Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MULTI-SCALE CROSS-ATTENTION VISION TRANSFORMER FRAMEWORK WITH CLASS-IMBALANCE OPTIMIZATION FOR AUTOMATED DIABETIC FOOT ULCER DIAGNOSIS

GOWRI MANOHARI V[1], MERCY PAUL SELVAN[2]*

[1]Assistant Professor, Department of Computer Science Engineering, Sathyabama Institute of Science and Technology, Chennai,
[2]Professor, Department of Computer Science Engineering, Sathyabama Institute of Science and Technology, Chennai
gowrimanohari22@gmail.com, mercypaulselvan1@outlook.com

***Corresponding author: Mercy Paul Selvan**

ABSTRACT

Diabetic foot ulcers (DFUs) are a serious complication of diabetes that, if left untreated, can result in the loss of a lower limb. In contrast to traditional clinical approaches for classifying DFUs, automated methods based on deep learning architectures have shown promising results. However, current deep learning techniques frequently fail to capture both global contextual information and fine-grained local characteristics, which reduces generalization in real-world scenarios. This study suggests CrossViT versions with multi-scale feature learning and a weighted cross-entropy loss function to successfully handle class imbalance in order to address these issues. This paper presents a novel approach that improves feature extraction and representation learning on DFU images by utilizing CrossViT across multiple scales. The multi-scale CrossViT architecture enables the model to learn both local lesion features and global image-level representations, which is crucial for detecting ulcer patches of varying sizes, textures, and colors. To further address class imbalance in DFU datasets, a weighted cross-entropy loss (WCEL) is employed during training to emphasize underrepresented classes, such as early-stage or small ulcers. This enhances the model's sensitivity and overall diagnostic accuracy. Experimental results demonstrate that the proposed CViT-WCEL approach achieves an accuracy of 98.19% and outperforms conventional CNN and single-scale transformer models in terms of F1-score, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). These findings highlight the potential of the proposed method as a scalable and efficient solution for reliable and early DFU diagnosis in real-world clinical environments.

Keywords: *Diabetic Foot Ulcer, CrossViT, Multi-scale Vision Transformer, Weighted Cross-Entropy, Deep Learning, Medical Image Analysis, Automated Diagnosis, Explainable AI*

1. INTRODUCTION

DFU stands for serious foot injuries caused by diabetes. Between 2000 and 2021, the number of people with diabetes rose from 151 million to over 537 million, with over 422 million suffering from the condition in 2014. Between 2000 and 2021, the number of persons over 18 who had diabetes rose by 10.5% [1]. According to Table 1, 630 million people worldwide are expected to have diabetes by the end of 2035.

Table 1: Diabetes And Diabetic Foot Ulcer Prevalence In 2021: Facts And Estimations.

At the glint	Year 2000	2021	2030	2045
The number of diabetics in India (in millions)	32.7	74.2	101	124.9
15% of people get diabetic foot ulcers (in million)	22.65	80.55	96.45	117.45
The quantity of individuals with diabetes (in million)	151	537	643	783

Global adult population (19–79 years old) (in billion)	3.2	7.9	8.6	9.5
The prevalence of diabetes (in percentage)	4.6%	10.5%	11.3%	12.2%
Global adult population (19–79 years old) (in billion)	3.2	7.9	8.6	9.5

Furthermore, eighty percent of these patients reside in underdeveloped nations, which have fewer medical facilities and lower awareness of patient care issues [2]. Diabetic foot ulcers (DFUs), which affect 15% to 25% of diabetic individuals, can result in lower limb amputations, hospital stays, and even death if they are not treated. Limb amputation is a possible outcome of DFU infection, and patients who suffer amputations have a much worse survival rate. This condition adversely affects social interactions, quality of life, and livelihood [3]. Another severe complication is gangrene, which is a result of tissue damage due to illness. DFUs should be anticipated to be even more common in the future. Over one million high-risk diabetic patients who may develop DFUs lose part of their foot annually because of unavailability of resources and medical support. Reportedly, there is one diabetic foot surgery every twenty seconds. In Figure 1, Illustrations (a) to (h) show an example of a normal, healthy foot, and (i) to (p) show diabetic foot ulcers [4].

DFUs are extremely diverse concerning their size, shape, texture and color and conventional models are not especially good at capturing the local and global features. In many rural or resource-poor areas, there is usually no access to sophisticated diagnostic methods and qualified physicians, which can lead to delay in diagnosis or misdiagnosis. Furthermore, medical image data tends to be biased in terms of common cases of ulcer, fewer cases of severe or uncommon cases, and traditional models are biased to majority classes and ignore important features. Current deep learning methods also lack generalization due to differences in lighting, skin colour, and image acquisition devices. Thus, a powerful, scalable, and smart system that can overcome challenges related to multi-scale feature extraction, class imbalance, and variability is needed to help make early and accurate DFU diagnosis.

Figure 1: Images of diabetic feet (a)–(h) a healthy, normal foot. (i)– (p) The foot is affected by diabetic foot ulcers

The proposed CrossViT variants with different scales and weight values of cross-entropy loss constitute a promising transformer model for DFU classification. The model overcomes the limitations of single-scale vision models by effectively capturing detailed lesion-specific textures and the surrounding anatomical information through multi-scale feature learning employing parallel CrossViT branches. Additionally, class imbalance in DFU datasets is addressed by class-balanced weighted cross-entropy loss, which increases the model's sensitivity and robustness for various DFU kinds. Our findings show that the multi-scale variants of CrossViT demonstrate better classification accuracy, suggesting the model's robustness to varying DFU cases. The research offers a clinically relevant contribution by facilitating fast, accurate and efficient DFU diagnosis with deep vision transformers.

1.1 The main contribution of this research

- To develop an innovative approach that utilizes CrossViT variations across multiple scales to improve feature extraction and representation learning from DFU images.
- The multi-scale CrossViT structure allows the model to simultaneously learn global

contextual and local lesion features that are important in the detection of ulcer regions, which are of varying size, texture, and color.

- We use a Weighted Cross-Entropy Loss (WCEL) which emphasizes those classes that are underrepresented, e.g. early-stage or microscopic ulcers, to overcome the problem of imbalance of classes in DFU datasets.

The remainder of the document is structured as follows: The literature review is in Section 2, the suggested model is in Section 3, the data analysis is in Section 4, and the conclusion is in Section 5.

2. LITERATURE SURVEY

The authors [5] used two convolutional neural networks (CNNs) that were modified using a multilevel refinement technique to classify DFU images into four groups: ischemia, infection, none, and both ischemia/infection. Batch normalization, transfer learning, and a fully linked layer with varying numbers of neurons form the foundation of the selected multilevel design. A dataset of 8,242 photos was subjected to data augmentation techniques in order to increase performance, address class imbalance, and lessen overfitting. The proposed method outperformed earlier methods in terms of the number of image classes, achieving accuracy, Kappa, and F1-score values of 95.91%, 93.28%, and 95.10%, respectively, using five-fold cross-validation.

Nagararajan et al. [6] proposed a hybrid of Sparrow Search Optimization (SSO) and deep learning (SSODL-DFUDC) to recognize and classify diabetic foot ulcers. The SSODL-DFUDC method can be used to classify DFUs. It is based on Inception-ResNet-v2 to produce feature vectors in this manner. The optimal hyperparameters of the Inception-ResNet-v2 model are determined with the help of the SSO approach, which enhances DFU classification. This approach is better than the conventional trial and error method of hyperparameter tuning, which is time consuming and not accurate. In addition, stacked sparse autoencoder (SSAE) model is used to classify DFUs. The outcomes of the expansive experiments have shown that the SSODL-DFUDC system has a high performance in comparison with the current deep learning approaches.

Anastasios Doulamis et al. [7] suggested a non-invasive photonic approach to the treatment of diabetic foot ulcers (DFUs). The instrument assessed the state of an ulcer by means of thermal and hyperspectral imaging. Oxyhemoglobin and deoxyhemoglobin biomarkers were estimated with this photonic imaging method. This system was combined with signal processing algorithms on the basis of deep learning which reduced noise and improved the accuracy of pixels with the help of super-resolution methods.

esides the ensemble method suggested by the winning team of the DFUC2020, Yap et al. [8] consider numerous deep learning algorithms, including three types of Faster R-CNN, YOLOv3, YOLOv5, EfficientDet, and a new Cascade Attention Network. Each method is described in detail in terms of the model architecture, training parameters, and additional procedures such as preprocessing, data augmentation, and post-processing. Each of the methods is discussed in the research. Both methods have a post processing stage to reduce false positive and a data augmentation stage to increase the number of training photos. The Deformable Convolution-based Faster R-CNN variation achieved a higher F1-score of 0.7434 and a mean average precision (mAP) of 0.6940, compared to the others. The authors conclude that although multi-deep learning ensemble methods could be used to boost the F1-score, the mAP might not be affected.

In a more recent article, Goyal [9] proposed computer vision (CV) algorithms to automatically identify the diabetic foot ulcers (DFUs) of various grades and severity. The authors primarily used machine learning methods to differentiate between DFU areas and healthy areas of the foot area to identify features that can lead to a false alarm of DFUs. Moreover, the authors used Fully Convolutional Networks (FCNs) to segment the DFU regions and normal skin from the full-foot image. Finally, they adopted portable and robust deep learning methods to localise DFUs in the foot for tele-diabetic monitoring.

Current diabetic foot ulcer (DFU) diagnosis systems have some limitations despite technological innovations. Photonic methods using thermal and hyperspectral data offer rich biomarker information but require costly and sophisticated equipment, limiting their use. Deep learning techniques like Faster R-CNN, YOLO variants and EfficientDet perform well but need large datasets and complex pre-processing. They can require extensive data augmentation and ensemble methods for accuracy. They also struggle with changes in ulcer presentation, illumination and patient characteristics. Their computational demands also

restrict their real-time and low-resource use. Thus, current systems are not scalable, accessible and efficient for clinical use.

2.1 Problem Identification for Existing System

- Manual visual inspection, which is subjective, laborious, and prone to diagnostic inaccuracy, is a major component of current DFU diagnosis techniques.
- Multi-scale ulcer patterns are difficult for traditional machine learning and CNN-based models to capture, which reduces accuracy in complex or nuanced DFU situations.
- Most existing systems fail to handle class imbalance effectively, resulting in poor detection of minority ulcer categories, such as severe or atypical lesions.
- Current automated DFU systems lack robustness across diverse patient groups, imaging conditions, and ulcer severity levels.

2.2 Research Gap

Existing research on DFU diagnosis still has some limitations. Most of the current methods employ conventional CNN models that are unable to learn multi-scale features of the lesion, such as fine edges, surrounding skin, and ulcer sizes. Furthermore, current studies do not effectively balance the number of samples from normal skin, mild ulcer and severe ulcer classes, causing biased modelling and inferior performance of these clinically relevant classes. The majority of research is trained and tested on small datasets, limiting their generalisation to various practical use cases. Moreover, there is a lack of transformer and cross-attention models tailored for DFU images, and very few studies present reliable and explainable models for clinicians' confidence. These challenges suggest the need for sophisticated and multi-scale deep learning approaches that account for class imbalance to ensure reliable DFU diagnosis.

3. PROPOSED METHODOLOGY

Here, we present a new approach that employs different sizes of CrossViTs to improve feature extraction and representation learning of DFU images. The CrossViT architecture with multi-scale features allows the model to learn both contextual cues and lesion-specific features, which are crucial for detecting ulcer areas of varying sizes, textures and colours. A weighted cross-entropy loss function (WCEL) is utilized to prioritize under-represented classes, including tiny or early-stage ulcers, in order to address the class imbalance problem in the DFU data set. Figure 2 depicts the architecture of the suggested approach.

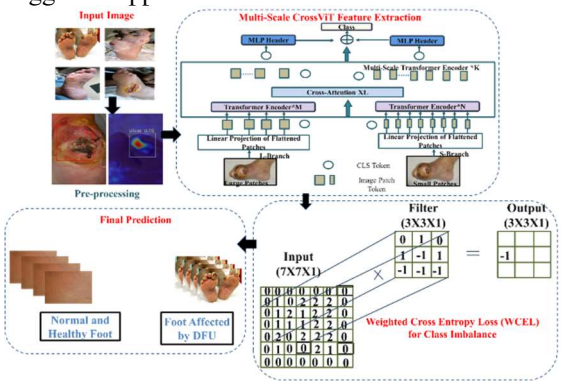


Figure 2: Architetcure diagram of proposed Model

3.1 Dataset Description

The more compact and complex binary classification data set is that of DFU or "DS2" in Fig. 3. The DS2 is from the diabetic unit of Nasiriyah Hospital in southern Iraq. The data set contains 1,055 patches, 543 normal patches and 512 ulcerated patches. For this data set, 754 full-foot (224 x 224) pixel images were taken for each patch using electronic devices, such as tablets and mobile phones. The images were then labelled as either healthy or ulcerated DFU skin.

Figure 3: Sample Images from Dataset

3.3 Data Splitting and Cross-Validation

The dataset has been cross-validated using five-fold strategy to ensure a solid and impartial performance assessment. Under this approach, the entire dataset was divided into five parts. Every iteration was based on four subsets used to train and a subset used to validate. This was repeated 5 times wherein each subset was used as the validation set once.

Using this cross-validation scheme enables to comprehensively test the power and extrapolative ability of the model by testing the model with multiple data partitions, thereby reducing the risk of

overfitting and providing more consistent and credible results.

3.3 Pre-processing

Before feeding DFU images into the CrossViT model, preprocessing is essential to ensure high-quality and uniform inputs. The preprocessing pipeline consists of three main steps. First, image cleaning removes low-quality images that are blurred, overexposed or underexposed, duplicated, or contain artifacts. Given an image set $I=\{I_1,I_2,....,I_N\}$, the cleaned set I_{clean} is:

$$I_{clean}=\{I_i\in I\mid Q(I_i)\geq\tau\} \quad (1)$$

Where $Q(I_i)$ is a quality metric for image I_i, and τ is a predefined threshold. Second, ROI extraction crops images to focus on the foot and ulcer regions, removing irrelevant background. For an image I_i with coordinates (x,y) and ROI dimensions of width w and height h, the cropped image I_i^{ROI} is:

$$I_i^{ROI}=I_i[x:x+w,y:y+h] \quad (2)$$

Finally, Resizing & Normalization adjusts all images to a fixed resolution $H\times W$ (e.g., 224×224) and scales pixel intensities. Min-max normalization maps pixel values to [0, 1]:

$$I_i^{norm}(x,y,c)=\frac{I_i^{resized}(x,y,c)-I_{\min}}{I_{\max}-I_{\min}} \quad (3)$$

Alternatively, standardization using channel-wise mean μ_c and standard deviation σ_c can be applied:

$$I_i^{norm}(x,y,c)=\frac{I_i^{resized}(x,y,c)-\mu_c}{\sigma_c} \quad (4)$$

These steps standardise and de-noise the inputs to allow the CrossViT model to learn both global and local representations of lesions.

3.4 Feature Extraction using Multi-Scale CrossViT

In the proposed approach, input images $I\in ¡^{H\times W\times C}$ are processed through multi-scale CrossViT blocks to extract both global and local features, enabling effective DFU detection [10]. Figure 2 shows the diagram for Multi-Scale CrossViT.

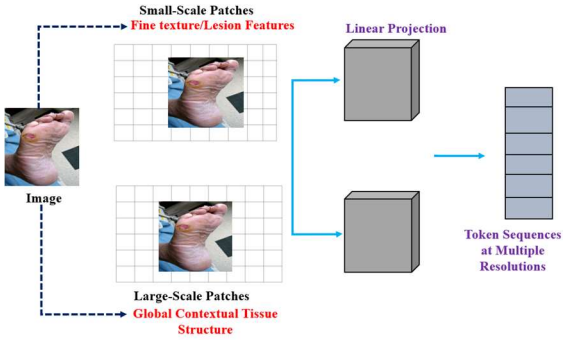


Figure 4: Architecture diagram for Multi-Scale CrossViT

For global context learning, the image is divided into large patches of size $P_g\times P_g$ to capture the overall spatial structure of the foot and the ulcer.

$$X_g=Flatten\left(Patch\left(I,P_g\right)\right)+E_g \quad (5)$$

Where X_g is the global patch embedding and E_g is the positional encoding. For local lesion representation, smaller patches $P_l\times P_l$ are used to detect fine-grained ulcer characteristics, such as subtle textures and color variations.

$$X_l=Flatten\left(Patch\left(I,P_l\right)\right)+E_l \quad (6)$$

With X_l representing the local patch embeddings and E_l representing their positional encoding, the multi-scale attention mechanism integrates global and local features through cross-attention.

$$CrossAttention(Q,K,V)=Soft\max\left(\frac{QK^{\Gamma}}{\sqrt{d}}\right)V \quad (7)$$

$$X_{ms}=CrossAttention\left(X_lW_q,X_gW_k,X_gW_v\right)+X_l \quad (8)$$

Where X_{ms} represents the multi-scale fused features, W_q,W_k,W_v denotes the learnable parameters used to estimate queries, keys, and values, and d is the dimensionality of the query/key vectors. The combination of these global and local characteristics allows the model to be sensitive and specific in detecting coarse and fine details of

lesions at the same time and enhances its general classification ability.

3.5 Classification with Weighted Cross-Entropy Loss (WCEL)

The proposed model employs WCEL as a classification tool to handle the issue of class imbalance that is frequently present in the DFU datasets, with early-stage or small ulcers frequently being underrepresented [11].The loss function penalizes misclassified samples from minority classes, thereby enabling the model to learn more discriminative features and improving its sensitivity to infrequent lesion types. Given a training sample with true class label $y \in \{1,2,...,K\}$, predicted probabilities p_k, and class-specific weights w_k, the weighted cross-entropy loss is defined as:

$$L_{WCEL} = -\sum_{k=1}^{K} w_k \cdot y_k \cdot \log(p_k) \quad (9)$$

Where $y_k = 1$ if the sample belongs to class k, otherwise 0. The class weight w_k is typically computed as the inverse of class frequency to ensure higher importance for minority classes:

$$w_k = \frac{N}{K \cdot n_k} \quad (10)$$

Where n_k is the number of samples in class k, K is the total number of classes, and N is the size of the entire dataset. The adoption of these weights enables the Weighted Cross-Entropy Loss (WCEL) to emphasize underrepresented ulcer categories, resulting in improved detection accuracy, higher sensitivity, and more balanced performance across the full range of diabetic foot ulcer (DFU) severities.

3.6 Advantages of Proposed Method

- Preserves global foot features and local ulcer features with multi-scale CrossViT features.
- Increases the ability to detect small, subtle, and early DFUs with improved local lesion details.
- Increases model robustness and transferability on various DFU datasets, including complex real-world clinical images.
- Minimises confusion among minority classes to ensure accurate diagnosis for early detection and prompt treatment.

4. RESULT AND DISCUSSION

The outcomes of the suggested CViT-WCEL model on the classification of diabetic foot ulcers are shown in this section. Metrics such as accuracy, F1-score, sensitivity, specificity, and AUC-ROC are evaluated for the suggested model using five-fold cross-validation. The accuracy, stability, and generalization performance of the model are demonstrated by these results. To demonstrate the contribution of weighted loss, cross-attention, and multi-scale feature extraction to improving classification accuracy, more comparisons and ablation studies are also provided.

4.1 Training Protocol

A set of hyperparameters was used to train the CrossViT-WCEL model in order to achieve optimal performance and steady convergence. We employed an Adam optimizer with default momentum terms ($\beta_1 = 0.9$, $\beta_2 = 0.999$), a batch size of 32, and an initial learning rate of 0.0001. To prevent overfitting and improve generalization, a learning rate scheduler was employed to dynamically modify the learning rate in response to the model's performance on the validation set. Once convergence was achieved, training was stopped early after 100 training epochs. Augmentation techniques employed to resist picture distortions of diabetic foot ulcer (DFU) pictures include rotation, flipping and scaling.

All the experiments were carried out in Python using deep learning libraries such as PyTorch and TensorFlow, and additional libraries such as NumPy, OpenCV and Scikit-learn for data pre-processing and evaluation. To speed up the model's training and inference, the application was run on a PC equipped with an NVIDIA GP, 16 GB of RAM, and an Intel Core i7 processor. We also cross-checked the experiments on cloud computing platforms such as Google Colab for reproduction and parallel processing. This setup was capable of handling the multi-scale transformer model and a rigorous evaluation in multiple folds.

4.2 Evaluation Metrics

To assess the model's performance, we calculated the accuracy, F1-score, specificity, sensitivity, and

area under the receiver operating characteristic (AUC-ROC) curve. For each group, the results were calculated over five folds.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

When there is class imbalance, the F1 score the harmonic mean of precision and recall should be used to evaluate a model's performance. It also ensures accurate recording of false positives and false negatives.

True positives and false negatives are also considered during evaluation.

$$F1-Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

Specificity measures how well a model avoids false positives in order to assess its accuracy in identifying negative cases. Making ensuring that people who are not ill are not mistakenly diagnosed as ill is particularly crucial.

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

The true positive rate also known as recall or sensitivity measures a model's ability to correctly identify real positive cases. It displays the percentage of sick or targeted samples that the system correctly detects.

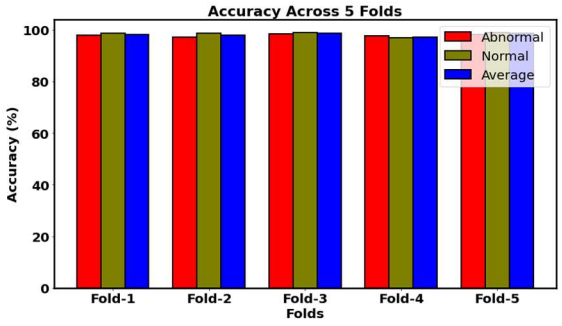
$$Sensitivity = \frac{TP}{TP + FN} \quad (14)$$

AUC-ROC, or Area Under the Receiver Operating Characteristic Curve, gauges how well a model can distinguish between positive and negative classes. A higher AUC indicates superior discriminative ability across various categorization thresholds.

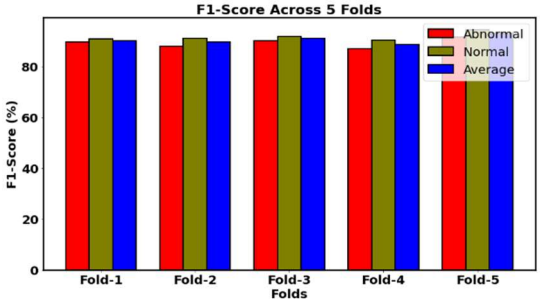
$$AUC = \int_0^1 TPR\left(FPR\right) d\left(FPR\right) \quad (15)$$

Table 2: Classifier results from the suggested approach with distinct metrics and folds.

Class	Accuracy	F1-Score	Specificity	Sensitivity	AUC-ROC
		Fold-1			
Abnormal	97.89	89.56	87.22	91.23	96.78
Normal	98.78	90.88	89.77	90.79	96.78
Average	98.33	90.22	88.49	91.01	96.78
		Fold-2			
Abnormal	97.19	87.98	91.56	94.19	91.94
Normal	98.66	91.11	89.33	86.65	91.94
Average	97.92	89.54	90.44	90.42	91.94
		Fold-3			
Abnormal	98.55	90.22	94.56	95.81	89.91
Normal	98.98	91.76	95.56	95.91	89.91
Average	98.76	90.99	95.06	95.86	89.91
		Fold-4			
Abnormal	97.65	86.87	88.39	93.23	90.61
Normal	96.98	90.43	92.23	90.65	90.61
Average	97.31	88.65	90.31	91.94	90.61
		Fold-5			
Abnormal	98.33	91.54	94.17	86.44	96.63
Normal	98.99	94.45	92.51	94.51	96.63
Average	98.66	92.99	93.34	90.47	96.63



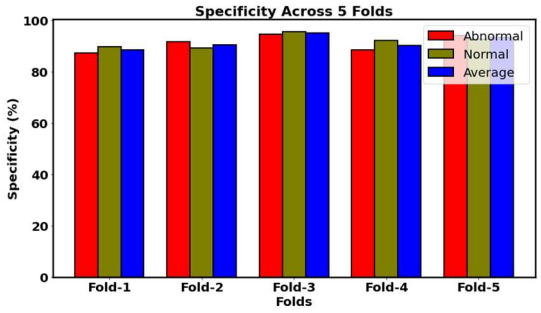
4(a) Accuracy Analysis across 5 Folds



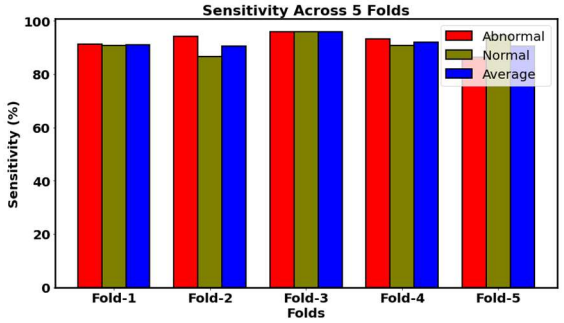
4(b) F1-Score Analysis across 5 Folds

Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10

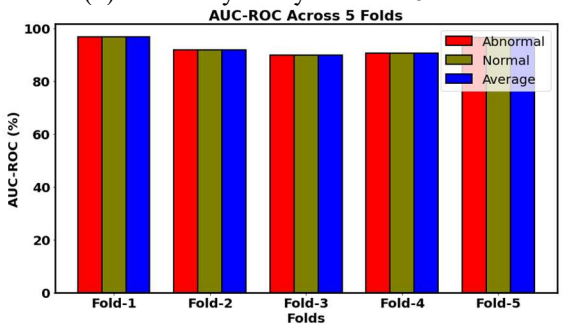
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



4(c) Specificity Analysis across 5 Folds



4(d) Sensitivity Analysis across 5 Folds



4(e) AUC-ROC Analysis across 5 Folds

Figure 5: Classifier results from the suggested approach with distinct metrics and folds

The final results of the classifier employing the suggested method are displayed in Table 2 and Figure 4 in terms of five performance metrics: sensitivity, accuracy, F1-score, specificity, and AUC-ROC. These results show the model performs extremely well in both the Normal and Abnormal classes, with an average accuracy across the folds from 97.31% to 98.76%. The F1-score and specificity values demonstrate a balanced classification with few false positives and false negatives and an effective separation between the two classes. The model has successfully detected the true positives in every fold, according to the sensitivity. Finally, the AUC-ROC values, which are greater than 89% for all folds, indicate the excellent discriminative power of the classifier. These results demonstrate the superior generalisation ability of the adopted method and its stability and robustness under multi-fold analysis.

4.3 Training Validation Accuracy and Loss

The preprocessed DFU images were used to train the CrossViT-WCEL model. It was trained to discriminate between normal and ulcerated skin. The weighted loss function helped the training process by having a greater penalty for misclassifying minority or harder-to-classify examples.

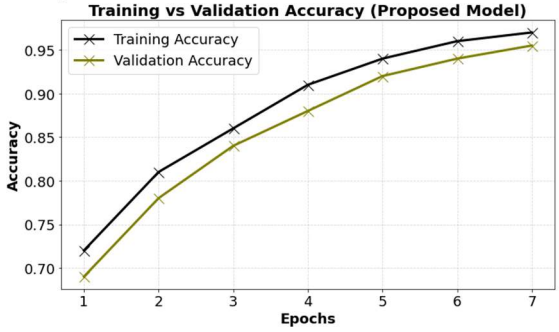


Figure 6: Training and Validation accuracy analysis

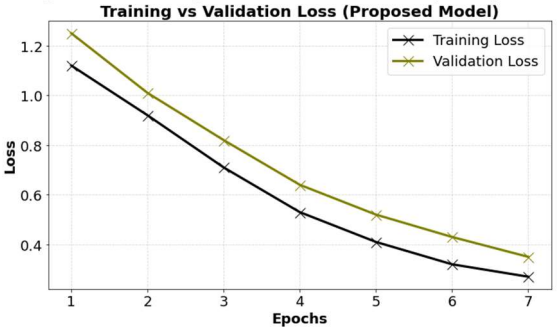


Figure 7: Training and Validation Loss Analysis

4.4 Confusion matrix

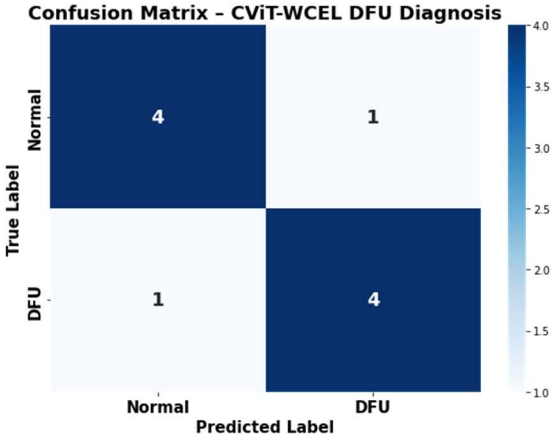


Figure 8: Confusion Matrix for the suggested method

The confusion matrix in Figure 7 displays the classification performance of the CViT-WCEL model for DFU diagnosis.

The model has high discriminative power between the two classes, with four normal cases and four DFU cases correctly predicted among the ten test samples. There is limited overlap and confusion between the classes, with one normal instance misclassified as DFU and one DFU instance misclassified as normal. The equal number of correct predictions for both classes suggests that the model has successfully learned multi-scale representations of the lesions. Furthermore, the low number of misclassifications indicates that the combination of CrossViT and weighted cross-entropy loss for class re-weighting and lesion variability were successful.

4.5 Discussion

The new CViT-WCEL model has been shown to be superior to existing state-of-the-art DFU detection models; however, it is clear that, while there are commonalities, there are also key differences. The proposed model shows good accuracy - comparable to previous models such as DFU_QUTNet and DFU_FNet - suggesting deep learning models can be used to detect DFU. However, the proposed CrossViT model extracts global and local features of the lesion through multi-scale attention rather than just local features, as in CNN models. This results in improved sensitivity and specificity, particularly in detecting early or small DFUs, which are prone to be missed by current models.

A significant difference is the improved accuracy (98.19%) of the proposed model over the 92-95% reported in the previous studies. This is due to two main reasons: (i) the proposed model incorporates multi-scale feature learning, which results in a better representation of the various ulcer patterns, and (ii) it employs weighted cross-entropy loss, which addresses the class imbalance problem (which is not sufficiently tackled in previous studies). By contrast, many models heavily depend on data augmentation or ensemble learning to address imbalance, adding to the computational burden while not addressing class sensitivity.

However, it should be noted that some of the previous studies, particularly large-scale datasets or ensemble object detection methods, show very good localization performance which is not explicitly tested in this study. Also, the dataset size, data acquisition and evaluation metrics might affect the results and make it difficult to compare. Therefore, it appears that although this approach outperforms other approaches in terms of classification accuracy, it should be evaluated on a larger and more diverse dataset.

4.6 Ablation Study

This subsection describes the ablation study to assess the contribution of various parts of the CViT-WCEL architecture. The contributions of feature fusion, cross-attention, and the weighted cross-entropy loss are evaluated by individually removing or altering components. The findings show the improvement in classification accuracy provided by each module, and confirm the benefits of the combined architecture.

Table 3: Ablation Research for Accuracy Analysis

Model Variant	Multi-Scale Fusion	Cross-Attention	WCEL	Accuracy (%)
Baseline CNN	✗	✗	✗	88.78
Vanilla ViT	✗	✗	✗	95.19
CrossViT (Single-Scale)	✗	✓	✗	95.99
CrossViT + Multi-Scale	✓	✓	✗	97.54
Proposed	✓	✓	✓	98.19

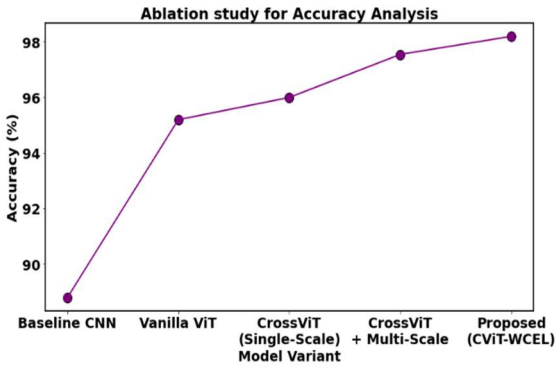


Figure 9: Ablation study for Accuracy Analysis

The ablation study that evaluates the impact of multi-scale fusion, cross-attention, and WCEL on the model's accuracy is also displayed in Table 3 and Figure 8. The plain CNN has the lowest accuracy (88.78%), suggesting its inability to model the features of diabetic foot ulcers (DFU). The plain Vision Transformer (ViT) achieves higher accuracy (95.19%) because it can model the features more effectively. The addition of cross-attention in the single-scale CrossViT model

increases the accuracy to 95.99%. Adding multi-scale fusion boosts the accuracy to 97.54% and demonstrates the role of multi-resolution information of DFU. The full model, incorporating all the components and WCEL, achieves the best accuracy of 98.19%, showing the synergy of each module and the superiority of the proposed CViT-WCEL approach.

Table 4: Comparison of DFU Detection Techniques.

Author	Method	Accuracy (%)
Alzubaidi et al. [12]	DFU_QUTNet	94.50%
Santos et al. [13]	DFU-VGG	93.45%
Rubavathi et al. [14]	FastCNN	92.90%
Fadhel et al. [15]	DFU_FNet	94.70%
Our Method	CViT-WCEL	98.19%

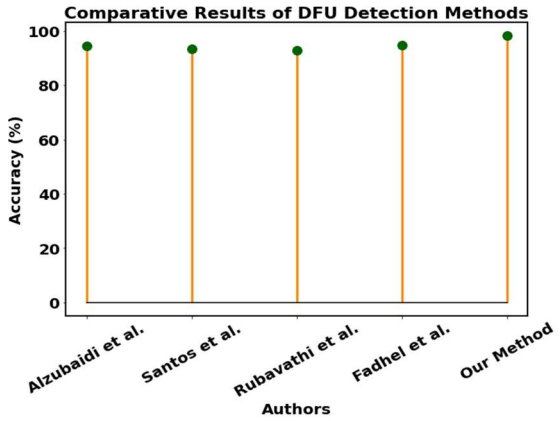


Figure 10: Comparative DFU Detection Technique Results

Table 4 and Figure 9 illustrate the effectiveness of current cutting-edge DFU detection techniques on the suggested CViT-WCEL model. In terms of DFU detection, baseline techniques including DFU_QUTNet (94.50%) by Alzubaidi et al., DFU-VGG (93.45%) by Santos et al., FastCNN (92.90%) by Rubavathi et al., and DFU_FNet (94.70%) by Fadhel et al. demonstrate similar outcomes. But the proposed CViT-WCEL model surpasses all existing methods, with an accuracy of 98.19%, demonstrating a significant advancement in feature extraction and overcoming class imbalance. This performance suggests that the proposed approach is reliable and efficient to detect DFU.

5. LIMITATIONS AND FUTURE WORK

The model has been tested on a small number of DFU datasets, which might not represent the diversity of ulcer shapes, lighting and skin types.

Because the multi-scale CrossViT versions are computationally costly, their deployment on medical devices with limited resources is challenging.

Clinical data, which could improve diagnostic accuracy, is not used in this method, which only uses image diagnosis.

Explainability is only visual (attention maps) and lacks further interpretability tools for clinical adoption.

6. CONCLUSION

The findings demonstrate the effectiveness and accuracy of the CrossViT variations, weighted cross-entropy loss, and multi-scale feature fusion techniques for the detection of diabetic foot ulcers. The suggested CViT-WCEL framework efficiently addresses the heterogeneity of DFU appearances and the class imbalance commonly observed in ulcer datasets by incorporating global contextual understanding and fine-grained lesion representations. The tests demonstrate that the model's learning power is significantly enhanced by the employment of multi-scale fusion and cross-attention mechanisms, and that the addition of weighted cross-entropy loss increases its sensitivity to early stage and minority classes. Overall, the suggested model outperforms CNNs and single-scale transformer models on important performance indicators, suggesting it could be useful for diagnostics. Future work will investigate the incorporation of multi-modal data such as thermal images and medical data with RGB images to improve the early diagnosis and to give a better understanding. Also, the use of self-supervised and few-shot learning could be employed to address the data scarcity and enhance the model's generalization ability.

DECLARATION

Data Availability: The following information was supplied regarding data availability:
https://www.kaggle.com/datasets/laithjj/diabetic-foot-ulcer-dfu

Funding Statement: This research received no external funding.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Ethical Approval: The declaration is "Not Applicable".

REFERENCES

[1]. D. Abdissa, (2020). "Prevalence of diabetic foot ulcer and associated factors among adult diabetic patients on follow-up clinic at jimma medical center, southwest Ethiopia, 2019," an institutional-based cross-sectional study, J. Diabetes Res. 2020 pp. 1–6.

[2]. A. Pourkazemi, (2020). "Diabetic foot care: knowledge and practice," BMC Endocr. Disord. vol. 20, no. 1, pp. 1–8.

[3]. P. N. Thotad, G. R. Bharamagoudar, & B. S. Anami, (2023), "Diabetic foot ulcer detection using deep learning approaches," Sensors International, vol. 4, pp. 100210.

[4]. https://www.kaggle.com/datasets/laithjj/diabetic-foot-ulcer-dfu

[5]. Dos Santos, E. S., Veras, R. D. M. S., Dos Santos, F. D. C. T., Ito, M., Bianchi, A. G. C., Aires, K. R. T., & Tavares, J. M. R. (2025). "Enhancing Diabetic Foot Ulcer Classification Through Fine-Tuned Multilevel CNN," IEEE Access.

[6]. S. Nagaraju, K. V. Kumar, B. P. Rani, E. L. Lydia, M. K. Ishak, I. Filali, & S. M. Mostafa, (2023). "Automated diabetic foot ulcer detection and classification using deep learning," IEEE Access, vol. 11, pp. 127578-127588.

[7]. A. Doulamis, (2021). A non-invasive photonics-based device for monitoring of diabetic foot ulcers: architectural/sensorial components & technical specifications," Inventions, vol. 6, no. 2, pp. 27-27.

[8]. M. H. Yap, R. Hachiuma, A. Alavi, R. Brüngel, B. Cassidy, M. Goyal, & E. Frank, (2021). "Deep learning in diabetic foot ulcers detection: A comprehensive evaluation," *Computers in biology and medicine,* vol. 135, pp. 104596.

[9]. M. Goyal, "Novel computerised techniques for recognition and analysis of diabetic foot ulcers," Ph.D. thesis, Manchester Metropoli tan Univ., 2019. [Online]. Available: https://e-space.mmu.ac.uk/625105/1/Thesis_Manu_Revised.pdf

[10]. C. F. R. Chen, Q. Fan, & R. Panda, (2021). "Crossvit: Cross-attention multi-scale vision transformer for image classification," In Proceedings of the IEEE/CVF international conference on computer vision, pp. 357-366.

[11]. Y. Ho, & S. Wookey, (2019). "The real-world-weight cross-entropy loss function: Modeling the costs of mislabelling," IEEE access, vol. 8, pp. 4806-4813.

[12]. L. Alzubaidi, M.A. Fadhel, S.R. Oleiwi, O. Al-Shamma, J. Zhang, (2020). "DFU_QUTNet: diabetic foot ulcer classification using novel deep convolutional neural network," Multimed. Tools Appl. vol. 79, pp. 15655–15677.

[13]. F. Santos, E. Santos, L.H. Vogado, M. Ito, A. Bianchi, J.M. Tavares, R. Veras, 2022. "DFU VGG, a Novel and Improved VGG-19 Network for Diabetic Foot Ulcer Classification," 29th International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, 2022, pp. 1–4.

[14]. C.Y. Rubavathi, J. Diofrin, (2023). "Diabetes Foot Ulcer Diagnosis using Fast Convolution Neural Network," 2023 International Conference on Networking and Communications (ICNWC), IEEE, pp. 1–5.

[15]. M.A. Fadhel, L. Alzubaidi, Y. Gu, J. Santamaría, Y. Duan, (2024). "Real-time diabetic foot ulcer classification based on deep learning & parallel hardware computational tools," Multimed. Tools Appl, pp. 1–26.