72% sensitivity, and 93.83% specificity. The studys limitations include that the best accuracy (95.81%) was achieved with dual-foot input, excluding amputees or unilateral cases, and it is unclear whether augmentation was applied pre-split, which may affect the generalizability of results.

- Another study [17] evaluated five ImageNet-pretrained ConvNeXt configurations (Tiny, Small, Base, Large, XLarge) to automatically extract features from the same IEEE dataset (244 diabetic, 90 control). After normalization, two augmentation pipelinesconventional (rotation ś10ř, horizontal flip) and fancy PCA color augmentationwere implemented, labeled as First and Second study respectively. These augmentations were then classified using fully connected layers (FC), linear regression (LR), and Support Vector Machine (SVM). Their proposed model achieved 100% accuracy using ConvNeXt XLarge for feature extraction, Linear Regression as the classifier, and Fancy PCA as the augmentation method. The studys limitations include the lack of validation metrics despite creating a separate validation set, unclear attribution of AUC results (test vs. validation), absence of hyperparameter tuning details, and unusually perfect scores for a small dataset, which raise concerns about potential data leakage and overfitting.

- A separate study [18] proposed multi-class classification of Diabetic Foot Ulcers using the same IEEE dataset. They defined a six-class grading scheme (Class 0: Control, Classes 15: Severity Grades for diabetic foot) and used the Thermal Change Index (TCI), the mean absolute temperature deviation per angiosome region from norms established in the control group. Class imbalance was addressed via augmentation methods (rotation, scaling, flip, shear, contrast, brightness). Using stratified 5-fold cross-validation and reporting metrics per class, their custom CNN achieved high performance (98.27% mean accuracy, 96.84% sensitivity, 98.92% specificity), significantly outperforming AlexNet (92.08% accuracy). Limitations include severe skew in raw data (Class 3: <40 samples vs. Class 5: >80), reliance on augmented rather than genuinely diverse thermal patterns, TCI-based grading not being clinically validated, and splitting a limited dataset into six classes increasing the risk of overfitting and reducing robustness and generalizability.

- Utilizing numeric data, a study [19] created an ML pipeline to identify diabetic foot ulcers (DFUs) early using the IEEE dataset. They used direct temperature CSV files and thermal images to analyze temperature distribution, employing SIFT and SURF descriptors with Bag of Features (BOF) to extract features, which were then input into seven ML classifiers (SVM, XGBoost,

PNN, LDA, KNN, RF, DT). Best performance was achieved with SURF-BOF and SVM (97.81% accuracy, 97.81% sensitivity, 97.9% precision, 97.815% specificity, AUC 0.9995), outperforming pretrained DL networks (ResNet-50, AlexNet, EfficientNetB0). Limitations include detecting diabetic foot rather than actual ulceration (dataset labels do not indicate ulcers), unsupported claims about patch-level annotation necessity, lack of an independent validation set for hyperparameter tuning, and potential inflation of metrics due to reliance on test folds for optimization.

- A study [20] developed two multiclass classification models for DFU severity staging using the IEEE dataset. Thermal images were segmented via K-means clustering into five temperature-based regions to extract regional descriptors and a Cluster Thermal Index (CTI). Model 1 performed direct 4-class classification (Healthy, Class 1 Mild, Class 2 Moderate, Class 3 Severe), while Model 2 used a hierarchical approach: Stage 1 used Logistic Regression (89.2% accuracy, 92.3% sensitivity, 80.6% specificity), and Stage 2 used SVM with a quadratic kernel (94.0% average F1-score). Model 2 outperformed Model 1, especially in distinguishing Healthy from Class 1. Limitations include the reliance on CTI, which is not medically validated, thresholds being statistically derived and not clinically grounded, sole dependence on temperature values instead of spatial thermal patterns, and ignoring clinically proven diagnostic thermal pattern information.

- Using both ML and DL, another study [21] was conducted on plantar thermograms. For ML (SVM/MLP), histogram-based segmentation optimized via Differential Evolution (DE) extracted handcrafted features from 110 thermograms. DL models (AlexNet, GoogleNet, DFTNet) used raw thermogram patches augmented via rotation/flipping. DFTNet (6 conv. layers + 1 FC) achieved 94.53% average accuracy across five classes, outperforming SVM (87.08%), MLP (81.67%), AlexNet (81.18%), and GoogleNet (81.84%). Limitations include reliance on TCI-based 5-class grading, lack of clinical validation, small dataset size (110 thermograms) despite augmentation, likely imbalanced class distribution, and unusually high accuracy metrics, raising concerns about reliability.

- Another study [22] proposed a computer-aided detection system for DFUs using texture and temperature asymmetry. Plantar regions were segmented via region-growing algorithms, with 11 ROIs manually selected per foot, extracting 12 GLCM-based texture features and 2 temperature features. Asymmetry between ipsilateral and contralateral ROIs was analyzed, showing significant thermal-texture differences. An SVM classifier achieved 95.61% accuracy, 96.5% sensitivity, and 92.41% specificity. Limitations include the very small cohort (60 individuals), reliance on multiple ROIs per subject which inflates dataset size without increasing inter-individual variability, and failure in cases where disease progression is symmetric between feet.

- Using multi-level thermographic data, another study [23] proposed an automatic DFU recognition system using the IEEE dataset (334 thermograms from 122 diabetic and 45 control subjects), augmented via rotation/flipping to 1000

full images and 3000 angiosome patches. Four handcrafted features were extracted and fed into seven ML classifiers across three data levels (combined, full-image, patch-only). For DL, ResNet50, DenseNet121 transfer learning, and a custom 7-layer CNN were used. Best ML result was 85.6% accuracy (XGBoost on full-image data), whereas CNN achieved 97.1% accuracy, 0.97 sensitivity, 95% specificity, and 0.976 AUC. Limitations include rotational augmentation distorting foot anatomy (feet are not rotationally symmetric), no reserved validation set or hyperparameter tuning, and potential overfitting due to small effective dataset.

- A study [24] proposed early DFU detection using multi-modal thermal and RGB images (STANDUP dataset). Images were preprocessed (resizing, rotation, flipping, normalization) and segmented using SegNet, achieving the highest IoU (0.95). ResNet-50 extracted modality-specific features, which were concatenated and classified using a simple neural network, reporting training accuracy 99.10%, validation 96.38%, and overall accuracy 98.55% with F1-score 0.9511. Limitations include no mention of data splitting, ambiguous validation metrics, misleading overall accuracy including training data, no steps to address data scarcity, potential overfitting due to heavy models without proper regularization, and lack of test results making reported performance unreliable.

- A multi-dataset study [25] utilized the IEEE dataset along with their own dataset (35 healthy and 35 diabetic individuals). Images were segmented into four foot regions, and temperature distributions and color texture features were extracted. Four feature combinations were tested with SVM and KNN classifiers, with SVM achieving 92.86% accuracy, 96.55% sensitivity, 84.62% specificity, 93.33% precision, and 94.92% F1-score using the RBT feature set. Limitations include no mention of validation, small sample size risking overfitting, absence of DL methods to simplify feature extraction, lack of augmentation or class imbalance handling, and limited inter-individual variability, reducing reliability.

- Another study [26] proposes a novel hybrid diagnostic framework for the early detection of Diabetic Foot Ulcers (DFUs) using plantar thermograms. The key innovation is the integration of handcrafted features, extracted using the ORB (Oriented FAST and Rotated BRIEF) algorithm and Bag of Features (BOF) method, with deep feature representations from pretrained convolutional neural networks (ResNet50, AlexNet, EfficientNet). This fused feature set is fed into a custom lightweight Deep Neural Network (DNN) classifier. The model was trained and tested on a public dataset of 1670 thermal images. It achieved excellent performance with an accuracy of 98.51%, precision of 100%, sensitivity of 98.98%, and an AUC of 1.00. An ablation study confirmed the superiority of the hybrid (ORB+DL) approach over using either handcrafted or deep features alone. The authors highlight that their model is the first to integrate ORB-based descriptors with deep CNN features for thermal-image-based DFU detection, resulting in a system that is both highly accurate and suitable for real-time clinical deployment due to its computational efficiency.

Authors of another study [27] proposed a feature selection framework for early DFU detection using plantar thermograms. Four selection methods (Lasso, RF, Concrete Dropout, Variational Dropout) were applied to an extended, multi-source dataset. The top-ranked features were used to train an SVM classifier. Performance degraded when merging datasets (STANDUP and IEEE), indicating sensitivity to data heterogeneity from different acquisition protocols. While SMOTE was used, the underlying class imbalance (more diabetic cases) persists in source data.

While existing studies have reported promising results in thermal DFU classification, several methodological limitations remain. Almost all of the works rely on single-dataset evaluations, which restrict understanding of model generalizability across different populations and imaging devices. Dataset size is often augmented to compensate, yet excessive augmentation risks altering pathological features critical for accurate classification. Several studies also rely on computationally intensive architectures (e.g.,ConvNeXt XLarge [17], ResNet50[24]) or complex handcrafted feature extraction methods (e.g., [19], [22], [25]), which may hinder efficient implementation and continuous learning from newer datasets on cost-effective devices.

The only study [27] utilizing the combined dataset approach (STANDUP + IEEE) faces performance degradation when evaluating on the combined dataset due to data heterogeneity.

Moreover, none of the CNN-based approaches use Grad-CAM or similar techniques to verify that models focus on relevant ulcer regions rather than background artifacts, a likely concern given the small datasets further limiting the robustness of reported results. Instances of near-perfect metrics (eg.,[17] or very high metrics on limited datasets also suggest potential overfitting or data leakage and cannot be verified without Grad-CAMs.

These observations highlight persistent gaps in dataset diversity, model generalization, validation rigor, and explainability, motivating the development of approaches that enforce cross-subject evaluation and systematic assessment of model focus to improve clinical reliability.

2.4 Summary of Key Findings

Overall, these studies suggest that deep learning with pretrained architectures can yield very high results on the IEEE dataset. Machine Learning models also perform on-par with Deep Learning models using manually extracted features as well as temperature data for binary classification. However, some results with perfect and near-perfect scores on this limited dataset hints greatly at overfitting issues and data-leakage, despite their measures to counter the lack of data and maintain splits. As the STANDUP dataset is fairly new, it is not widely utilized in classification tasks. However, a study [27] combined the STANDUP dataset with the IEEE dataset but achieved poor performance due to data heterogeneity, a key limitation we address

in this research.

Study	Task	Proposed Model	Acc.	Rec.	Prec.	F1	Spec.	AUC
[16]	Binary	AdaBoost + RF	96.71%	96.71%	96.70%	96.70%	94.58%	–
		MobileNetV2	95.81%	95.81%	95.84%	95.82%	94.24%	–
[17]	Binary	ConvNeXt XLarge + Fancy PCA + LR	100%	100%	100%	100%	100%	100%
[18]	Multiclass (6-class)	Custom CNN	98.27%	96.84%	96.26%	96.21%	98.92%	–
[19]	Binary (Early DFU)	SURF-BOF + SVM	97.81%	97.81%	97.90%	97.93%	97.82%	99.95%
[20]	Multiclass (4-class)	Hierarchical LR + SVM	93.20%	86.90%	86.70%	86.70%	–	–
[21]	Multiclass (5-class)	DFTNet	94.53%	95.34%	94.01%	94.57%	93.75%	94.55%
[22]	ROI Classification	SVM (Texture & Temp)	95.61%	96.50%	–	–	92.41%	–
[23]	Binary	Custom CNN	97.00%	97.00%	82.30%	89.10%	95.80%	97.60%
[24]	Binary	ResNet-50 + Neural Network	98.55%	–	–	–	–	–
[25]	Binary	SVM (RBT Features)	92.86%	96.55%	93.33%	94.92%	84.62%	–
		KNN (RBT Features)	85.71%	89.66%	87.50%	88.57%	76.92%	–
[26]	Binary	Hybrid ORB+DL $\rightarrow$ DNN	98.51%	100%	98.98%	98.97%	98.98%	1.00
[27]	Binary (STANDUP)	Feature Selection $\rightarrow$ SVM	91.46%	96.08%	86.04%	90.70%	–	–
[27]	Binary (IEE+STANDUP)	Feature Selection $\rightarrow$ SVM	86.19%	92.50%	79.31%	85.13%	–	–

Table 2.1: Performance comparison of diabetic foot classification studies.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

3.1.1 Functional Requirements

- Input: plantar thermogram images from FLIR E60 / FLIR ONE Pro or equivalent thermal cameras; support for both Rainbow HC (IEEE) and white-hot (STANDUP) formats.
- Preprocessing: colormap conversion pipeline to standardize Rainbow HC colormapping with White-Hot colormap, cropping and mask-based padding to preserve plantar patterns.
- Model inference: support for lightweight CNNs (GhostNet_100, EfficientNet-B0, MobileNetV3-Small, ShuffleNetV2, ResNet-18) with ability to run on an edge/mobile device or a small GPU server.
- Explainability: Grad-CAM visualizations for each prediction to confirm model focus on plantar regions.
- Evaluation: support for class-wise metrics (accuracy, precision, recall, specificity, AUC, F1) and composite scoring described in Chapter 5.

3.1.2 Non-fucntional Requirements

- Performance: compute suitable for near-real-time screening on capable edge hardware (i.e., model sizes and FLOPs constrained as reported).
- Robustness: person-wise splitting and ID tracking to avoid data leakage; augmentation and domain balancing to improve generalizability.
- Reproducibility: provide scripts for preprocessing (including colormap conversion), training, and Grad-CAM generation; document dataset splits and seeds.

3.1.3 Hardware and software resources

- Typical development machine: modern x86 workstation; for training, CUDA supported GPU with at least 8-12 GB VRAM recommended, to locally run experiments described.

3.2 Societal Impact

3.2.1 Health benefits

Early screening using thermal imaging can reduce incidence of undetected diabetic foot complications, potentially lowering ulceration and amputation rates by enabling earlier clinical referral. The studys focus on low-cost imaging (FLIR ONE Pro) improves accessibility for low-resource settings such as rural and underdeveloped areas.

3.2.2 Accessibility

Use of cost-effective equipment and efficient, lightweight models promotes broader screening availability in clinics and community health centers, contributing to preventive care coverage. This research helps demonstrate the potential of implementing deep-learningbased healthcare for diabetic foot patients worldwide, ensuring accessibility is not limited by heavy or expensive equipment.

3.3 Environmental Impact

3.3.1 Computational footprint

Training of CNNs has a carbon cost. Utilizing lightweight models (GhostNet_100, EfficientNet-B0 etc.) reduces compute and energy consumption over heavier architectures, which can be significant when viewed from a collective point of view.

3.4 Ethical Issues

3.4.1 Privacy and consent

Plantar thermograms are medical-adjacent data. The datasets used were publicly available for use with informed consent, explicit data usage permissions, and anonymization[15][13].

3.4.2 Clinical responsibility

Models are diagnostic aids, not final clinical decisions. Clear user guidance and disclaimers must accompany any deployment. The classification alone must never be used as the final diagnosis.

3.5 Risk Management

3.5.1 Data leakage

Mitigation via ID-wise/group-wise splitting and deterministic seeds used to prevent data leakage between data splits and to make sure unaugmented images were used in validation.

3.5.2 Model overfitting

Use of augmentation, cross-validation and composite scoring to detect and mitigate overfitting. Grad-CAM used to detect overfitting to noise, making sure the selected models are robust, unbiased and focused on clinically relevant features.

3.5.3 Dataset Heterogeneity

Use of color-conversion pipeline harmonized the datasets, ensuring models learned generalized features instead of dataset-specific information.

3.6 Economic Analysis

3.6.1 Operational Cost Considerations

Primary costs include cloud and GPU resources for model training and periodic re-training, maintenance of the inference pipeline, and clinical expenses associated with routine follow-up examinations following initial screening.

3.6.2 Cost−Benefit Framework

Replacing costly medical thermography equipment (typically several thousand EUR) with smartphone-compatible thermal imaging devices (EUR 350) can substantially reduce procurement expenses while preserving sufficient diagnostic and screening capabilities for low-resource settings. The integration of heterogeneous datasets further enables the development of models that generalize across device types, improving scalability and maximizing return on investment.

3.6.3 Economic Impact

Early detection significantly reduces downstream costs associated with advanced treatments and amputations. It also minimizes the risk of misdiagnosis and repeated clinical visits, enabling faster intervention and preventing disease progression.

Chapter 4

Methodology

4.1 Methodology Overview

This research evaluates a variety of lightweight CNNs in order to create a feasible and generalizable DFU thermal pattern classification pipeline for lightweight/mobile devices. A data transformation pipeline was been used to normalize the data type, i.e transform the IEEE dataset from Rainbow HC to White-Hot thermal to match the palette of the STANDUP Dataset. Specific IDs have been assigned to each individual to help identify them in order to prevent data leakage. An augmentation pipeline was then used to split the data into an 80:20 ratio for training and testing splits respectively. The split was done ID-wise to prevent different images from a single patient from appearing in multiple splits. Augmentation was then applied on the training split only. 20% of the training data was reserved for validation in the main training pipeline. Suffixes were assigned to augmented images in order to identify them and prevent them from being shared into the validation split. An ID filter was be used in the main training pipeline to ensure that a single individual, regardless of different images, remains in either training or validation. This ensured protection against data leakage. Firstly, well-established hyperparameters were used in training the models. A second training run was done using hyperparameter tuning with Optuna. Both methods used the same final evaluation strategy, utilizing Grad-CAM to visualize the regions of interest during testing. Usage of Grad-CAM clarified whether the models focused on the plantar region or overfitted on background information, giving a strong understanding of model behavior, a blackbox shortcoming that was not addressed in other studies.

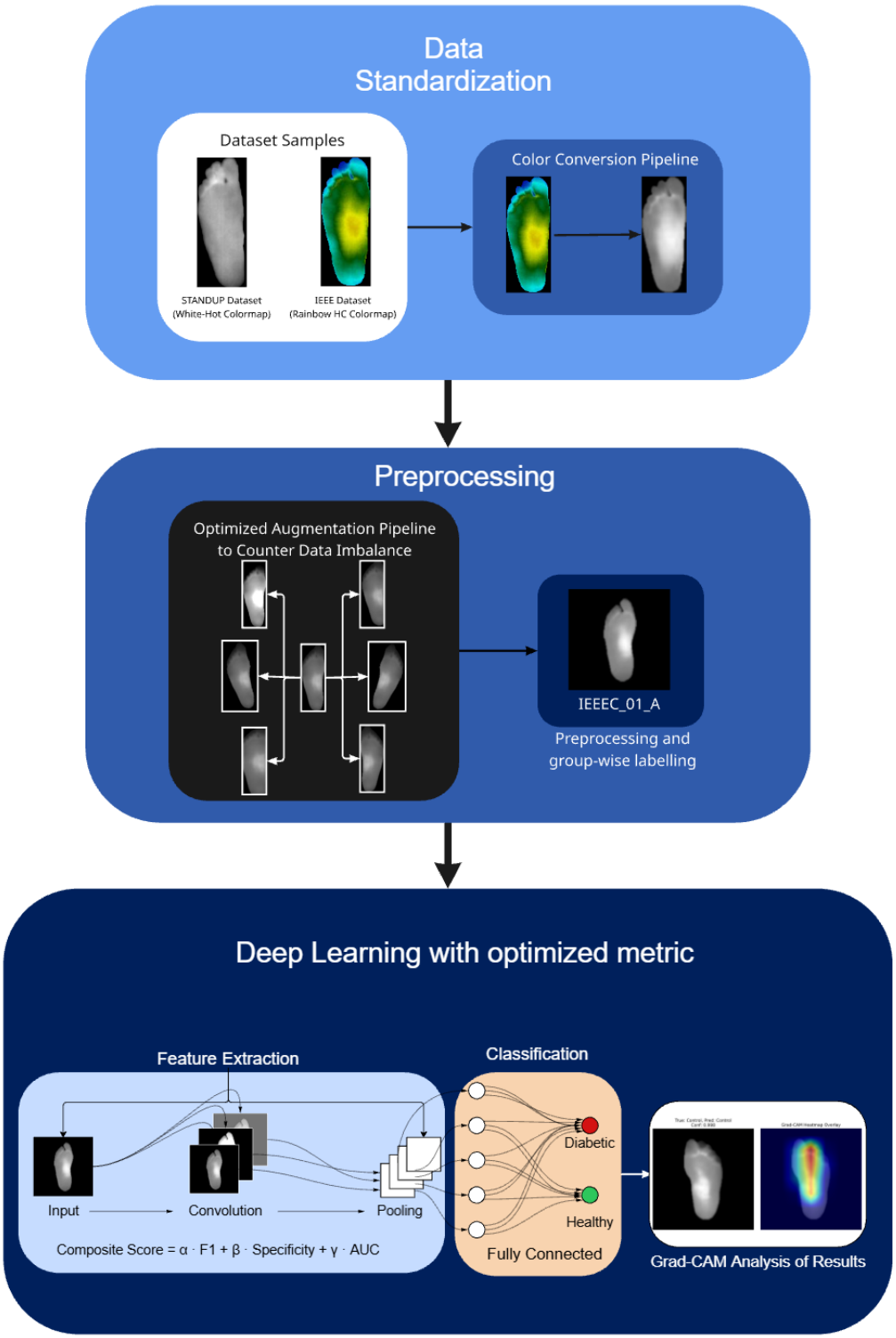

Figure 4.1: Schematic demonstration of methodology overview.

4.2 Computing Environment and Software

All experiments were conducted on a single development workstation equipped with an AMD Ryzen 5 2600 CPU, 16 GB of system RAM, and an NVIDIA GeForce RTX 3070 GPU; code development and debugging were performed in Visual Studio Code. Model training and evaluation were implemented in Python using CUDA-accelerated PyTorch and standard imaging library (Pillow). Training runs used deterministic settings where feasible (e.g., fixed seeds) and all experiments were logged with timestamps, and saved checkpoints so that reported runtimes and resource usage can be reproduced on comparable hardware.

4.3 Pipeline Design

4.3.1 Dataset Processing

The IEEE dataset and STANDUP dataset are the primary datasets of the process. The STANDUP dataset contains thermal images captured using the FLIR ONE Pro camera, which is a smartphone-compatible camera priced at approximately EUR 350. The dataset contains thermal images of 62 healthy individuals as the control group and 120 diabetic individuals, with two images from most individuals [15]. The IEEE thermogram dataset consists of 167 individuals, 122 of whom are diagnosed with diabetes and 45 non-diabetic individuals being the control group [13] The IEEE dataset goes through a colormap conversion pipeline before going into the final augmentation pipeline along with the STANDUP Dataset. The STANDUP dataset had RGB images and thermograms for each image however, the RGB images did not align with the thermograms, making it unfeasible to use the RGB images as masks for segmentation. Instead, the thermal images had to be manually segmented using editing software.

4.3.1.1 Colour Conversion Pipeline

A simple grayscale conversion on the IEEE dataset does not retain spatial information about temperature distribution, as different colors can map to the same grayscale value. To address this, we implemented a colormap conversion pipeline with standard RGB-to-HSV conversion and custom adjustments for thermogram interpretation. The pipeline converts Rainbow HC thermograms into white-hot grayscale by first converting each image to HSV and extracting the hue channel $hue \in [0, 360°]$. The hue scale starts from red at 0°, passes through yellow (60°), green (120°), cyan (180°), blue (240°), magenta (300°), and finally pinkish-red at 359°, looping back to red at 360°. Black background pixels are masked, and a Gaussian blur is applied to smooth transitions. Each processed frame is saved as a grayscale thermal image, preserving the underlying heat distribution while standardizing the colormap.

Since hue values wrap around the color wheel, they are shifted and normalized into the unit interval:

$$nh = \frac{(hue + 30) \bmod 360}{360} \tag{4.1}$$

This normalization ensures that the red/yellow region, which indicates hotter areas in Rainbow HC thermograms, is consistently positioned within the unit domain, preserving spatial correspondence across samples. A base grayscale inversion is then applied so that lower hues (near red/yellow) correspond to higher brightness levels, following the widely adopted white-hot thermal imaging convention:

$$base = 1 - nh \tag{4.2}$$

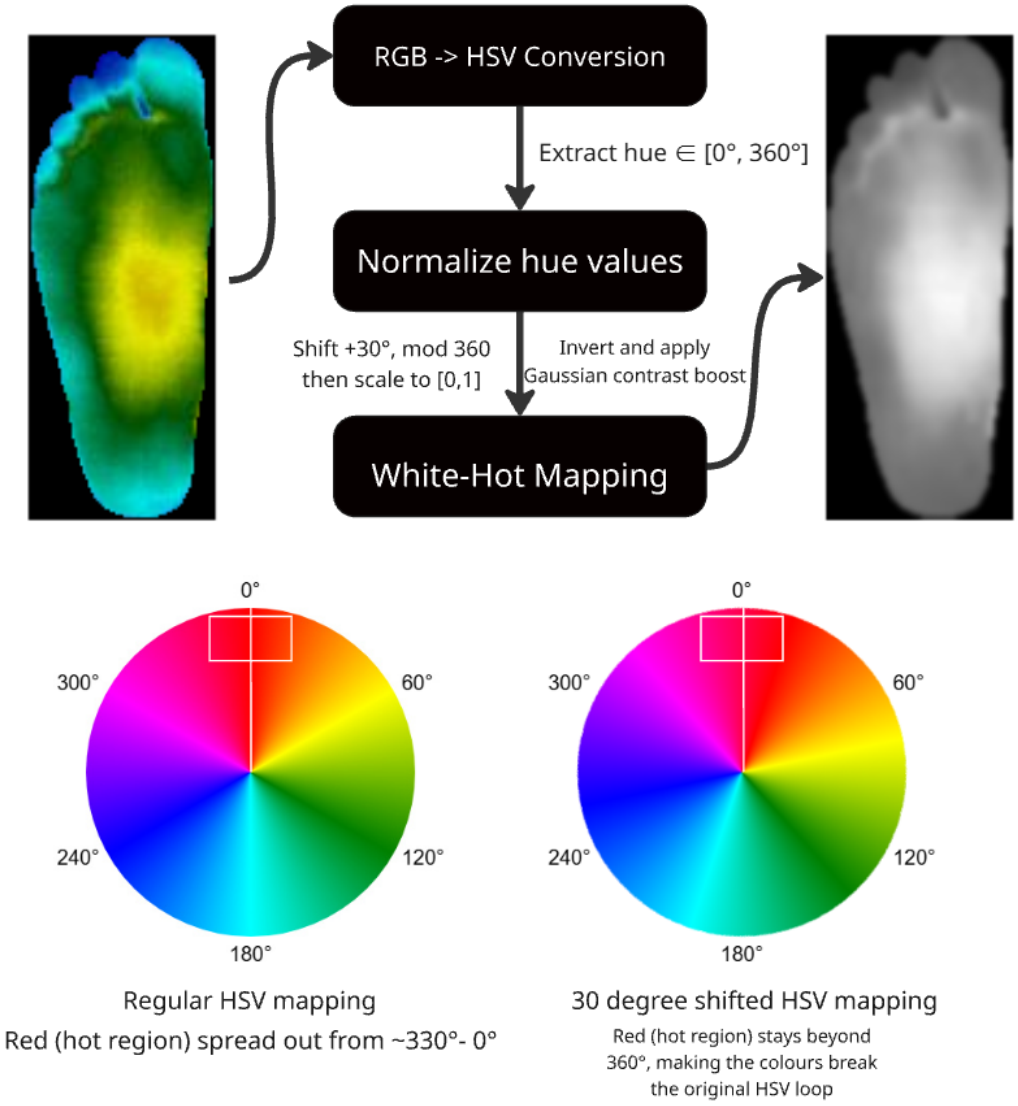

Figure 4.2: Colour conversion pipeline.

To emphasize and visually isolate high-temperature regions, a Gaussian-based contrast boost is applied around the red/yellow hue range:

$$gaussian = A \cdot \exp\left(-\frac{1}{2}\left(\frac{nh - c}{w}\right)^2\right), \quad A = 0.5,\ c = 0.1,\ w = 0.1 \tag{4.3}$$

Here, the Gaussian center c defines the target hotspot hue, the width w restricts enhancement to a narrow temperature band, and the amplitude A controls the strength

of contrast amplification. This selectively increases brightness and sharpness of thermal hotspots while suppressing enhancement in mid- and low-temperature regions, improving hotspot separability without saturating the image or distorting background information.

Finally, the grayscale intensity is clipped to the $[0, 1]$ range for numerical stability and scaled to 8-bit precision for storage and compatibility with standard computer vision frameworks:

$$gray = \text{clip}(base + gaussian, 0, 1) \times 255 \tag{4.4}$$

This pipeline ensures thermal and spatial information is preserved while bringing IEEE images into the same White-Hot format as the STANDUP dataset.

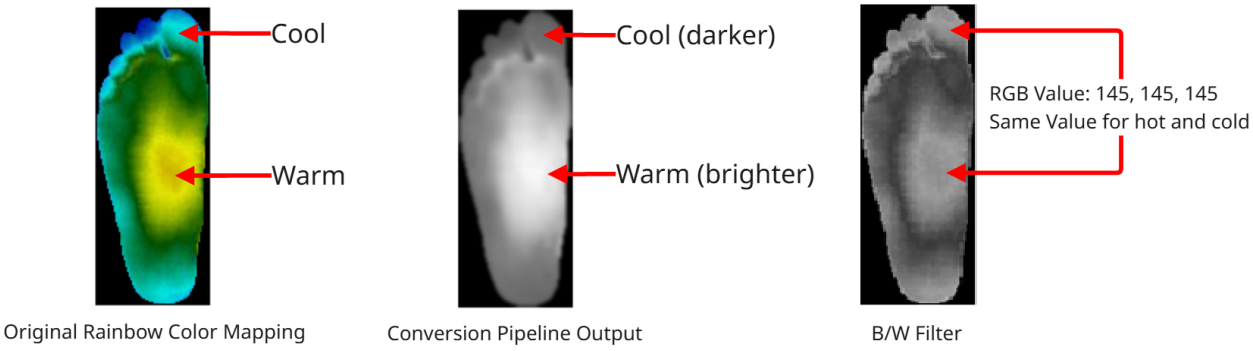

Figure 4.3: Comparison of the proposed conversion pipeline with a simple grayscale filter.

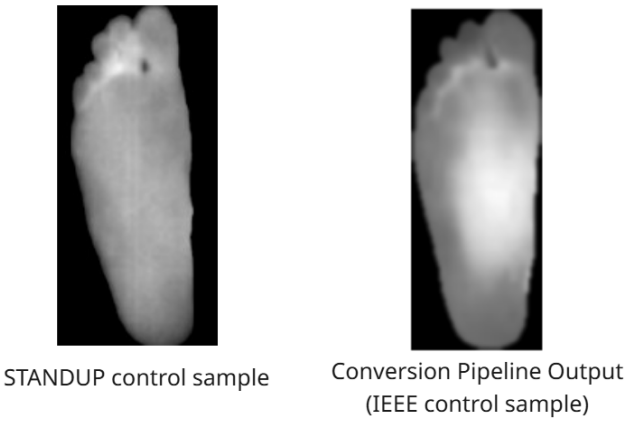

Figure 4.4: The White-Hot mapping standardizes the IEEE dataset with the STANDUP dataset.

4.3.1.2 Augmentation Pipeline

The combined dataset was split into an 80:20 ratio for training and testing respectively, after which augmentation was applied to the images based on their IDs. As different classes and domains (class type and dataset type) have different data distributions, augmentation was specially applied to each to balance the distribution. The initial class distribution was extremely spread out with 90 IEEE control samples (IEEEC), 240 Standup control samples (STC), 244 IEEE diabetic samples (IEEED) and 410 Standup diabetic samples (STD). All of the augmentations are applied once a reserved test split was created using a deterministic seed. The split occurs group wise (by personID) to ensure no data leakage occurs between splits, i.e all images of a single individual remain in one split. This was specially helpful for the STANDUP dataset where there are multiple pairs of the same person (creating a total of 4 separate images of a single individual).

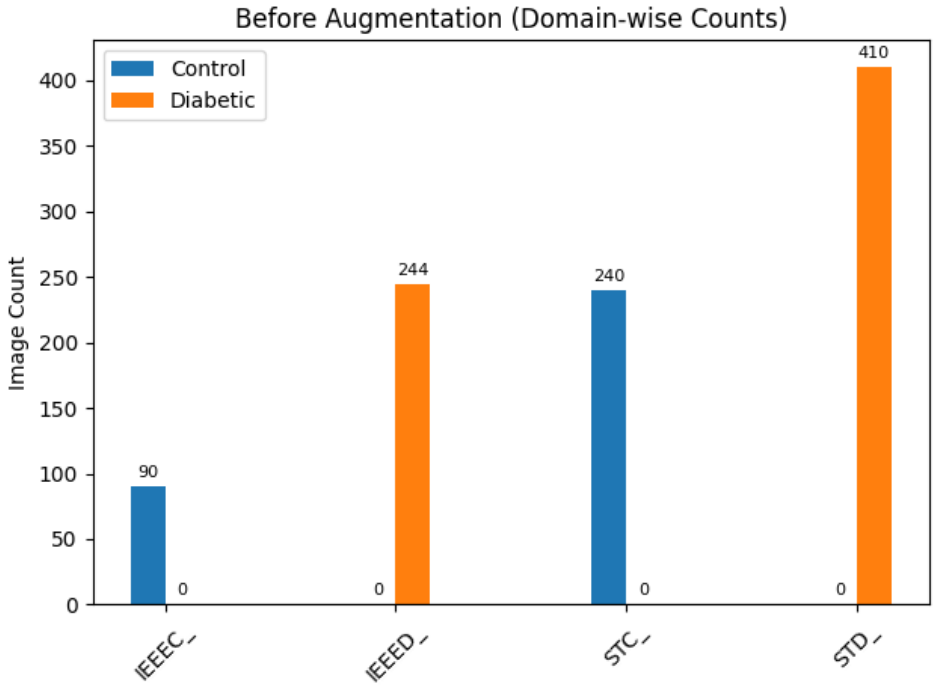

Figure 4.5: Data distribution before augmentation.

The domains are selectively augmented in the following manner:

- Heavier augmentations are applied to the IEEEC dataset, including flips, rotations ($\pm 10 - 15°$), brightness and contrast adjustments, flip+rotation combinations, perspective distortion, and affine transformations (translation, scaling, shearing). Each original image was transformed into seven augmented images, resulting in a total of 504 images in the training split.
- IEED and STC: Augmented only until they reach exactly 504 images each using horizontal flips and rotations.
- STD was the least augmented as it already contains a high number of images. Flips and rotations are applied to extend it to 504 images.

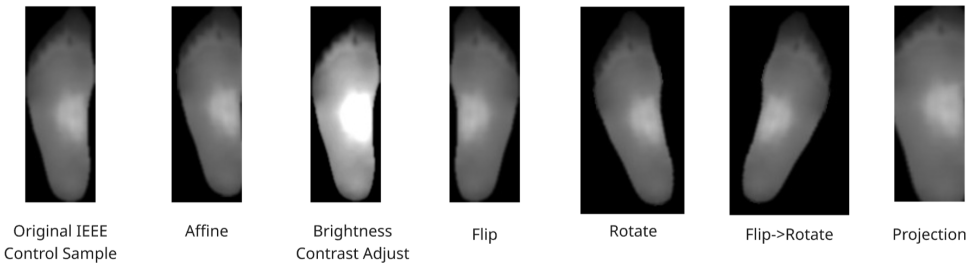

Figure 4.6: Augmentation Demonstration for IEEE Control.

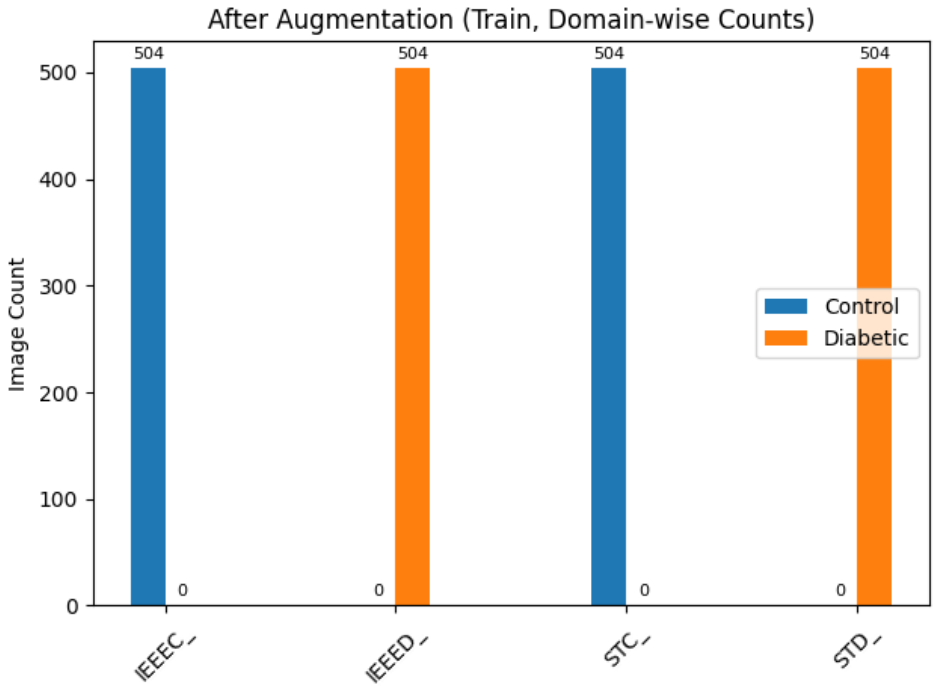

Figure 4.7: Data distribution for training after augmentation.

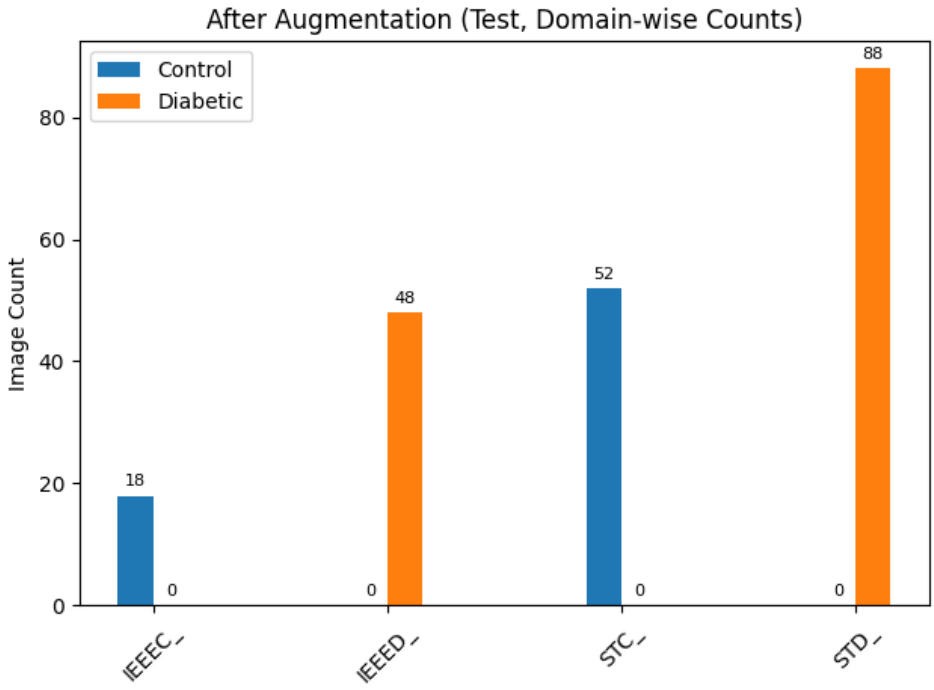

Figure 4.8: Data distribution for test split.

The test set was split domain-wise due to the variance in original image numbers. The final training dataset contains 504 images from each domain. The IEEE control domain has been extensively augmented to balance dataset variance. Further augmenting this domain may result in duplications or distortions in the images. Rest of the domains have been augmented to match the number of images in the IEEE control domain. A total of 1008 images per class were present in the training set, making a total number of 2016 training images from the original 984 images. For testing, using original images, a total of 70 control images and 136 diabetic images were reserved.

4.3.1.3 Preprocessing Pipeline

Images were first loaded in RGB format to ensure compatibility with pretrained networks. Each image was converted to a tensor, and a binary mask was created by identifying all non-zero pixels, representing the foot region. The tensor was then cropped to the bounding box of this mask to remove background noise. If the mask was empty, the full image was used as a fallback. The cropped image and mask were resized to 224x224 pixels with black padding to preserve the foots aspect ratio and prevent distortion. Both the image and mask were converted back to tensors, with the mask binarized using a 0.5 threshold. Labels are loaded as PyTorch tensors. The padding was also implemented after models were found to be overfitting on background information such as distance from foot region to borders. This issue was identified using Grad-CAM and after the padding was implemented, the issue was resolved. Person-level grouping ensures no data leakage across dataset splits.

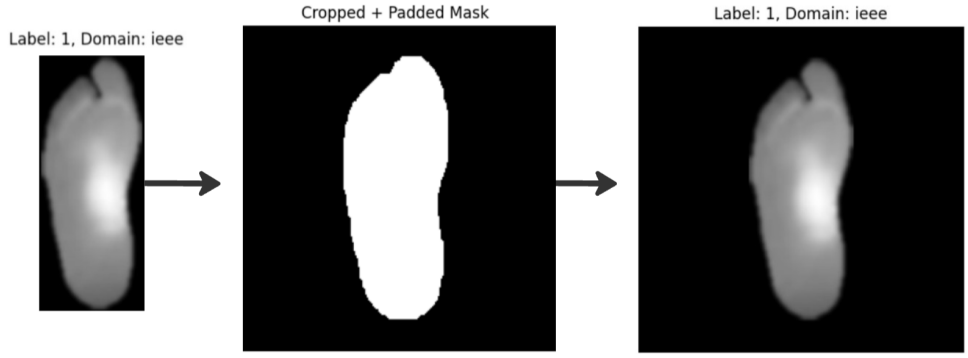

Figure 4.9: Image preprocessed to square dimensions while maintaining aspect ratio to retain thermal patterns and prevent overfitting.

4.3.2 Models

A number of light-weight models were carefully selected to balance performance with efficiency making sure the pipeline does not overfit by learning too in-depth features such as foot shape, angles or distance from borders. As our goal was to create a scalable pipeline that is efficient and functions with both expensive and cost efficient hardware, light-weight models are best-suited for our pipeline.

4.3.2.1 ResNet18

ResNet-18, is a convolutional neural network designed to address the issue of training very deep models. The authors proposed the concept of residual learning, where the network learns residual functions with respect to an identity mapping, using skip connections to allow direct flow of gradients through the network. ResNet-18 consists of 18 layers, with each block containing two 3x3 convolutional layers followed by a skip connection. The model is highly efficient, achieving excellent performance on benchmarks like ImageNet [28]. Although this model is the heaviest from our choice of models, we included it incase our pipeline is adapted when more datasets become available or cost effective devices become more strong and optimized.

4.3.2.2 EfficientNetB0

EfficientNet-B0 was introduced to address the challenge of scaling CNNs effectively. The authors identified that balancing network depth, width, and input resolution yields better accuracy and efficiency than scaling a single dimension. They performed a multi-objective neural architecture search (NAS) optimizing both accuracy and FLOPs, which produced the EfficientNet-B0 baseline architecture. EfficientNet-B0's design uses mobile inverted bottleneck (MBConv) layers augmented with squeeze-and-excitation (SE) blocks to enhance feature learning. The authors then applied a compound scaling coefficient to uniformly scale up this baseline in depth, width, and resolution, yielding the EfficientNet-B1 to B7 family. EfficientNet-B0 delivers 76.3% top-1 ImageNet accuracy with only 5.3 million parameters (0.39 billion FLOPs), matching ResNet-50s accuracy while using an order-of-magnitude fewer resources [29]. Efficiency combined with power makes it a suitable choice for our research.

4.3.2.3 MobileNetv3-Small

MobileNetV3-Small is one of two variants (Large and Small) proposed for optimizing mobile classification under low-resource constraints. The model uses hardware-aware neural architecture search (NAS) combined with the NetAdapt algorithm to produce an architectures for mobile phone CPU latency. It uses inverted residual (bottleneck) blocks (inherited from MobileNetV2) along with squeeze-and-excitation (SE) modules and efficient nonlinearities like h-swish and h-sigmoid for lower computational cost. The search also considers latency constraints, yielding a model that balances accuracy and speed. MobileNetV3-Small achieves 67.4% top-1 accuracy on ImageNet while reducing latency compared to MobileNetV2 and using fewer resources. The paper also reports MobileNetV3-Small being 4-5% more accurate yet somewhat faster than a MobileNetV2 model with similar latency [30]. This makes MobileNetV3-Small well-suited for our research where efficiency is critical. Most of the researches on early diabetic foot detection used the MobileNetv2 but left out the v3 small. Hence implementing this model will give us more insight into the field.

4.3.2.4 GhostNet_100

GhostNet introduces a novel Ghost module designed to exploit redundancy in feature maps by generating ghost feature maps via cheap linear operations rather than full convolutions. First, a small set of intrinsic feature maps is computed through conventional convolution; then, additional feature maps are derived via low-cost transforms (e.g. depthwise convolutions) applied to these intrinsic maps. This design reduces both parameters and FLOPs while maintaining expressivity. Ghost modules are assembled into Ghost bottlenecks (with expansion and compression stages) to build GhostNet architectures. The full GhostNet (often referred to as GhostNet_100) achieves around 75.7% top-1 accuracy on ImageNet with significantly fewer FLOPs compared to MobileNetV3-level models at similar resource budgets. Its plug-and-play modules can upgrade existing CNNs into more efficient architectures without drastic structural changes [31]. This is yet another promising model that remains unutilized in the early diabetic foot detection research field.

4.3.2.5 ShuffleNetV2_x1_0

ShuffleNetV2_x1_0 is the reference 1.0x width version of ShuffleNetV2, proposed to create lightweight CNNs that balance accuracy with real-world inference speed. Unlike many models optimized only for FLOPs, ShuffleNetV2 introduces practical design guidelines. It keeps equal channel widths to minimize memory access costs, avoids excessive group convolutions, reduces fragmentation in operations and limit element-wise operations. The architecture is built from Shuffle Units, which use depthwise separable convolutions and channel shuffle to ensure cross-channel information flow. The 1.0x variant achieves about 69.4% top-1 accuracy on ImageNet with just 2.3M parameters and 146M FLOPs, offering an excellent trade-off between efficiency and accuracy. Its design influenced subsequent efficient models like MobileNetV3 and GhostNet, establishing ShuffleNetV2 as a strong baseline for mobile and embedded applications [32].

4.3.3 Performance Reliability using Explainable AI

4.3.3.1 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is a widely used explainability technique for deep learning models in computer vision. It highlights image regions most influential for a models prediction by computing the gradient of the target class score with respect to feature maps in the last convolution layers. These gradients are globally averaged to produce importance weights, which are then combined with the feature maps to generate a heatmap over the input image. Grad-CAM is model-agnostic for CNN-based architectures and provides intuitive visual explanations for classification and localization tasks [33]. A lack of explainability with blackbox models is a crucial gap in existing research that facilitate unreliability in results. To bridge this gap, we implement Grad-CAM usage to verify whether our model overfit to noise or perform intuitions based on proper regions of interest. The regions of interest are then verfied through a licensed dermatologist to further validate the results.

.

Chapter 5

Result Analysis

5.1 Performance Evaluation

Initial evaluation using validation loss or F1-score alone led to models with high false-positive rates, which are clinically undesirable as they may prompt unnecessary follow-up procedures. To better align model selection with clinical priorities, emphasizing sensitivity (via F1) while controlling false positives (via specificity), we adopted a composite scoring metric.

$$\text{Composite Score} = \alpha \cdot F1 + \beta \cdot \text{Specificity} + \gamma \cdot \text{AUC} \tag{5.1}$$

The weights ($\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$) were determined through iterative sensitivity analysis to balance these competing objectives. Using this composite score as the primary evaluation metric resulted in more clinically balanced models with reduced false-positive rates.

5.2 Results

5.2.1 Performance Evaluation

All five architectures were trained using a single, consistent set of hyperparameters. Specifically, we used a learning rate of 1e-4, batch size of 32, Adam optimizer with weight decay 1e-5, and cross-entropy loss with class weighting. Models were trained for up to 50 epochs with early stopping (patience=15, delta=0.001) based on the composite score. Input images were resized to 224X224 with preserved aspect ratio and min-max normalized. Training used a person-level 80/20 train/validation split with no data leakage. The hyperparameters were selected based on common transfer learning practices and prior empirical experience with small-scale image classification tasks. Given the relatively limited dataset size (2,016 training images after augmentation), conservative defaults were favored to promote training stability and reduce overfitting. Preliminary hyperparameter tuning experiments did not result in consistent improvements across models. Consequently, the same parameter set was applied to all architectures to ensure reproducibility and fair comparison between models.

Models were evaluated across 5-folds for robustness check before going into final training. No leakage was confirmed between the cross-validation trial and final

training. GPU memory was cleared before and after each set of training to further ensure no data leakage occurred. Entire process was repeated with 12 times with different random seed values each time.

Model	Final best Val F1	Avg Val Composite (5-fold)	Avg Val F1 (5-fold)
GhostNet	0.9735 ± 0.0068	0.9604 ± 0.0097	0.9641 ± 0.0089
MobileNetV3-Small	0.9771 ± 0.0102	0.8983 ± 0.0214	0.8941 ± 0.0221
ShuffleNetV2-1x	0.9765 ± 0.0074	0.9630 ± 0.0086	0.9653 ± 0.0079
EfficientNet	0.9706 ± 0.0081	0.9643 ± 0.0072	0.9683 ± 0.0069
ResNet18	0.9602 ± 0.0095	0.9737 ± 0.0064	0.9749 ± 0.0061

Table 5.1: Final training and cross-validation metrics for each model (mean ± standard deviation across folds).

Model	Accuracy	Precision	Recall	F1	Specificity	AUC
ResNet18	0.946 ± 0.021	0.949 ± 0.017	0.977 ± 0.013	0.963 ± 0.015	0.888 ± 0.073	0.990 ± 0.011
EfficientNet-B0	0.949 ± 0.014	0.957 ± 0.015	0.980 ± 0.012	0.969 ± 0.010	0.880 ± 0.028	0.985 ± 0.006
ShuffleNetV2-1x	0.942 ± 0.012	0.950 ± 0.019	0.970 ± 0.012	0.960 ± 0.010	0.875 ± 0.032	0.971 ± 0.015
MobileNetV3-Small	0.922 ± 0.025	0.954 ± 0.032	0.930 ± 0.039	0.941 ± 0.026	0.891 ± 0.085	0.980 ± 0.006
GhostNet_100	0.960 ± 0.014	0.971 ± 0.015	0.979 ± 0.009	0.975 ± 0.009	0.943 ± 0.039	0.978 ± 0.010

Table 5.2: Overall model performance (mean ± standard deviation) across IEEE and STANDUP datasets.

Model	Domain	Accuracy	Precision	Recall	F1	Specificity	AUC
ResNet18	IEEE	0.928 ± 0.025	0.946 ± 0.021	0.963 ± 0.017	0.954 ± 0.016	0.852 ± 0.073	0.981 ± 0.016
	STANDUP	0.964 ± 0.012	0.952 ± 0.013	0.991 ± 0.006	0.971 ± 0.009	0.923 ± 0.025	0.998 ± 0.001
EfficientNet-B0	IEEE	0.915 ± 0.018	0.934 ± 0.018	0.964 ± 0.015	0.949 ± 0.011	0.778 ± 0.071	0.970 ± 0.009
	STANDUP	0.983 ± 0.014	0.979 ± 0.012	0.996 ± 0.006	0.988 ± 0.005	0.981 ± 0.013	0.999 ± 0.001
ShuffleNetV2-1x	IEEE	0.922 ± 0.017	0.944 ± 0.017	0.957 ± 0.015	0.950 ± 0.012	0.833 ± 0.039	0.945 ± 0.026
	STANDUP	0.961 ± 0.012	0.956 ± 0.020	0.982 ± 0.009	0.969 ± 0.008	0.917 ± 0.030	0.997 ± 0.002
MobileNetV3-Small	IEEE	0.871 ± 0.032	0.939 ± 0.049	0.871 ± 0.067	0.903 ± 0.043	0.833 ± 0.146	0.960 ± 0.010
	STANDUP	0.973 ± 0.017	0.969 ± 0.015	0.989 ± 0.011	0.979 ± 0.010	0.949 ± 0.024	0.999 ± 0.002
GhostNet_100	IEEE	0.932 ± 0.021	0.956 ± 0.021	0.958 ± 0.015	0.957 ± 0.014	0.911 ± 0.062	0.957 ± 0.019
	STANDUP	0.988 ± 0.006	0.986 ± 0.008	0.999 ± 0.003	0.992 ± 0.004	0.975 ± 0.016	0.999 ± 0.001

Table 5.3: Domain-specific model performance (mean ± standard deviation across 12 independent runs).

Model	Total Params (M)	FLOPs (G)	Reference
ResNet-18	11.18	1.80	[28]
EfficientNet-B0	4.01	0.39	[29]
ShuffleNetV2-1x	2.3	0.146	[32]
MobileNetV3-Small	2.4	0.044	[30]
GhostNet_100	5.2	0.141	[31]

Table 5.4: Comparison of chosen models showing total parameters and computational cost in FLOPs.

All the models demonstrated good results in both domains, with a slight better performance in the STANDUP domain. This resonates with the higher data diversity in the STANDUP dataset compared to the IEEE dataset. GhostNet_100 demonstrated superior performance over all the other models.

5.2.2 Model Interpretation and Comparative Analysis Using Explainable AI

Post-hoc analysis of Grad-CAM visualizations revealed overfitting in MobileNetV3-Small. The model frequently attended to background elements such as the distance from the foot to image borders, rather than plantar regions. This was evidenced by splotchy, diffuse attention patterns, indicating poor localization of clinically relevant features. ResNet18 demonstrated similar but partial focus on background information in a repeated patterns, suggesting overfitting to similar dataset specific cues. In contrast to MobileNetV3 and ResNet18, EfficientNet-B0, ShuffleNetV2-1x and GhostNet_100 maintained focused attention on plantar regions. GhostNet_100 demonstrated particularly reliable thermal pattern localization, confirming its suitability for diabetic foot diagnosis. Similar Grad-CAM visualizations were generated for all 12 runs in all models.

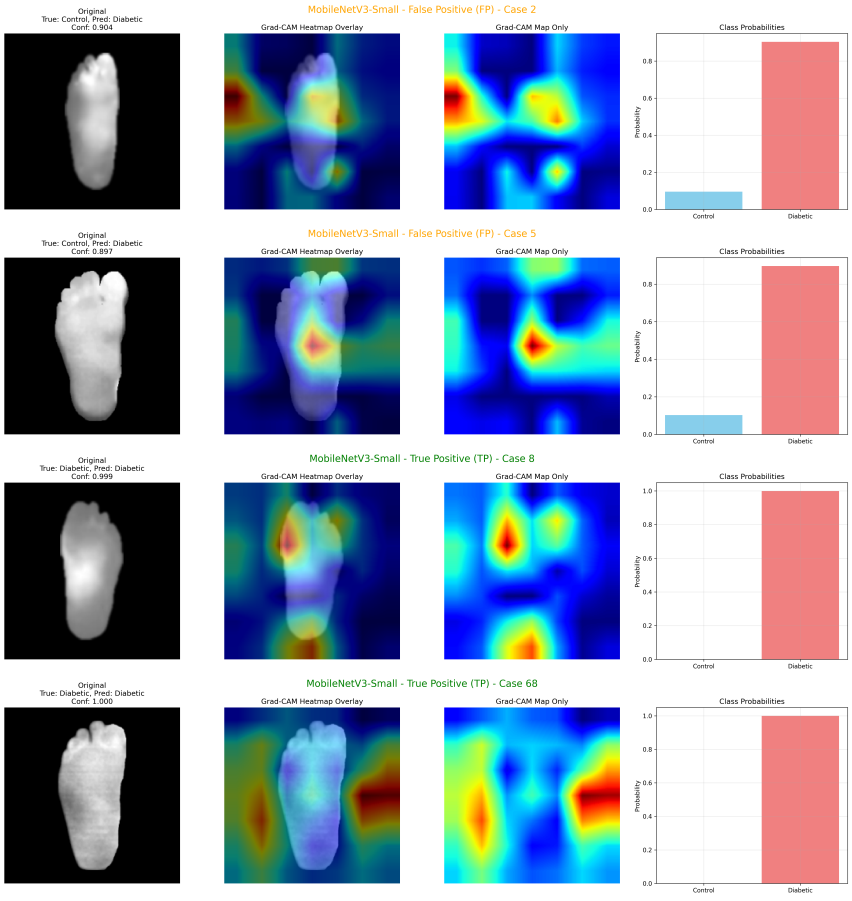

Figure 5.1: MobileNetV3 Grad-CAM demonstrating overfitting on noise.

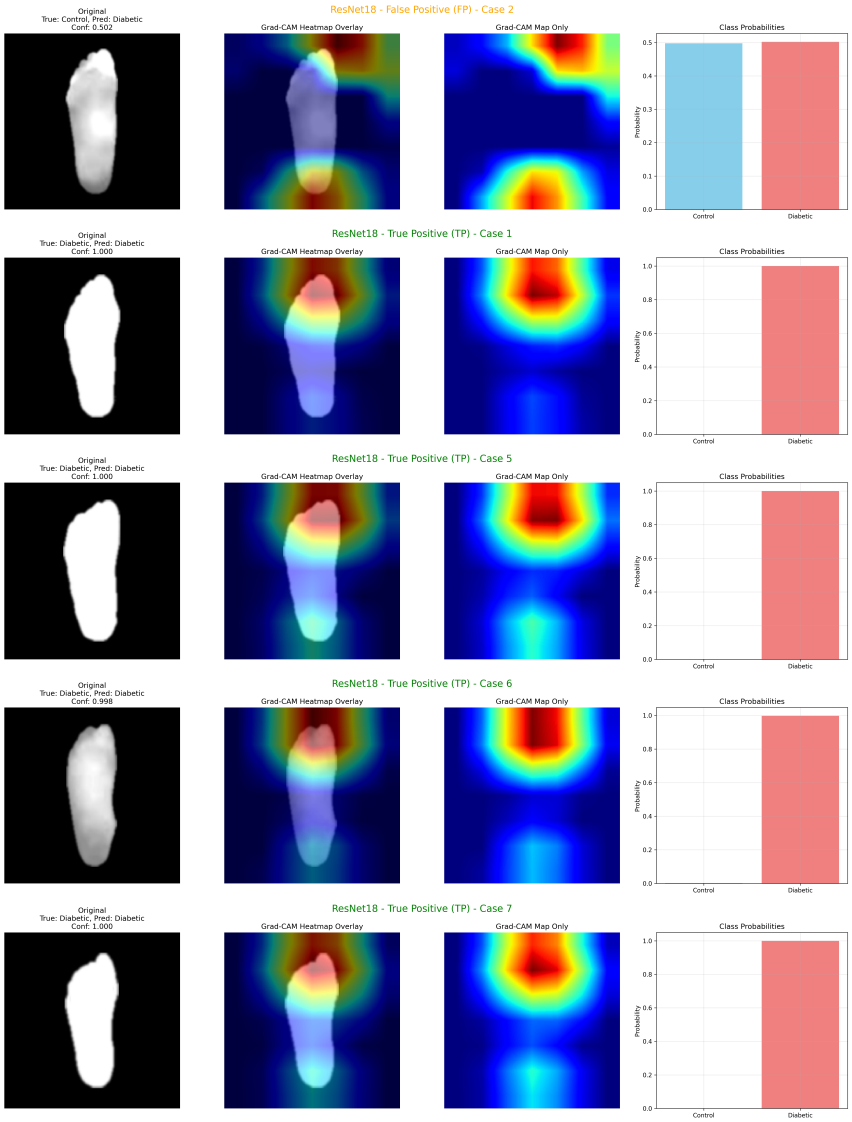

Figure 5.2: ResNet18 Grad-CAM demonstrating partial focus on background information.

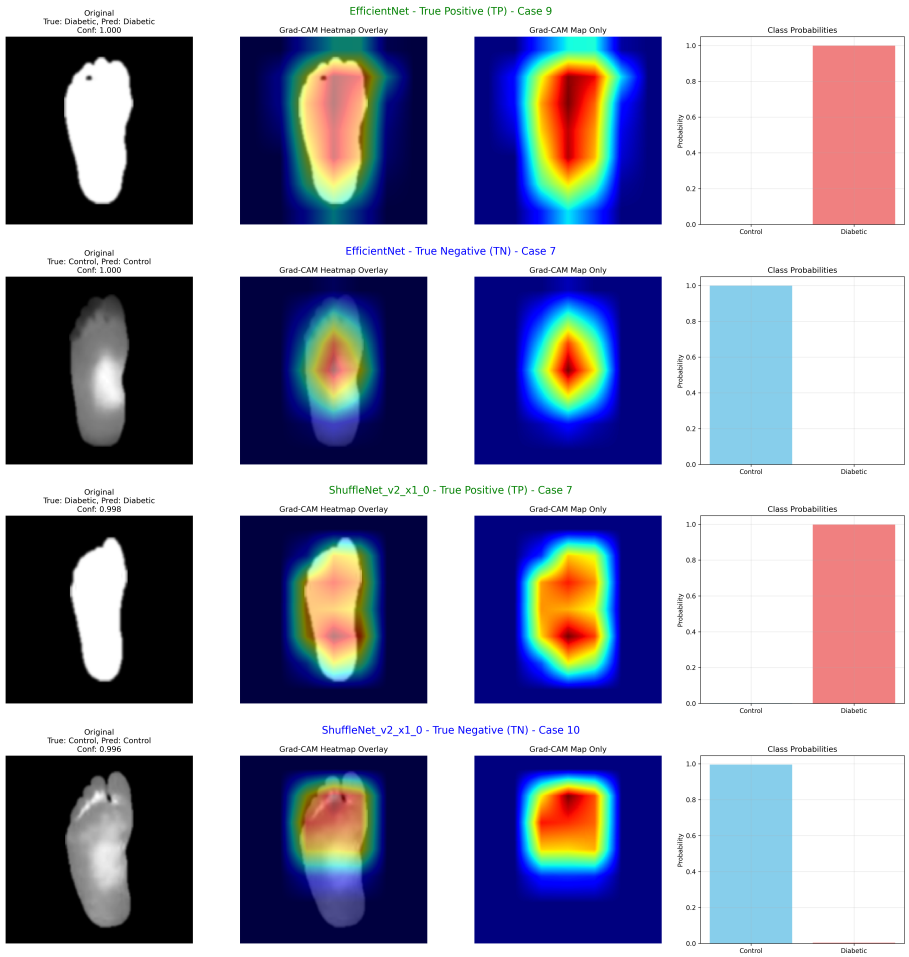

Figure 5.3: Grad-CAM outputs of EfficientNet-B0 and ShuffleNetV2-1x.

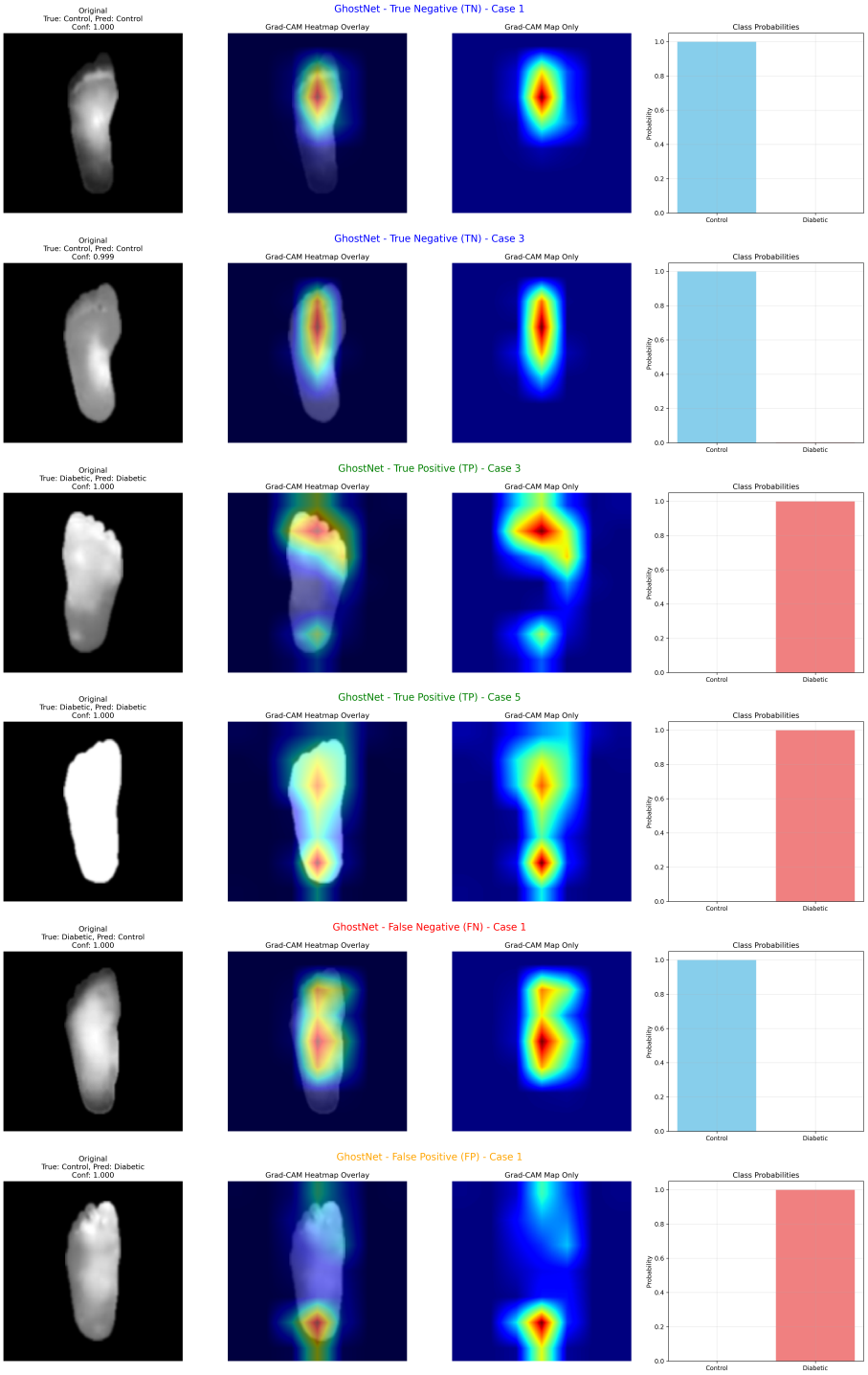

Figure 5.4: Grad-CAM outputs GhostNet_100.

5.2.3 Ablation Studies for Analysis of Dataset Diversity on Results

Two ablation studies were conducted to investigate the generalizability of the models across datasets. The first study investigates the performance difference when the models are trained only on the IEEE subset. The second study investigates the reliability of the results when training without colormap harmonization. The same data split was used as the main study.

5.2.3.1 Single vs Multi-Dataset Training

For the first study, the models were trained exclusively on the color-converted IEEE dataset and evaluating performance on the IEEE test domain. The results demonstrate that models trained on a single dataset generally exhibited degraded performance compared to multi-dataset training,a stark contrast to the existing research [27]. The IEEE dataset was selected as the baseline for this study due to its established prevalence in existing literature on plantar thermography analysis.

Model	Accuracy	Precision	Recall	F1	Specificity	AUC
ResNet18	0.909 ± 0.018	0.920 ± 0.014	0.958 ± 0.013	0.939 ± 0.012	0.778 ± 0.028	0.973 ± 0.011
EfficientNet-B0	0.909 ± 0.021	0.920 ± 0.024	0.958 ± 0.012	0.939 ± 0.019	0.718 ± 0.032	0.970 ± 0.014
ShuffleNetV2-1x	0.909 ± 0.019	0.938 ± 0.017	0.938 ± 0.016	0.938 ± 0.014	0.792 ± 0.033	0.983 ± 0.018
MobileNetV3-Small	0.818 ± 0.031	0.810 ± 0.036	0.979 ± 0.000	0.887 ± 0.028	0.389 ± 0.000	0.957 ± 0.021
GhostNet_100	0.924 ± 0.022	0.922 ± 0.020	0.979 ± 0.015	0.949 ± 0.017	0.778 ± 0.061	0.973 ± 0.016

Table 5.5: Performance on the IEEE subset after training on color-converted IEEE dataset only (mean ± standard deviation across 12 independent runs).

Model	Accuracy	Precision	Recall	F1	Specificity	AUC
ResNet18	+1.9%	+2.6%	+0.5%	+1.5%	+7.4%	+0.8%
EfficientNet-B0	+0.6%	+1.4%	+0.6%	+1.0%	+6.0%	0.0%
ShuffleNetV2-1x	+1.3%	+0.6%	+1.9%	+1.2%	+4.1%	−3.8%
MobileNetV3-Small	+5.3%	+12.9%	−10.8%	+1.6%	+44.4%	+0.3%
GhostNet_100	+0.8%	+3.4%	−2.1%	+0.8%	+13.3%	−1.6%

Table 5.6: Average change in IEEE-domain performance after training with both datasets.

This ablation study proves the effectiveness and cross-dataset generalizability and of our pipeline, where model performance across all models improved, with significant improvement for MobileNetV3-Small with our pipeline. Specificity improved largely for all models, improving reliability as false negatives are risky in clinical diagnostics. Grad-CAM visualizations remained similar to the main study in this IEEE-only training.

5.2.3.2 Effect of Colormap-Harmonization on Performance and Reliability

A second ablation study was carried out to demonstrate the effectiveness of our proposed pipeline. In this study, the IEEE dataset was left in its original color pallete (Rainbow HC) and trained alongside the STANDUP dataset.

Model	Accuracy	Precision	Recall	F1	Specificity	AUC	Composite
ResNet18	0.9709 ± 0.012	0.971 ± 0.010	0.9853 ± 0.008	0.9781 ± 0.009	0.9429 ± 0.015	0.995 ± 0.004	0.9661 ± 0.011
EfficientNet-B0	0.9612 ± 0.010	0.9507 ± 0.012	0.9926 ± 0.007	0.9712 ± 0.009	0.9000 ± 0.020	0.9941 ± 0.003	0.9478 ± 0.010
ShuffleNetV2-1x	0.9612 ± 0.009	0.9638 ± 0.011	0.9779 ± 0.010	0.9708 ± 0.008	0.9286 ± 0.014	0.9951 ± 0.005	0.9562 ± 0.009
MobileNetV3-Small	0.9709 ± 0.011	0.9643 ± 0.012	0.9926 ± 0.006	0.9783 ± 0.007	0.9286 ± 0.016	0.9982 ± 0.002	0.9619 ± 0.010
GhostNet_100	0.9612 ± 0.010	0.9776 ± 0.011	0.9632 ± 0.009	0.9704 ± 0.008	0.9571 ± 0.012	0.9964 ± 0.003	0.9646 ± 0.010

Table 5.7: Overall model performance (mean ± standard deviation across 12 independent runs) for models trained on both IEEE and STANDUP datasets without colormap-harmonization.

Model	Domain	Accuracy	Precision	Recall	F1	Specificity	AUC
ResNet18	IEEE	0.939 ± 0.012	0.958 ± 0.010	0.958 ± 0.008	0.958 ± 0.009	0.889 ± 0.015	0.973 ± 0.011
	STANDUP	0.986 ± 0.010	0.978 ± 0.009	1.000 ± 0.000	0.989 ± 0.008	0.962 ± 0.012	0.980 ± 0.010
EfficientNet-B0	IEEE	0.924 ± 0.011	0.922 ± 0.012	0.979 ± 0.007	0.949 ± 0.008	0.778 ± 0.020	0.962 ± 0.009
	STANDUP	0.979 ± 0.009	0.967 ± 0.008	1.000 ± 0.000	0.983 ± 0.007	0.942 ± 0.010	0.970 ± 0.008
ShuffleNetV2-1x	IEEE	0.939 ± 0.010	0.958 ± 0.011	0.958 ± 0.009	0.958 ± 0.008	0.889 ± 0.014	0.988 ± 0.010
	STANDUP	0.971 ± 0.009	0.967 ± 0.010	0.989 ± 0.006	0.978 ± 0.007	0.942 ± 0.012	0.966 ± 0.008
MobileNetV3-Small	IEEE	0.924 ± 0.011	0.922 ± 0.012	0.979 ± 0.007	0.949 ± 0.008	0.778 ± 0.016	0.988 ± 0.009
	STANDUP	0.993 ± 0.007	0.989 ± 0.006	1.000 ± 0.000	0.994 ± 0.005	0.981 ± 0.008	0.990 ± 0.006
GhostNet_100	IEEE	0.879 ± 0.013	0.935 ± 0.012	0.896 ± 0.010	0.915 ± 0.008	0.833 ± 0.011	0.973 ± 0.009
	STANDUP	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

Table 5.8: Domain-specific model performance (mean ± standard deviation across 12 independent runs) for models trained on both IEEE and STANDUP datasets without colormap-harmonization.

Despite the high performance without colormap-harmonization, Grad-CAM visualizations of these tests revealed that the models (with the exception of ShuffleNetV2-1x) were overfitting on background information, with EfficientNet-B0, MobileNetV3-Small and ResNet18 overfitting significantly more than the tests with the harmonized datasets. GhostNet_100 demonstrated partial focus on background information across the Grad-CAM visualizations in the STANDUP domain, a pattern that was previously absent in the colormap-harmonized studies (see Figure 14). This resonates with the perfect scores (100%) across all metrics in the STANDUP domain. GhostNet_100 also demonstrated much worse performance in the IEEE domain without colormap-harmonization. Although ShuffleNetV2-1x did not display overfitting in the Grad-CAM results, its results demonstrate that it is suited for small dataset specific tasks and does not generalize as well as GhostNet_100 when dataset-harmonization is used (see Table 3).

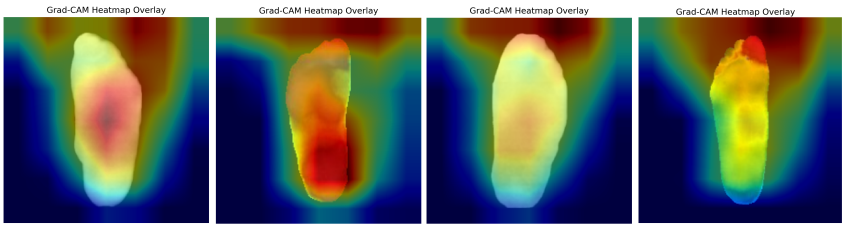

Figure 5.5: Grad-CAM outputs of EfficientNet-B0 in second ablation study.

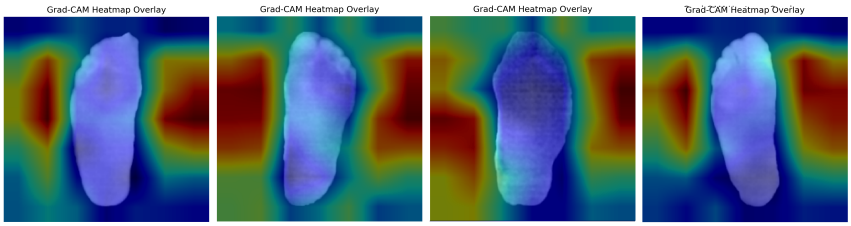

Figure 5.6: Grad-CAM outputs of MobileNetV3-Small in second ablation study.

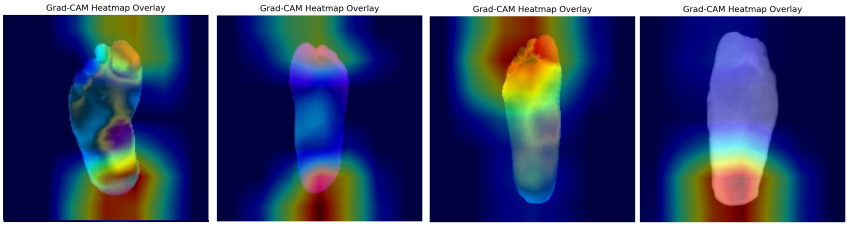

Figure 5.7: Grad-CAM outputs of ResNet18 in second ablation study.

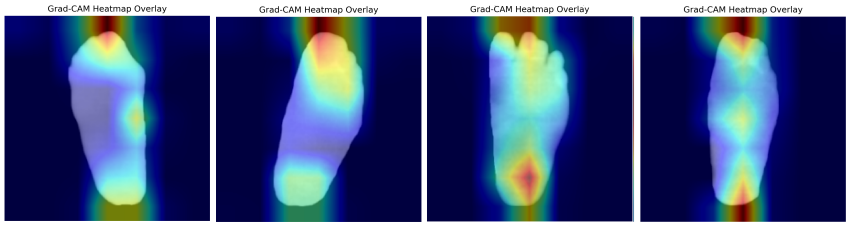

Figure 5.8: Grad-CAM outputs of GhostNet_100 in second ablation study.

This ablation study also demonstrates that high performance metrics do not always resonate with correct results as models may learn dataset specific features (as seen in the Grad-CAM visualisations above) and why it is important to incorporate Grad-CAMs in clinical studies.

5.3 Comparison with existing research

Compared to existing literature, our proposed pipeline using transfer learning with GhostNet_100 demonstrates competitive performance in binary DFU classification while maintaining a compact and efficient architecture suitable for handheld deployment. On the IEEE dataset, GhostNet_100 achieved an F1-score of 95.7% and a specificity of 91.1%. Grad-CAM evaluation of our pipeline confirmed that, when trained on harmonized datasets, the model consistently focused on physiologically relevant plantar regions, avoiding overfitting to background artifacts. This highlights the importance of cross-dataset evaluation, as training solely on one small dataset can yield misleadingly high performance (as demonstrated in the second ablation study). By contrast, our multi-dataset evaluation framework provides stronger evidence of generalizability, where GhostNet_100 maintained high performance on the independent STANDUP dataset, achieving 98.8% accuracy. This cross-dataset consistency, combined with the model's computational efficiency (5.2M parameters, 0.14 GFLOPs), demonstrates a practical balance between accuracy, interpretability, and real-world deployability.

Study	Task	Proposed Model	Accuracy	Recall	Precision	F1	Specificity	AUC
[16]	Binary	AdaBoost + RF	96.71%	96.71%	96.70%	96.70%	94.58%	–
		MobileNetV2	95.81%	95.81%	95.84%	95.82%	94.24%	–
[17]	Binary	ConvNeXt XLarge + Fancy PCA + LR	100%	100%	100%	100%	100%	100%
[19]	Binary	SURF-BOF + SVM	97.81%	97.81%	97.90%	97.93%	97.82%	99.95%
[22]	Binary	SVM (Texture & Temp)	95.61%	96.50%	–	–	92.41%	–
[23]	Binary	Custom CNN	97.00%	97.00%	82.30%*	89.10%	95.80%	97.60%
[25]	Binary	SVM (RBT Features)	92.86%	96.55%	93.33%	94.92%	84.62%	–
		KNN (RBT Features)	85.71%	89.66%	87.50%	88.57%	76.92%	–
[26]	Binary	Hybrid ORB+DL → DNN	98.51%	100%	98.98%	98.97%	98.98%	1.00
This study	Binary (IEEE subset)	GhostNet_100	**93.2%**	**95.8%**	**95.6%**	**95.7%**	**91.1%**	**95.7%**

Table 5.9: Model performance comparison of existing research vs proposed GhostNet_100 on the IEEE dataset. "*" indicates values calculated from other provided metrics.

Study	Task	Model Variant	Accuracy	Recall	Precision	F1	Specificity	AUC
[24]	Binary	ResNet-50	98.55%	–	–	–	–	–
[27]	Binary (STANDUP Subset)	Feature Selection → SVM	91.46%	96.08%	86.04%	90.70%	–	–
This study	Binary (STANDUP Subset)	GhostNet_100	**98.8%**	**99.9%**	**98.6%**	**99.2%**	**97.5%**	**99.9%**

Table 5.10: Performance comparison of existing research vs proposed GhostNet_100 on the STANDUP dataset.

Study	Task	Model Variant	Accuracy	Recall	Precision	F1	Specificity	AUC
[27]	Binary (IEE+STANDUP)	Feature Selection → SVM	86.19%	92.50%	79.31%	85.13%	–	–
This study	Binary (IEEE+STANDUP)	GhostNet_100	**96.0%**	**97.9%**	**97.1%**	**97.5%**	**94.3%**	**97.8%**

Table 5.11: Performance comparison of existing research vs proposed GhostNet_100 based pipeline on combined datasets (IEEE+STANDUP).

Model	Params (M)	FLOPs (G)	Ref
ResNet-50	25.56	4.09	[34]
DenseNet-201	20.01	4.29	[35]
DenseNet-121	7.98	2.83	[36]
Inception-v3	27.16	5.71	[37]
GoogLeNet (v1)	6.62	1.5	[38]
VGG-19	143.67	19.63	[39]
MobileNetV2	3.5	0.3	[40]
ConvNeXt-XL	350	60.9	[41]
AlexNet	61.1	0.71	[42]
GhostNet_100	**5.2**	**0.14**	**[31]**

Table 5.12: Comparison of efficiency between CNN models used in binary classification of early diabetic foot (those reported).

5.4 Theoretical Feasibility Analysis with Mobile Devices for GhostNet_100

5.4.1 Overview

The FLIR A600 used to capture the IEEE dataset must be connected to a separate host computer for operation and therefore can utilize heavyweight CNNs. Models such as the FLIR A500f can perform on-device classification but are cost-prohibitive for low-resource settings. Because we aim for a cost-effective pipeline suitable for deployment in limited resource environments like rural and underdeveloped areas, we adopt the FLIR ONE Pro adapter. As noted above, the STANDUP dataset uses the FLIR ONE Pro attached to a mobile phone [15]. In this section we analyze the feasibility of deploying GhostNet_100 with the FLIR ONE Pro adapter and the class of mobile devices required for real-time inference.

5.4.2 Existing metrics

In GhostNet_100's original CVPR report, the authors measured GhostNet_100 running inference in approximately 40ms on an ARM-based smartphone (TensorFlow Lite, single-thread) [31], corresponding to roughly 25 frames per second (FPS). Thus GhostNet_100 achieves about 25 FPS on mobile hardware circa 2019, outperforming comparable models such as MobileNetV3, which required 45 ms.

5.4.3 Current Advancements

Recent mobile deep-learning runtimes have substantially accelerated CNN inference. For example, a study [43] reports that MobileNetV2 latency on CPU was reduced from 203 ms to 21.9 ms on a high-end 2021 smartphone after runtime optimizations. As GhostNet_100 requires roughly half the FLOPs of MobileNetV2 (142M vs. 300M), a similar improvement implies an estimated latency of 10–15 ms (66–100 FPS) on comparable hardware. In practice, modern flagship SoCs (e.g., Snapdragon 8xx, Exynos 10xx, Apple A13/A14 and later) commonly enable GhostNet_100 to exceed 30–50FPS. Even on a single CPU core the model is lightweight;

multi-threading or GPU/NPU acceleration further increases throughput.

5.4.4 Conclusion on feasibility

GhostNet_100's 25 FPS baseline already exceeds the FLIR One Pro's thermal frame rate (8.7 Hz as stated on the official website) [44]. Therefore, a smartphone classifier using GhostNet_100 can readily keep pace with the incoming thermal stream. Even midrange SoCs (e.g., Snapdragon 712) should deliver well over the 9 FPS required. In summary, GhostNet_100 is feasible for FLIR One Probased thermal classification: it attains tens of FPS on typical 2018–2022 smartphones [31], [43], comfortably above the camera's 9-FPS output. Our theoretical analysis indicates that GhostNet_100 is expected to exceed the FLIR ONE Pro frame rate on midrange and flagship devices; however, on-device validation is required to confirm real-world FPS, latency, and thermal behavior.

5.5 Proposed Model and Pipeline

The proposed pipeline leverages GhostNet_100 as the main classification model while emphasizing methodological rigor and practical deployment. It uses a balanced composite metric for model training and integrates dataset standardization, controlled augmentation, and patient-level splitting to prevent data leakage, ensuring that the model learns clinically relevant features rather than dataset-specific artifacts. Grad-CAM evaluation confirms that the pipeline guides the model to focus on physiologically meaningful plantar regions, reducing background overfitting.

Model	Domain	Accuracy	Precision	Recall	F1	Specificity	AUC
GhostNet_100	IEEE+STANDUP	0.960 ± 0.014	0.971 ± 0.015	0.979 ± 0.009	0.975 ± 0.009	0.943 ± 0.039	0.978 ± 0.010
GhostNet_100	IEEE	0.932 ± 0.021	0.956 ± 0.021	0.958 ± 0.015	0.957 ± 0.014	0.911 ± 0.062	0.957 ± 0.019
GhostNet_100	STANDUP	0.988 ± 0.006	0.986 ± 0.008	0.999 ± 0.003	0.992 ± 0.004	0.975 ± 0.016	0.999 ± 0.001

Table 5.13: Mean performance metrics of proposed pipeline utilizing GhostNet_100 across 12 independent runs with different seeds.

By framing GhostNet_100 within this pipeline, the methodology ensures a balance of robust performance, interpretability, and computational efficiency suitable for mobile deployment. While other architectures (e.g., ResNet18, ShuffleNetV2-1x) achieve competitive performance, Grad-CAM visualizations revealed overfitting or some dependence on background regions, highlighting the risk of dataset-specific overfitting. In contrast, the proposed pipelines combination of person-wise splitting, controlled augmentation, and explainability checks strengthens model reliability and reproducibility.

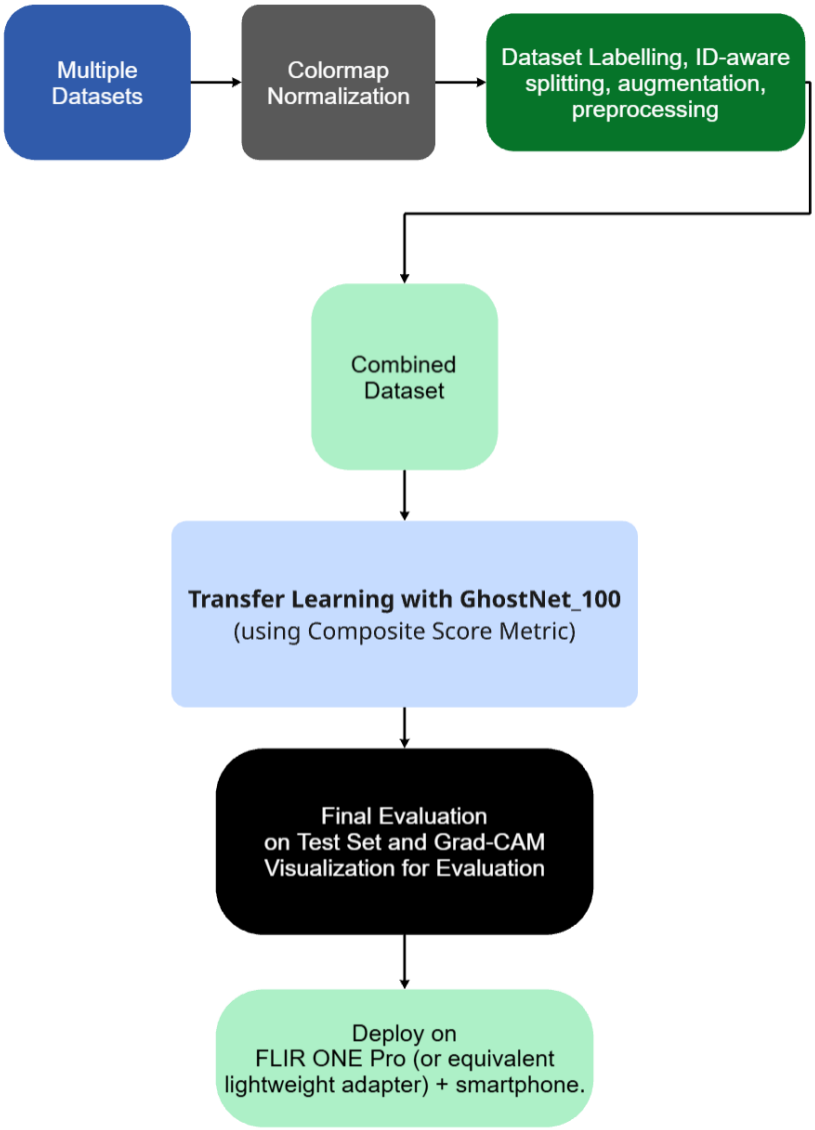

Figure 5.9: Overview of the proposed classification pipeline using GhostNet_100 with transfer learning.

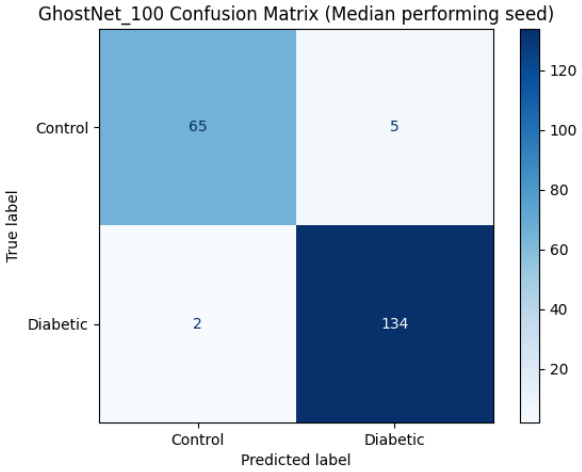

Figure 5.10: Confusion matrix for the median-performing model over 12 independent runs of GhostNet_100 using the proposed classification pipeline.

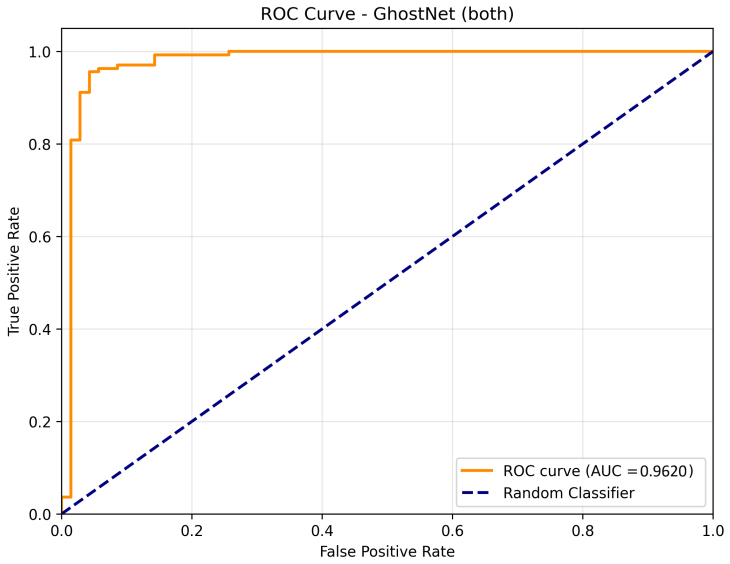

Figure 5.11: ROC curve averaged via interpolated TPR for GhostNet_100 with the proposed pipeline.

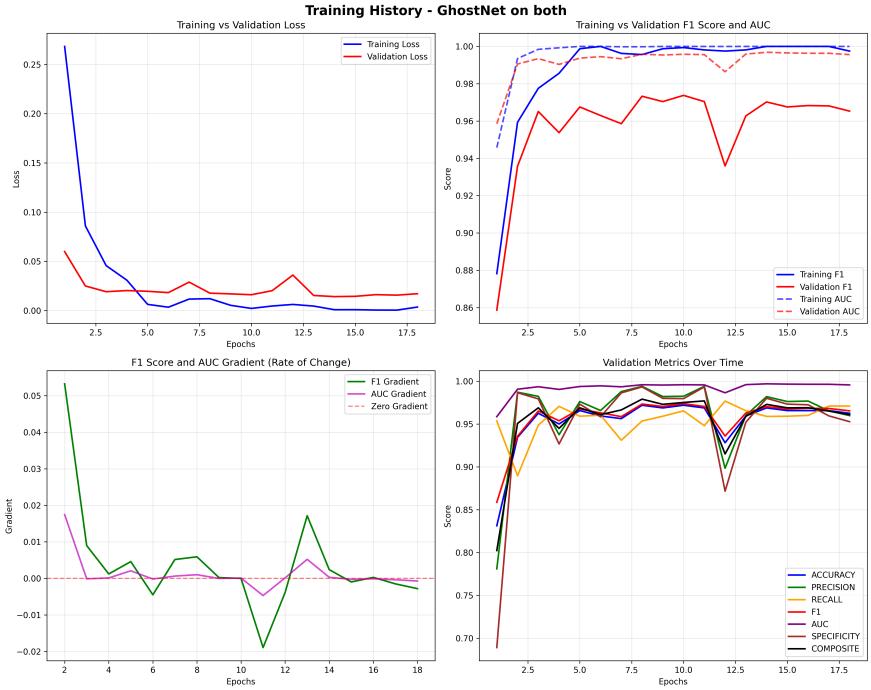

Figure 5.12: Mean training history for GhostNet_100 within the proposed pipeline.

5.6 Discussions

The experimental results demonstrate the effectiveness of our comprehensive pipeline in developing clinically viable models for diabetic foot classification. The adoption of a composite scoring metric ($\alpha = 0.5, \beta = 0.3, \gamma = 0.2$) proved crucial in aligning model selection with clinical priorities, effectively balancing high sensitivity with controlled false-positive rates.

GhostNet_100 emerged as the optimal architecture, achieving superior performance across both IEEE (F1: 95.7%, specificity: 91.1%) and STANDUP (F1: 99.2%, specificity: 97.5%) datasets while maintaining exceptional computational efficiency (5.2M parameters, 0.14 GFLOPs), making it ideal for mobile deployment in resource-limited settings.

The ablation studies revealed important insights: single-dataset training degraded performance, highlighting the need for cross-dataset training, while training without colormap harmonization yielded high metrics but produced unreliable models that overfitted to background artifacts. This emphasizes that performance metrics alone are insufficient indicators of clinical viability.

Grad-CAM analysis proved invaluable for validation, revealing crucial differences in model behavior. EfficientNet-B0, ShuffleNetV2-1x, and GhostNet_100 maintained focused attention on plantar regions, while MobileNetV3-Small and ResNet18 demonstrated concerning background overfitting. These visualizations confirm that model decisions are based on clinically relevant thermal patterns.

Additionally, our study achieved the highest results in comparison to studies on the combined IEEE+STANDUP datasets (see table 5.11).

The feasibility analysis shows GhostNet_100's computational profile comfortably exceeds the FLIR One Pro's 9 FPS capture rate, enabling real-time mobile deployment for point-of-care screening.

Our findings indicate that methodological rigor, including person-level splitting, controlled augmentation, dataset standardization, and composite metric optimizationis as critical as architectural selection for developing reliable clinical models. The combination of GhostNet_100 with our framework represents a promising approach for accessible, interpretable, and clinically reliable diabetic foot screening systems.

Chapter 6

Conclusion

6.1 Summary of Findings

This study successfully developed and evaluated a scalable deep learning pipeline for the early detection of diabetic foot complications using plantar thermography. The primary findings are as follows:

- Developed and evaluated a scalable deep learning pipeline for early detection of diabetic foot complications using plantar thermography.
- Integrated and standardized the high-specification IEEE dataset with the cost-effective STANDUP dataset using a novel colormap-harmonization pipeline to mitigate domain shift.
- Trained five lightweight CNN architectures (ResNet18, EfficientNet-B0, MobileNetV3-Small, ShuffleNetV2-1x, and GhostNet_100) using pretrained weights on the combined thermal dataset.
- Demonstrated that GhostNet_100 achieved the highest diagnostic performance on the combined test set, with mean values of 96.0% accuracy, 97.1% precision, 97.9% recall, 97.5% F1-score, and 94.3% specificity.
- Showed that multi-dataset training improved generalization and reduced overfitting compared to single-dataset training.
- Verified through Grad-CAM visualizations that GhostNet_100 consistently focused on clinically relevant plantar regions rather than background artifacts.
- Confirmed via ablation studies the effectiveness of dataset harmonization and pooling in improving generalization and preventing overfitting.
- Demonstrated high computational efficiency of GhostNet_100 (5.2M parameters, 0.14 GFLOPs), supporting its suitability for real-time mobile deployment with low-cost thermal cameras.

In summary, this research presents a reliable, efficient, and generalizable solution that addresses key limitations of previous studies, namely dataset scarcity, lack of model explainability, and high computational costs.

6.2 Contributions to the Field

This research makes several significant contributions to the field of AI-based diabetic foot detection:

- Outperforms previous methods, achieving new state-of-the-art results in combined dataset study.
- Proposes a novel pipeline that integrates and standardizes multiple plantar thermography datasets, addressing key challenges in domain generalization.
- Demonstrates that efficient architectures such as GhostNet_100 achieve strong diagnostic performance while remaining feasible for mobile deployment in resource-constrained settings.
- Enforces person-wise data splitting via ID tracking to prevent data leakage and ensure reliable evaluation.
- Utilizes a composite scoring metric to balance sensitivity and specificity, supporting clinically relevant model selection.
- Validates model predictions through systematic Grad-CAM analysis, confirming reliance on physiologically relevant features and enhancing explainability and clinical trustworthiness.
- Demonstrates the effective pairing of a low-cost thermal camera adapter with an efficient deep learning model, outlining a feasible pathway for early diabetic foot screening in primary-care and low-resource environments.

6.3 Recommendations for Future Work

While this study presents a significant step forward, several avenues remain for future research to build upon these findings:

1. **Expansion to Multi-Class Classification:** The current work is limited to binary classification (Diabetic vs. Control). Future studies should develop models capable of classifying the severity of diabetic foot complications (e.g., neuropathy, ischemia, ulceration stages) to provide more specific diagnostic information to clinicians.
2. **Incorporation of Clinical Metadata:** Integrating patient metadata such as HbA1c levels, age, sex, and duration of diabetes with the thermal images could create a multi-modal model, potentially improving predictive accuracy and clinical relevance.
3. **Prospective Clinical Validation:** The proposed pipeline requires validation in a prospective clinical trial with a larger and more diverse cohort. This would test its real-world efficacy, refine its performance, and establish clinical benchmarks.

4. **On-Device Deployment and Testing:** The theoretical feasibility of mobile deployment should be followed by practical implementation. Developing a smartphone application and conducting real-world latency and usability tests on various mobile devices would be a logical next step.

5. **Proper Labeling of Special Cases:** Some diabetic patients exhibit healthy feet while some healthy people exhibit plantar vasculopathy and neuropathy. Proper labelling for these special cases should be done within datasets to help models to be more stable and flexible.

6. The proposed color conversion pipeline offers a reusable framework for standardizing thermal datasets with colormap heterogeneity. It can be utilized in future work involving clinical thermography studies.

Bibliography

[1] P. Zimmet, K. G. M. M. Alberti, D. J. Magliano, and P. H. Bennett, "Diabetes mellitus statistics on prevalence and mortality: Facts and fallacies," *Nature Reviews Endocrinology*, vol. 12, no. 10, pp. 616–622, 2016. DOI: 10.1038/nrendo.2016.105 [Online]. Available: https://doi.org/10.1038/nrendo.2016.105

[2] S. Tesfaye et al., "Diabetic neuropathies: Update on definitions, diagnostic criteria, estimation of severity, and treatments," *Diabetes Care*, vol. 33, no. 10, pp. 2285–2293, 2010. DOI: 10.2337/dc10-1303 [Online]. Available: https://doi.org/10.2337/dc10-1303

[3] E. B. Jude, I. Eleftheriadou, and N. Tentolouris, "Peripheral arterial disease in diabetesa review," *Diabetic Medicine*, vol. 27, no. 1, pp. 4–14, 2010. DOI: 10.1111/j.1464-5491.2009.02866.x [Online]. Available: https://doi.org/10.1111/j.1464-5491.2009.02866.x

[4] A. Hingorani et al., "The management of diabetic foot: A clinical practice guideline by the society for vascular surgery in collaboration with the american podiatric medical association and the society for vascular medicine," *Journal of Vascular Surgery*, vol. 63, no. 2, Suppl. 3S–21S, 2016. DOI: 10.1016/j.jvs.2015.10.003 [Online]. Available: https://doi.org/10.1016/j.jvs.2015.10.003

[5] A. J. M. Boulton, L. Vileikyte, G. Ragnarson-Tennvall, and J. Apelqvist, "The global burden of diabetic foot disease," *The Lancet*, vol. 366, no. 9498, pp. 1719–1724, 2005. DOI: 10.1016/S0140-6736(05)67698-2 [Online]. Available: https://doi.org/10.1016/S0140-6736(05)67698-2

[6] D. G. Armstrong, A. J. M. Boulton, and S. A. Bus, "Diabetic foot ulcers and their recurrence," *The New England Journal of Medicine*, vol. 376, no. 24, pp. 2367–2375, 2017. DOI: 10.1056/NEJMra1615439 [Online]. Available: https://doi.org/10.1056/NEJMra1615439

[7] F. L. Game et al., "Effectiveness of interventions to enhance healing of chronic ulcers of the foot in diabetes: A systematic review," *Diabetes/Metabolism Research and Reviews*, vol. 32, no. Suppl. 1, pp. 154–168, 2016. DOI: 10.1002/dmrr.2707 [Online]. Available: https://doi.org/10.1002/dmrr.2707

[8] D. G. Armstrong, T. Tan, A. J. M. Boulton, and S. A. Bus, "Diabetic foot ulcers," *JAMA*, vol. 330, no. 1, p. 62, 2023. DOI: 10.1001/jama.2023.10578 [Online]. Available: https://doi.org/10.1001/jama.2023.10578

[9] J. J. Van Netten et al., "Prevention of foot ulcers in the atrisk patient with diabetes: A systematic review," *Diabetes/Metabolism Research and Reviews*, vol. 32, no. S1, pp. 84–98, 2015. DOI: 10.1002/dmrr.2701 [Online]. Available: https://doi.org/10.1002/dmrr.2701

[10] T. Mori et al., "Morphological pattern classification system for plantar thermography of patients with diabetes," *Journal of Diabetes Science and Technology*, vol. 7, no. 5, pp. 1102–1112, 2013. DOI: 10.1177/193229681300700502 [Online]. Available: https://doi.org/10.1177/193229681300700502

[11] L. Wu, R. Huang, X. He, L. Tang, and X. Ma, "Advances in machine learning-aided thermal imaging for early detection of diabetic foot ulcers: A review," *Biosensors*, vol. 14, p. 614, 2024. DOI: 10.3390/

[12] O. Kurkela, J. Lahtela, M. Arffman, and L. Forma, "Infrared thermography compared to standard care in the prevention and care of diabetic foot: A cost analysis utilizing real-world data and an expert panel," *ClinicoEconomics and Outcomes Research*, vol. 15, pp. 111–123, 2023. DOI: 10.2147/ceor.s396137 [Online]. Available: https://doi.org/10.2147/ceor.s396137

[13] D. Hernández-Contreras, H. Peregrina-Barreto, J. Rangel-Magdaleno, and F. Renero-Carrillo, *Plantar thermogram database for the study of diabetic foot complications*, Data formats: *.png, *.csv, *.xlsx, *.txt, *.m, IEEE Dataport, 2019. DOI: 10.21227/tm4t-9n15 [Online]. Available: https://dx.doi.org/10.21227/tm4t-9n15

[14] D. A. Hernandez-Contreras, H. Peregrina-Barreto, J. De Jesus Rangel-Magdaleno, and F. Orihuela-Espina, "Statistical approximation of plantar temperature distribution on diabetic subjects based on beta mixture model," *IEEE Access*, vol. 7, pp. 28 383–28 391, 2019. DOI: 10.1109/access.2019.2902502 [Online]. Available: https://doi.org/10.1109/access.2019.2902502

[15] D. Bouallal, A. Bougrine, R. Harba, and et al., "Standup database of plantar foot thermal and rgb images for early ulcer detection," *Open Research Europe*, vol. 2, p. 77, 2022, Version 1; peer review: 1 approved, 1 approved with reservations. DOI: 10.12688/openreseurope.14706.1 [Online]. Available: https://doi.org/10.12688/openreseurope.14706.1

[16] A. Khandakar et al., "A machine learning model for early detection of diabetic foot using thermogram images," *Computers in Biology and Medicine*, vol. 137, p. 104 838, 2021. DOI: 10.1016/j.compbiomed.2021.104838 [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.104838

[17] O. Fasihi-Shirehjini and F. Babapour-Mofrad, "Effectiveness of convnext variants in diabetic feet diagnosis using plantar thermal images," *Quantitative InfraRed Thermography Journal*, pp. 1–18, 2024. DOI: 10.1080/17686733.2024.2310794 [Online]. Available: https://doi.org/10.1080/17686733.2024.2310794

[18] S. Muralidhara, A. Lucieri, A. Dengel, and S. Ahmed, "Holistic multi-class classification & grading of diabetic foot ulcerations from plantar thermal images using deep learning," *Health Information Science and Systems*, vol. 10, no. 1, 2022. DOI: 10.1007/s13755-022-00194-8 [Online]. Available: https://doi.org/10.1007/s13755-022-00194-8

[19] M. H. Alshayeji, S. C. Sindhu, and S. Abed, "Early detection of diabetic foot ulcers from thermal images using the bag of features technique," *Biomedical Signal Processing and Control*, vol. 79, p. 104 143, 2022. DOI: 10.1016/j.bspc.2022.104143 [Online]. Available: https://doi.org/10.1016/j.bspc.2022.104143

[20] V. Filipe, P. Teixeira, and A. Teixeira, "Automatic classification of foot thermograms using machine learning techniques," *Algorithms*, vol. 15, no. 7, p. 236, 2022. DOI: 10.3390/a15070236 [Online]. Available: https://doi.org/10.3390/a15070236

[21] I. Cruz-Vega, D. Hernandez-Contreras, H. Peregrina-Barreto, J. De Jesus Rangel-Magdaleno, and J. M. Ramirez-Cortes, "Deep learning classification for diabetic foot thermograms," *Sensors*, vol. 20, no. 6, p. 1762, 2020. DOI: 10.3390/s20061762 [Online]. Available: https://doi.org/10.3390/s20061762

[22] J. Saminathan, M. Sasikala, V. Narayanamurthy, K. Rajesh, and R. Arvind, "Computer aided detection of diabetic foot ulcer using asymmetry analysis of texture and temperature features," *Infrared Physics & Technology*, vol. 105, p. 103 219, 2020. DOI: 10.1016/j.infrared.2020.103219 [Online]. Available: https://doi.org/10.1016/j.infrared.2020.103219

[23] I. Khosa, A. Raza, M. Anjum, W. Ahmad, and S. Shahab, "Automatic diabetic foot ulcer recognition using multi-level thermographic image data," *Diagnostics*, vol. 13, no. 16, p. 2637, 2023. DOI: 10.3390/diagnostics13162637 [Online]. Available: https://doi.org/10.3390/diagnostics13162637

[24] A. R. Singh, S. R. Durgesh, and A. V. Vibro, "Early detection of diabetic foot ulcer with multi-modal images using federated learning," in *Proceedings of the International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, Erode, India, 2025, pp. 964–969. DOI: 10.1109/ICMSCI62561.2025.10894410 [Online]. Available: https://doi.org/10.1109/ICMSCI62561.2025.10894410

[25] S. Jayapal, N. B. Murugesan, and S. Mohan, "Detection of diabetic foot using statistical features," *Diabetic Foot - Recent Advances*, 2022. DOI: 10.5772/intechopen.106457 [Online]. Available: https://doi.org/10.5772/intechopen.106457

[26] A. S. Eldin, A. S. Ahmoud, H. M. Hamza, and H. Ardah, "Enhancing early detection of diabetic foot ulcers using deep neural networks," *Diagnostics*, vol. 15, p. 1996, 2025. DOI: 10.3390/diagnostics15161996 [Online]. Available: https://doi.org/10.3390/diagnostics15161996

[27] N. Arteaga-Marrero, A. Hernandez-Guedes, J. Ortega-Rodriguez, and J. Ruiz-Alzola, "State-of-the-art features for early-stage detection of diabetic foot ulcers based on thermograms," *Biomedicines*, vol. 11, p. 3209, 2023. DOI: 10.3390/biomedicines11123209 [Online]. Available: https://doi.org/10.3390/biomedicines11123209

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[29] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97, PMLR, 2019, pp. 6105–6114. DOI: 10.48550/arXiv.1905.11946 [Online]. Available: https://doi.org/10.48550/arXiv.1905.11946

[30] A. Howard et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/papers/Howard_Searching_for_MobileNetV3_ICCV_2019_paper.pdf

[31] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1580–1589. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Han_GhostNet_More_Features_From_Cheap_Operations_CVPR_2020_paper.html

[32] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. DOI: 10.48550/arXiv.1807.11164 [Online]. Available: https://doi.org/10.48550/arXiv.1807.11164

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74 [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

[34] PyTorch Foundation. "Resnet50 torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html

[35] PyTorch Foundation. "Densenet201 torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.densenet201.html

[36] PyTorch Foundation. "Densenet121 torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.densenet121.html

[37] PyTorch Foundation. "Inception v3 torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.inception_v3.html

[38] PyTorch Foundation. "Googlenet torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.googlenet.html

[39] PyTorch Foundation. "Vgg19 torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.vgg19.html

[40] PyTorch Foundation. "Mobilenet v2 torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.mobilenet_v2.html

[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. "A convnet for the 2020s." Accessed: 2025-10-25. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.pdf

[42] PyTorch Foundation. "Alexnet torchvision models documentation." Accessed: 2025-10-25. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.alexnet.html

[43] Q. Zhang et al., "A comprehensive benchmark of deep learning libraries on mobile devices," in *Proceedings of the ACM Web Conference 2022 (WWW '22)*, 2022, pp. 1–10. DOI: 10.1145/3485447.3512148 [Online]. Available: https://xumengwei.github.io/files/WWW22-MobileDLLibs.pdf

[44] F. Systems, *Flir one pro thermal imaging camera for smartphones*, Accessed: 2025-10-26, 2025. [Online]. Available: https://www.flir.com/products/flir-one-pro/#:~:text=Adjustable%20MSX%20distance%200,EST%20Mode%20No%20Shutter%20Automatic%2FManual