Deep Learning Strategies for Diabetic Foot Ulcer Diagnosis: An Empirical Comparison of Transformer and CNN Architectures

Sakshi R[1], *Thiriloksha* S[2], and *Tephillah* S[3]

[1]Department of ECE St. Joseph's Institute of Technology Chennai, Tamil Nadu, India sakshiravi464@gmail.com
[2]Department of ECE St. Joseph's Institute of Technology Chennai, Tamil Nadu, India thiriloksha@gmail.com
[3]Department of ECE St. Joseph's Institute of Technology Chennai, Tamil Nadu, India tephisam@gmail.com

Abstract. Diabetic Foot Ulcers (DFU) are among the most serious complications faced by diabetic patients. What usually starts as a small injury can worsen and in many cases lead to lower limb amputation. However with modern technologies like deep learning early screening and identification of DFU has become more achievable. This helps doctors and make their job easy. This study presents a practical comparison between ResNet18 a CNN and DeiT-Small a Vision Transformer. Both these models were tuned using transfer learning on datasets containing images of both DFU affected feet and healthy feet. There was a clear performance gap between both these models. DeiT-Small reached an impressive test accuracy of 99.17% while ResNet18 achieved 85.40% accuracy. To understand their performance better visualization tools like Grad-CAM and transformer attention maps were used. For future work, the study aims to create more consistent data splits, perform deeper architectural ablation and integrate IOT based sensors to support real time DFU monitoring.

1. Introduction

Diabetic Foot Ulcers (DFUs) are the most severe and costly complications of diabetes, affecting millions of people world- wide and contributing significantly to lower limb amputations rate [1]. Early and accurate diagnosis of DFU lesions is very important to initiate the timely intervention and improve the outcome [2]. Traditional and manual assessment methods are often subjective, as they are time consuming, and require specialized medical expertise which are very scarce in many resource limited settings. This increases the critical need for automated diagnostic tool assistance [3].

To face this challenge, a deep learning techniques have gained immense prominence in recent years, particularly Convolutional neural networks (CNNs), have emerged as powerful tools for automated medical image analysis [4]. Architectures like ResNet have historically excelled in DFU classification tasks. However, the recent introduction of the Vision Transformer (ViT) architecture, which takes global advantage of global context through self-attention mechanisms [5],[6]. The inherent differences in local feature extraction in CNN versus the global feature extraction in ViT suggests that one approach may be fundamentally better suited. The main objective of this study is to conduct an empirical comparison between CNN (ResNet18) and ViT (DeiT-Small) models for image classification, analyzing accuracy, precision and recall to identify the superior deep learning algorithm.

2. Related Work and Background

Deep learning for detecting diabetic foot ulcers has evolved swiftly, with majority of its approaches technically centered on CNN [1]. Initially, transfer learning became the go-to strategy to tackle the limited data in medical imaging. This often involved using ImageNet pre-trained models with CNN backbones like VGG, Inception and ResNet. For instances ResNet based models for specific tasks managed to achieve classification accuracy between 95% to 98% by effectively capturing local features of ulcers, such as wound boundaries, exudate and surrounding erythema as noted in real-time DFU classification systems [7].

The introduction of the transformer architecture to computer vision in 2020 marked a exemplar shift. Unlike CNNs, which use local convolutional filters, ViTs process images as sequence of patches and global contextual relationships across the entire image [8]. This shift has rapidly influenced medical image analysis, with models integrating on ViTs for tasks ranging from retinopathy detection [9] to chemotherapy cardiotoxicity prediction [10].

Even though both CNN and ViT have theoretical advantages, there is still a debate about which model works better for very little training data. Unlike CNN, ViTs need large datasets because they don't have built-in knowledge. So they learn everything from raw data [6]. Data-efficient Image Transformers (DeiT), a special version of ViT, can learn effectively using less data itself

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

and can spot tiny, spread-out visual signs, which helps to detect diabetic foot ulcers at an early stage [5], [6]. Recent studies show that pre-trained ViTs on huge datasets can perform very well on medical analysis because they generalize better and give more stable results [6]. This work extends prior comparative studies by specifically analyzing and comparing the performance of CNNs and ViTs on a diabetic foot ulcer dataset.

3. Methodology and Implementation

This section explains how the experiment was carried out, the datasets used, how the images were processed, what models were trained, and what evaluation measures were used.

3.1. Dataset and Preprocessing

A custom image dataset containing photos of both diabetic and non-diabetic feet was used. The images are grouped into two classes: feet with visible diabetic foot ulcers and feet that appear normal. The dataset introduced demonstrates the importance of standardized DFU image collection [11]. This supports our preprocessing choices for consistent model performance. Fig. 1 shows example images from both groups to demonstrate the two classes.

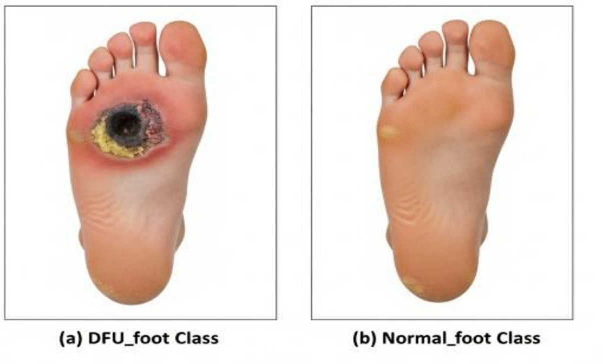

Fig. 1: Sample images from the dataset

3.1.1. Data Splitting: Two independent data-splitting strategies were used because the CNN and Vision Transformer pipelines were developed in different systems that process data differently.

3.1.1.1. CNN (ResNet18) Split: The dataset was split into 80% training, 10% validation, and 10% testing, resulting in 1052 training images, 226 validation images, and 226 test images, with the final test set being evenly balanced between DFU and normal classes. Stratification was necessary to avoid class imbalance effects, particularly because DFU recognition is clinically sensitive to false negatives.

3.1.1.2. ViT (DeiT-Small) Split: For the Vision Transformer experiment, the dataset was processed in a separate notebook implementation where an 80–20 data split strategy was used. Unlike the CNN pipeline, this split did not include a separate internal validation fold; instead, the validation and testing occurred over the same held-out portion. The final reported ViT testing was conducted on a 121-image set, consisting of 61 DFU images and 60 Normal images, maintaining balance between positive and negative clinical cases.

While both models were trained on the same original dataset, the resulting splits are not identical due to differences in preprocessing pipelines, augmentation libraries, and automated dataset handling logic built into each implementation. The decision was not intentional experimental bias but a practical consequence of developing the two models independently. Although both used different split strategies, both test sets had enough images and balanced classes, so the evaluation results were still fair and meaningful. Even though using two different splits is a limitation, the results are still useful, and future work will use a single split to allow more direct and accurate comparisons.

3.1.2. Image Preprocessing and Augmentation: All the input images were resized to a uniform dimension of 224×224 pixels. Normalization was applied to leverage the knowledge encoded in the pre-trained weights.

CNN augmentation using torchvision.transforms, a data augmentation pipeline was integrated to improve the generalization of the model. Transforms include resizing the image to 224×224 pixels, random cropping, horizontal flipping, and a random rotation of ±10°. In addition, Color Jitter is adopted to provide variations in brightness, contrast, saturation (0.2), and hue (0.05). Normalization is carried out on all images, considering standard ImageNet statistics: mean as 0.485, 0.456, 0.406, and std as 0.229,0.224, 0.225. The augmentation strategy increased variability in data, further reducing overfitting of the dataset.

Data augmentation was performed for the ViT model using the Albumentations library. The pipeline involved resizing to size 224 × 224, random horizontal flip (p =0.5), shift-scale-rotate transformation with respective limits of 0.1, 0.1, and 15° (p= 0.5), adding photometric variation with random brightness and contrast adjustments (p =0.2), with mean and std of (0.5, 0.5, 0.5). The normalization can be calculated using the following formula:

$$x_{norm} = \frac{x - \text{mean}}{\text{std}} \tag{1}$$

This ensures models converge steadily during the fine-tuning process for both CNN and ViT pipelines, augmentation strategies were also crafted to introduce variability without altering any key medical image features. The value of precise wound boundary extraction in DFU analysis supports the need for accurate preprocessing and model selection [12].

3.2. Deep Learning Architectures

Both models employed Transfer Learning using weights pre-trained on the large ImageNet dataset, a technique essential for medical imaging tasks with limited proprietary data. The pretrained networks extract highly discriminative DFU features. This supports our choice of transfer-learning based architectures for reliable DFU classification [13].

3.2.1. Convolutional Neural Network (CNN):

Backbone: ResNet18 was selected as it provides an effective compromise between representational capacity and computational efficiency.
Modification: The final fully connected layer (model.fc) was replaced with a customized binary classifier composed of a Linear layer (512→512), ReLU activation, Dropout (0.5), and a final Linear layer mapping to the two output classes (DFU foot and Normal foot).
Training Parameters: The model was trained using Cross- Entropy Loss and the Adam optimizer with an LR of ($1 \times 10{-}3$) and weight decay of ($1 \times 10{-}4$), using a batch size of 16 over 5 epochs.

3.2.2. Vision Transformer (ViT):

Backbone: DeiT-Small (Data-efficient Image Transformer) was chosen. DeiT is a high-performing ViT variant known for achieving competitive, making it highly effective for transfer learning on small datasets. The specific variant used was 'deit small patch16 224'.
Modification: The native classification head (model.head) was replaced with a simple linear layer mapping the ViT's feature dimension to the 2 output classes.
Training Parameters: The training setup employed Cross-Entropy Loss as the objective function and the Adam optimizer. The model was trained using a batch size of 16.

3.3. Evaluation Metrics

The model performance was evaluated on the basis of independent test sets using high standard metrics that are chosen for their clinical relevance:
Accuracy: Overall fraction of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, and FN denotes False Negatives.
Precision (Positive Value): The ratio of correctly predicted positive observations (TP) to the total predicted positive observations (TP + FP). It is essential for minimizing false alarms and clinical follow-ups.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall (Sensitivity): The ratio of correctly predicted positive observations (TP) to all observations in the actual positive class (TP + FN). It plays a crucial role in minimizing missed diagnoses.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

F1-Score: The weighted average of precision and recall.

$$F_{1\text{-Score}} = \frac{2 \cdot (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{5}$$

Confusion Matrix (CM): A visualization graph comparing the values of true labels to predicted labels.

4. Results and Analysis

This section presents the quantitative performance comparison, provides a qualitative analysis using Explainable AI (XAI), and discusses the real-world implications of the findings.

4.1. Quantitative Evaluation

The comparative evaluation clearly established the ViT architecture as the superior classifier for this DFU detection. In Table I, the key performance metrics are summarized.

Table 1 Performance Comparison of CNN (ResNet18) and ViT (DeiT-Small) Models

Metric	CNN (ResNet18)	ViT (DeiT-Small)
Test Samples	226	121
Overall Accuracy (%)	85.40	99.17
DFU Precision	0.78	0.98
DFU Recall	0.99	1.00
Normal Precision	0.99	1.00
Normal Recall	0.71	0.98
Macro Avg F1-Score	0.85	0.99

The results in Table 1 show that the DeiT-Small Vision Transformer performed much better with an accuracy of 99.17%. Whereas the ResNet18 achieved 85.40% accuracy on its 226 test images. This difference is also reflected in the learning curves shown in Fig. 2. The ViT model not only converges faster but also maintains a lower validation loss, indicating better generalization. On the other hand, the CNN shows a noticeable gap between its training and validation performance, indicating that it is less stable and more prone to overfitting.

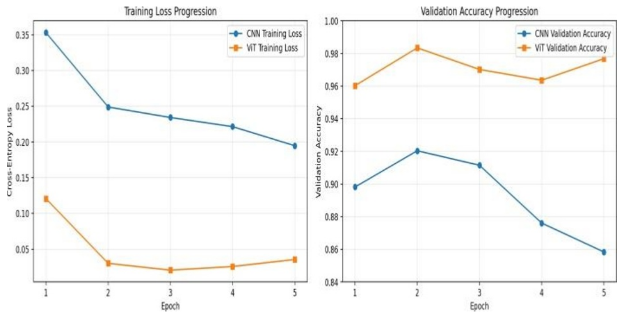

Fig. 2: Comparative training loss and validation accuracy curves showing over 5 epochs for the CNN and ViT models.

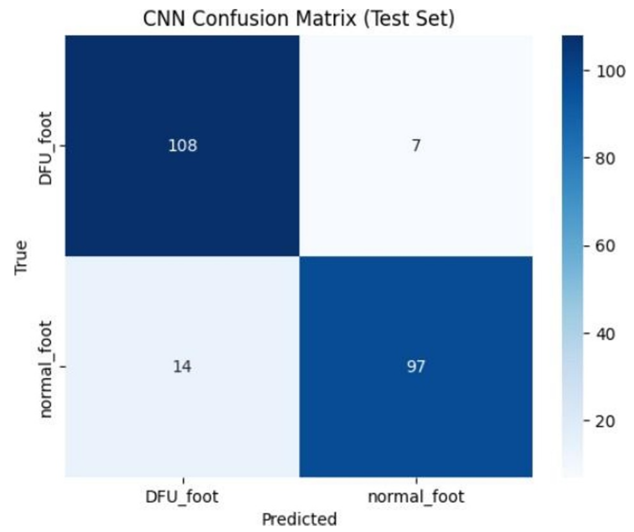

Fig. 3: Confusion matrix for the CNN model

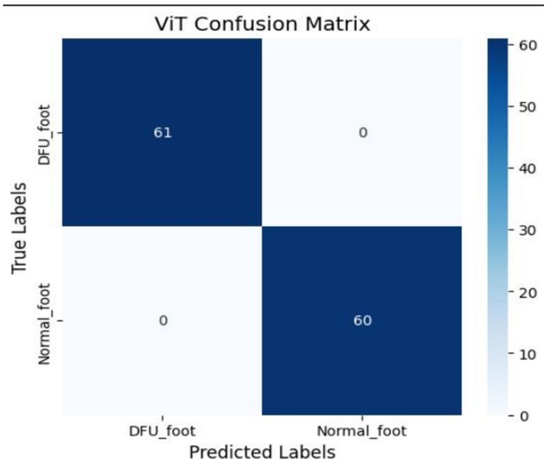

Fig. 4: Confusion matrix for the ViT model

The confusion matrices in Fig. 3 and Fig. 4 shows the performance difference between the two architectures. On a test set of 226 samples, the ResNet18 CNN was able to correctly classify 108 DFU cases and 97 normal cases but generated 14 false negatives and 7 false positives. Meanwhile, the DeiT- Small Vision Transformer, on a test set of 121 samples, 61 DFU cases and 60 normal cases were correctly identified, yielding zero false positives and only one false negative. In practical terms, the absence of false-positive and lower number of false-negative in transformer predictions approach indicates the much better reliability and suitability of the method for clinical practice in terms of DFU classification.

Finally, Fig. 5 and Fig. 6 show the ROC curves for both models. The ResNet18 CNN achieved an AUC of 0.990, showing high but not complete class separability. In comparison, the DeiT-Small Vision Transformer achieved an AUC of 1.000, indicating perfect discrimination between DFU and normal cases.

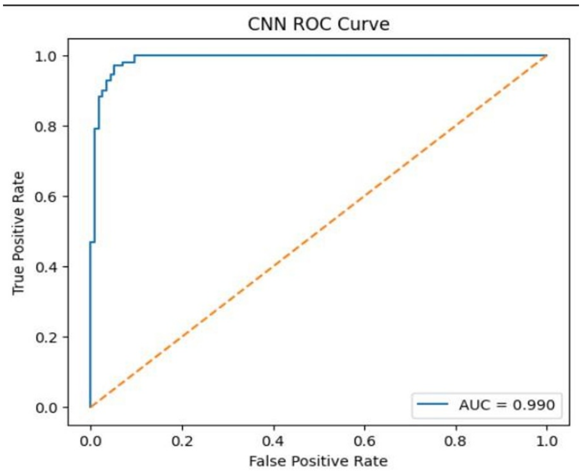

Fig. 5: Receiver operating Characteristics curve for CNN

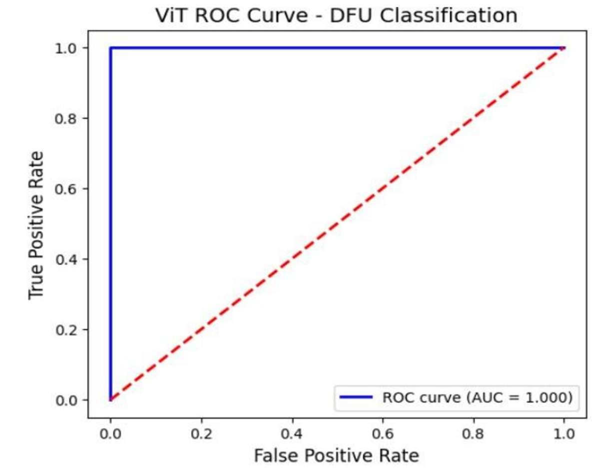

Fig. 6: Receiver operating characteristics curve for ViT

The ROC curve for the DeiT-Small models ascends sharply towards the upper left corner of the graph. This suggests that it can distinguish between positive and negative cases with high confidence, even as decision thresholds change. Its path is almost perfect, showing little overlap between class distributions. This matches with its AUC score of 1.000, the model has consistent and reliable decision boundaries throughout the evaluation range.

4.2. Qualitative Analysis

To gain insight into the decision-making behaviour of both models, explainability techniques were applied using Gradient- weighted Class Activation Mapping (Grad-CAM) for CNN and ViT, as shown in Fig. 7.

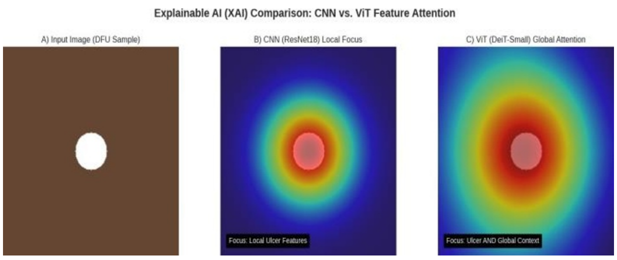

Fig. 7: Comparative visualization of feature importance. (a) Input image (b) CNN local focus (c) ViT global focus

The Grad-CAM visualization for the CNN, shown in (Fig. 7(b)) highlights a strong activation around the central ulcer area. This pattern matches the design of convolutional networks, which prioritize more on local feature extraction. On the other hand attention roll out map for the DeiT- Small model in (Fig. 7(c)) displays a wider global focus feature extraction technique, while still paying a significant and keen attention to the ulcer area, the vision transformer also considers surrounding tissue features and the overall foot structure. This undoubtedly proves that the ViT captures a very long-range spatial relationships across the image by allowing it to represent features with more context during classification.

4.3. Discussion and Real-World Implications

The performance gap between the two models could be explained by their core architectural principles while CNNs primarily focus on learning local patterns using fixed-size convolution filters, ViT use self-attention mechanisms that enable them to model global relationships across the entire image. This broader view helps the ViT to perform better, showing that understanding of the wider visual context, such as surrounding tissue condition, foot structure and overall morphology, is essential for DFU classification.

This characteristic is particularly important in medical imaging, where subtle clinical markers such as early inflammation, discoloration, or texture variation may extend beyond the boundary of ulcer. The ViT's global attention mechanism is therefore better aligned with the diagnostic complexity of DFU assessment. In addition, utilizing pretrained transformer-based models supports strong generalization, even when labeled datasets are limited, which is a common constraint in clinical applications.

Looking ahead, combining the ViT based classification system with real-time IoT sensor data such as plantar pressure or change in temperature opens up a path for practical use. A system like this could help shift DFU management from reacting to ulcers after they are visible to prevent them at early stage. The approach using thermal-image based DFU detection suggests potential for multimodal systems. Incorporating such thermal cues in future models could improve sensitivity, especially in early-stage ulcer formation [14]. By combining continuous physiological monitoring with automated image based prediction, a unified platform could enable real-time risk scores and generate early warnings. Fig. 8 shows one possible set up for how this kind of system might work in practice.

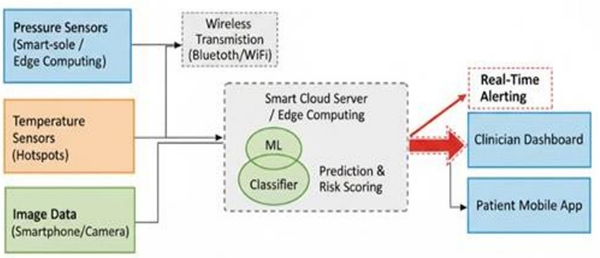

Fig. 8: Proposed Multi-Modal DFU Monitoring System Architecture.

5. Conclusion and Future Work

This study compared the performance of ResNet18 and the DeiT-Small Vision Transformer for diabetic foot ulcer (DFU) classification. The DeiT-Small model demonstrated superior performance, achieving 99.17% accuracy with Zero false positives and only one false negative. In contrast to that, ResNet18 model have achieved 85.40% accuracy and exhibited a higher rate of wrong classification. This supports the idea that optimized architectures can outperform traditional CNNs in medical imaging tasks [15]. These findings indicate that transformer-based architectures, with their ability to learn global spatial relationships through self-attention, are better suited for DFU recognition than approaches based on convolution that rely primarily on localized feature extraction.

Future work will focus on evaluating both architectures under identical non-pretrained conditions to isolate the impact of base architecture versus transfer learning and use a unified split to eliminate dataset-split variability and enable a strict model- to-model comparison. Additionally, a deployment-focused extension is planned through integration with Internet-of-Things (IoT) sensor data, including plantar pressure and thermal monitoring.

This results shows, the vision transformers as a promising foundation for next-generation DFU screening systems, with a great potential to improve early diagnosis, reduce preventable complications, and support scalable AI-assisted diabetic care system to enable accessible, real-time DFU assessment for both clinicians and patients for maintaing the medical history and to keep people self-aware of the complications that can be raised due to poor maintenance.

References

Journal articles

1. Bouatmane, A. Daaif, A. Bousselham, B. Bouihi, O. Bouattane, A Multimodal Deep Learning Model Integrating CNN and Transformer for Predicting Chemotherapy-Induced Cardiotoxicity. IEEE Access. **13**, 57568–57588 (2025).
2. A.R.W. Sait, R. Nagaraj, Diabetic Foot Ulcers Detection Model Using a Hybrid Convolutional Neural Networks–Vision Transformers. Diagnostics. **15**(6), 736 (2025).
3. B. Cassidy, N.D. Reeves, J.M. Pappachan, D. Gillespie, C. O'Shea, S. Rajbhandari, A.G. Maiya, E. Frank, A.J. Boulton, D.G. Armstrong, B. Najafi, J. Wu, R.S. Kochhar, M.H. Yap. The DFUC 2020 Dataset: Analysis Towards Diabetic Foot Ulcer Detection. touchREV Endocrinol. **17**(1), 5-11 (2021).
4. I. Kocak, S. E. Bayramog˘lu, N. Sayın, L. Thalib, Detection of Plus Disease in Retinopathy of Prematurity Using Deep Learning Models: Evaluating Vision Transformers and ResNet Architectures. Int. J. Imaging Syst. Technol. **35**, (2025). doi.org/10.1002/ima.70174.
5. J. Zhang, Y. Qiu, L. Peng, Q. Zhou, Z. Wang, M. Qi, A Comprehensive Review of Methods Based on Deep Learning for Diabetes-Related Foot

Ulcers.Frontiers in Endocrinology (Lausanne). **13**, 945020 (2022). doi: 10.3389/fendo.2022.945020.

6. M. Ahsan, S. Naz, R. Ahmad, H. Ehsan, A. Sikandar. A Deep Learning Approach for Diabetic Foot Ulcer Classification and Recognition. Information, **14(**1), 36 (2023).
7. M. A. Fadhel, L. Alzubaidi, Y. Gu et al. Real-time Diabetic Foot Ulcer Classification Based on Deep Learning and Parallel Hardware Computational Tools. Multimedia Tools and Applications, **83**, 70369–70394 (2024).
8. M. S. A. Toofanee et al. DFU-SIAM: A Novel Diabetic Foot Ulcer Classification with Deep Learning. IEEE Access, **11**, 98315–98332 (2023).
9. A. Galdran, G. Carneiro, M. A. G. Ballester, Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification. in Diabetic Foot Ulcers Grand Challenge (DFUC 2021), Lecture Notes in Computer Science. **13183**, (2022). doi: 10.1007/978- 3-030-94907-5 2.
10. A. Kumar, L. Nelson, S. Singh, ResNet-50 Transfer Learning Model for Diabetic Foot Ulcer Detection Using Thermal Images, 2023 2nd International Conference on Futuristic Technologies (INCOFT). Belagavi, Karnataka, India. 1-5 (2023). doi: 10.1109/INCOFT60753.2023.10425447.
11. E. M. D'Sa, M. Shetty, S. B. Shetty, D. Prabhu, A. Shetty, Diabetic Foot Ulcer Detection Using Deep Learning: A Focus on EfficientNet for Enhanced Accuracy, 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE). Nitte, India. 701-705 (2025). doi: 10.1109/AIDE64228.2025.10987389
12. R. R, L. M, C. J. Karthik, C. Y. Sai, K. Dhanush, M. Mahesh, A Real-Time IoT and Machine Learning-based Smart Insole System for Early Detection and Prevention of Diabetic Foot Ulcers. 2025 7th International Conference on Inventive Material Science and Applications (ICIMA), Namakkal, India, 870-876 (2025). doi: 10.1109/ICIMA64861.2025.11074003.
13. R. Mohan, N. S. Madhava Raja, R. Damasevicius, D. Taniar, S. Prabha, V. Rajinikanth, Vision Transformer Based Diabetic Foot-Ulcer Detection: A Study. in Proc. 2024 Tenth Int. Conf. Bio Signals, Images, and Instrumentation (ICBSII), Chennai, India, 1–4 (2024). doi:10.1109/ICBSII61384.2024.10564047.
14. S. Khalil, S. Y. Arafat, R. Muzaffar, Wound Detection and Size Estimation From DFU Images, 2024 19th International Conference on Emerging Technologies (ICET). Topi, Pakistan. 1-7 (2024). doi: 10.1109/ICET63392.2024.10935182.
15. T. M. M. S. Mohamed, A. Huong, K. Tay, C. C. Chew, Pretrained Deep Learning Networks for Diabetic Foot Ulcer (DFU) Classification. 2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina. 1-6(2025). doi: 10.1109/INFOTEH64129.2025.10959253.