Reliable Multimodal Skin Lesion Analysis using Dermoscopic Images and Clinical Metadata

1st Sonia Bouzidi
Research Groups in Intelligent Machines: REGIM-Lab, ENIS,
Sfax, Tunisia
Sonia.bouzidi.doc@enetcom.usf.tn

2nd Imen Jdey
Research Groups in Intelligent Machines: REGIM-Lab, ENIS,
Sfax, Tunisia
Imen.jdey@ieee.org

3rd Fadoua Drira
Research Groups in Intelligent Machines: REGIM-Lab, ENIS,
Sfax, Tunisia
Fadoua.drira@enis.tn

***Abstract*—Reliable decision-making in medical applications has become increasingly crucial, particularly in skin lesion analysis, where misdiagnosis can lead to severe health consequences. While recent advances in computer vision have achieved high performance, relying solely on image-based models often overlooks essential clinical context. To address this limitation, we propose a novel multimodal framework that integrates dermoscopic image analysis with structured clinical metadata to improve both classification performance and decision reliability. The proposed approach combines an optimized Vision Transformer with a Deep Fuzzy Neural Network (ViT-DFNN) for image-based representation learning and a TabNet model for tabular data, enabling effective fusion of heterogeneous information through a decision-level weighted strategy. Experimental results on the HAM10000 dataset demonstrate the effectiveness of the proposed framework, achieving an accuracy of 93.15%, AUC of 96.1%, precision of 92.95%, sensitivity of 93.30%, specificity of 98.85%, and G-mean of 96%. These results confirm that integrating clinical metadata significantly enhances predictive performance and robustness compared to unimodal approaches. To ensure transparency, the framework incorporates Explainable Artificial Intelligence (XAI) techniques, where Score-CAM highlights important image regions and Global Feature Importance (GFI) identifies influential clinical variables. Overall, the proposed method provides a reliable and interpretable solution for AI-assisted skin lesion diagnosis.**

***Index Terms*—Skin lesion analysis, ViT-DFNN, TabNet, multimodal framework, XAI techniques.**

I. Introduction

In recent years, the integration of multimodal data fusion techniques [7], [17] has emerged as a transformative approach in healthcare, enabling improved diagnostic accuracy and personalized treatment strategies [11]. In dermatology, accurate skin lesion analysis [1] remains particularly challenging due to the high variability in lesion appearance, patient demographics, and clinical context. This variability often leads to uncertainty in diagnosis, where relying solely on dermoscopic images [18] may result in incomplete or unreliable predictions.

Traditional AI-based approaches [20] for skin lesion classification have predominantly focused on image data. Although ViTs [2], [14], [15] have demonstrated strong performance by capturing both local and global contextual information [24] through patch-based representations, they inherently ignore valuable clinical metadata such as age, sex, and lesion location. This limitation restricts their ability to model the full clinical context, which is crucial for accurate and reliable diagnosis.

To overcome this limitation, multimodal learning [23] has emerged as a promising direction by integrating heterogeneous data sources. In particular, combining ViT-based image analysis with tabular models such as TabNet allows the model to jointly exploit visual patterns and structured clinical features. TabNet, with its sequential attention mechanism, effectively identifies the most relevant clinical attributes, thereby complementing image-based predictions and reducing uncertainty. Beyond performance, explainability is a critical requirement for medical AI systems, as clinical adoption depends on the ability to understand and trust model decisions [1]. XAI techniques play a key role in this context. For image-based models, methods such as Score-CAM provide visual explanations by highlighting important regions influencing predictions. For tabular data, GFI enables the identification of the most influential clinical variables, enhancing transparency and supporting clinical validation.

In this work, we propose a novel multimodal framework that explicitly addresses these challenges by integrating dermoscopic images and clinical metadata. The proposed approach combines a ViT-DFNN architecture, which enhances feature representation and models uncertainty in image data, with a TabNet model for structured clinical data analysis. A decision-level fusion strategy is employed to aggregate predictions from both modalities, improving robustness and reliability. Furthermore, explainability is ensured through the integration of Score-CAM for image explanations and GFI for tabular data analysis.

The proposed framework is evaluated on the HAM10000 dataset, demonstrating its effectiveness in improving both classification performance and decision reliability.

The remainder of this paper is organized as follows. Section 2 reviews related work in skin lesion analysis with multimodal learning. Section 3 details the proposed framework and its main components. Section 4 presents and discusses the experimental results and their implications. Finally, Section 5 concludes the paper and outlines future research directions.

The Ministry of Higher Education and Scientific Research of Tunisia (LR11ES48).

II. Literature Review

In this section, we present the recently proposed multimodal framework, specifically applied to skin lesion analysis, to illustrate the efficacy of these models for skin lesion prediction, are summarized in table VI.

All the following studies evaluated their proposed multimodal deep learning methods with skin lesion dataset by combining dermoscopic images with patient metadata. Yasmin et al. [13] introduced a multimodal architecture integrating a ViT for visual feature extraction and a multilayer perceptron (MLP) for patient metadata. A fuzziness-based semi-supervised deep learning (FSSDL) approach was used to leverage low-quality or unlabeled samples during training. Evaluation on PAD-UFES-20 (images + metadata) showed an F1-score of 89% and a G-Mean of 90%. Ahmad et al. [7] proposed a deep learning-based multi-sensor fusion framework for melanoma detection that integrates dermoscopic images processed by a multi-layer CNN with clinical metadata analyzed via a fully connected network (FCN). The model achieved an accuracy of 94.5% and an F1-score of 94%, demonstrating improved performance over single-modality approaches. Pavan et al. [12] proposed a hybrid framework combining a ViT and MLP to exploit both visual and tabular information. On the ISIC 2019 dataset, the model achieved an overall accuracy of 87%, with F1-scores of 92% for benign lesions and 56% for malignant lesions. Adebiyi et al. [6] proposed a multimodal deep learning framework based on the ALBEF architecture, combining dermoscopic images encoded by ViT with clinical metadata encoded by BERT. The multimodal approach achieved an accuracy of 94.11% and an AUCROC of 94.26%. Wang et al. [8] proposed the IM-CNN model, which integrates dermoscopic images processed through EfficientNet-B5 with patient metadata analyzed using a fully connected network, achieving an accuracy of 95.1%, a sensitivity of 83.5%, a specificity of 93.2%, an AUC of 97.8%, and an AUC_SEN_80 of 96.0%.

Ahammed et al. [10] proposed a multimodal deep learning approach that combines CNN-based encoders for dermoscopic images with linear layers processing clinical metadata. Their method was evaluated on the ISIC-DICM-17K dataset, a balanced collection of 17,060 images with associated metadata. Incorporating metadata improved model performance, with the best transfer learning model achieving an AUC of 88.51%, showing better distinction between melanoma and non-melanoma cases.

Ou et al. [9] developed a multimodal fusion model (MMF-Net) combining smartphone-acquired clinical images processed with ResNet-50 and patient metadata processed with a fully connected network, trained and evaluated on the PAD-UPES-20 dataset, achieving an average accuracy of 76.8%, balanced accuracy of 77.5%, and AUC of 94.7%.

Cai et al. [11] proposed a multimodal Transformer. The approach employs a ViT as the backbone to extract deep image features, while a Soft Label Encoder (SLE) is designed to embed the metadata. The model was evaluated on a private skin disease dataset and the ISIC 2018 dataset, achieving an accuracy of 81.6% on the private dataset and 93.81% accuracy with an AUC of 99% on ISIC 2018.

III. Methodology

This study proposes a multimodal framework for skin lesion analysis that integrates dermoscopic images with structured clinical metadata, as illustrated in Figure 1.

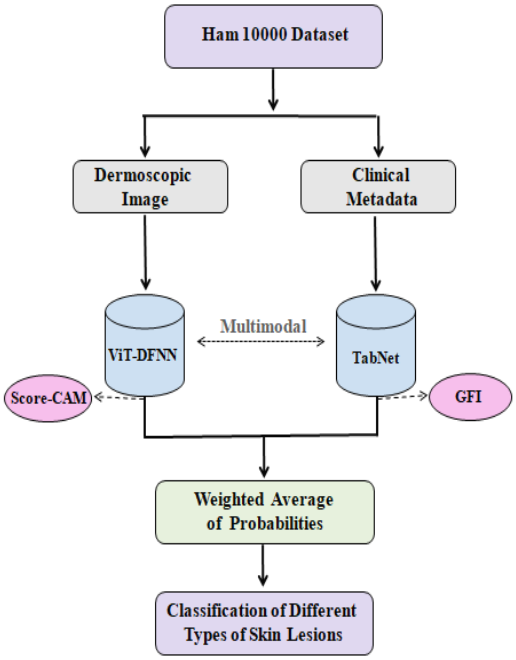

Fig. 1. Proposed multimodal framework using ViT-DFNN and TabNet.

A. ViT-DFNN Based Classification for HAM10000 Image Data

This study introduces **ViT-DFNN**, a hybrid deep learning architecture combining a ViT with a DFNN [5]. DFNN layers are integrated within the ViT encoder blocks to refine feature representations, reduce redundancy, and explicitly model uncertainty in features.

The classification pipeline of our ViT-DFNN branch proceeds as follows:

1) Image Patching and Embedding: Given an input image $x \in \mathbb{R}^{H\times W\times C}$, the image is split into N non-overlapping patches of size $P \times P$:

$$N = \frac{H \times W}{P^2}. \tag{1}$$

Each patch is flattened and projected into a D-dimensional embedding via a trainable linear projection E:

$$X_p^i E \in \mathbb{R}^D. \tag{2}$$

A learnable class token X_{class} and positional embeddings E_{pos} are added:

$$Z_0 = [X_{class}; X_p^1 E; \ldots; X_p^N E] + E_{pos}. \tag{3}$$

2) Transformer Encoder Blocks with DFNN Integration: The embedded patches Z_0 are processed by L transformer encoder blocks. Each block contains a Multi-Head Self-Attention (MHSA) followed by a feedforward MLP. The DFNN is applied after MHSA to refine features (see algorithm 1):

$$Z'_l = \text{MHSA}(\text{LN}(Z_{l-1})) + Z_{l-1}, \quad l = 1, \ldots, L, \tag{4}$$

$$Z_l = \text{MLP}(\text{LN}(Z'_l)) + Z'_l. \tag{5}$$

In the Multi-Head Self-Attention (MHSA) mechanism, the input sequence $Z_{l-1} \in \mathbb{R}^{N\times d}$ is first projected into queries Q, keys K, and values V using learnable weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_h}$:

$$Q = Z_{l-1}W_Q, \quad K = Z_{l-1}W_K, \quad V = Z_{l-1}W_V, \tag{6}$$

where d_h is the dimensionality of each attention head.

The attention scores are computed as a scaled dot-product between queries and keys:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V. \tag{7}$$

Algorithm 1 DFNN Integration within a Transformer Encoder Block

Input: Z_{l-1} (previous layer output), M (number of fuzzy rules), w_i (rule weights), c_i (centers), σ_i (spread), ϵ (small constant)
Output: Z_l^{refined} (refined feature)
$Z'_l \leftarrow \text{MHSA}(\text{LN}(Z_{l-1})) + Z_{l-1}$
for $i = 1$ to M **do**
  $\mu_i \leftarrow \exp\left(-\frac{\|Z'_l - c_i\|^2}{2\sigma_i^2}\right)$
end for
$Z_l^{\text{defuzz}} = \frac{\sum_{i=1}^{M} w_i \mu_i Z'_l}{\sum_{i=1}^{M} w_i \mu_i + \epsilon}$
$Z_l^{\text{refined}} \leftarrow \text{MLP}(\text{LN}(Z_l^{\text{defuzz}})) + Z_l^{\text{defuzz}}$
Return Z_l^{refined}

3) Classification Head: The class token from the final encoder block is passed to a branch-specific MLP classification head. Softmax produces class probabilities:

$$Z_{final}^{img} = \text{Softmax}(\text{MLP}_{classification}(Z_L^{refined})). \tag{8}$$

4) Score-CAM Integration: To enhance interpretability, **Score-CAM** is applied to highlight the regions of the input image that contribute most to the model's prediction. Let $Z_{L,n}$ denote the n-th patch feature map from the final ViT-DFNN encoder block after DFNN refinement. Each feature map is first upsampled to the input image size:

$$\hat{Z}_{L,n} = \text{Upsample}(Z_{L,n}). \tag{9}$$

A masked version of the input image is then created by element-wise multiplication:

$$I_n = I \odot \hat{Z}_{L,n}, \tag{10}$$

where I is the original input image. The model is applied to each masked image to obtain a class score for the target class c:

$$S_n^c = f_c(I_n), \tag{11}$$

where f_c denotes the model's output for class c. Normalized weights for each patch are computed as:

$$\alpha_n^c = \frac{S_n^c}{\sum_{m=1}^{N} S_m^c}, \tag{12}$$

and the final class-specific heatmap is obtained by aggregating the weighted feature maps:

$$L^c = \sum_{n=1}^{N} \alpha_n^c \, \hat{Z}_{L,n}. \tag{13}$$

Finally, L^c is upsampled to match the original image resolution for visualization, where high values indicate regions that strongly influence the model's decision.

B. TabNet-Based Analysis on HAM10000 Tabular Data

In addition to dermoscopic image classification, our study leverages the structured metadata of the HAM10000 dataset to perform lesion diagnosis from clinical tabular data. To model these data, we adopt **TabNet**, a deep learning architecture designed for tabular inputs. Unlike traditional multilayer perceptrons, TabNet employs a sequential attention mechanism that adaptively selects the most relevant features at each decision step.

1) Input Representation: Let $X \in \mathbb{R}^{N\times d}$ denote the tabular input matrix, where N is the number of samples and d is the number of features. Categorical variables (e.g., sex, lesion localization) are encoded via embedding layers, while numerical features (e.g., age) are normalized. The target variable corresponds to the lesion diagnosis class.

2) TabNet Decision Mechanism: TabNet operates through a sequence of T decision steps. At each step t, an attentive transformer computes a sparse feature mask $M^{(t)}$ that highlights the most informative features:

$$M^{(t)} = \text{Sparsemax}(A^{(t)}), \tag{14}$$

where $A^{(t)}$ represents the learned attention scores. The masked features are processed by a feature transformer to generate a step-specific output $D^{(t)}$. The final decision representation is obtained by aggregating all decision steps:

$$D = \sum_{t=1}^{T} D^{(t)}. \tag{15}$$

This mechanism allows TabNet to learn feature importance adaptively while maintaining interpretability.

3) Prediction Output: The aggregated decision representation D is passed through a fully connected classification layer followed by a softmax activation to produce class probabilities:

$$\mathbf{p}^{tab} = \text{Softmax}(WD + b), \tag{16}$$

where W and b are the learnable parameters of the classification layer. The predicted lesion class is determined by selecting the class with the highest probability.

4) Global Feature Importance: To ensure interpretability, TabNet naturally provides feature selection masks at each decision step. The global feature importance (GFI) for feature j is computed by averaging its attention contribution across all samples and decision steps:

$$\text{GFI}_j = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} M_{i,j}^{(t)}, \tag{17}$$

where N is the number of samples and T the number of decision steps. This global aggregation highlights the most influential clinical attributes driving predictions, enhancing transparency and supporting clinical interpretability.

C. Parallel Multimodal Fusion of Image and Tabular Analysis

To further improve diagnostic reliability and exploit complementary information sources, this work proposes a parallel multimodal fusion strategy that integrates predictions from dermoscopic images and structured clinical metadata. The two modalities are processed independently using dedicated deep learning models, namely optimized ViT-DFNN for image-based analysis and TabNet for tabular prediction, ensuring modality-specific feature learning without cross-modal interference.

Let x^{img} denote a dermoscopic image and x^{tab} its corresponding clinical metadata. The image-based branch produces a class probability vector $\mathbf{p}^{img} \in \mathbb{R}^C$ using the ViT-DFNN architecture, while the tabular branch generates $\mathbf{p}^{tab} \in \mathbb{R}^C$ through TabNet, where C represents the number of diagnostic classes. Since both models operate in parallel and are optimized separately, the fusion is performed at the decision level to preserve interpretability and stability.

The final multimodal prediction is obtained through a weighted probabilistic aggregation defined as:

$$\mathbf{p}^{fusion} = \alpha \mathbf{p}^{img} + (1 - \alpha)\mathbf{p}^{tab}, \tag{18}$$

where $\alpha \in [0, 1]$ controls the relative contribution of each modality. Given the superior discriminative power of dermoscopic images, a higher weight is assigned to the image-based predictions, while tabular metadata act as a complementary source that refines and stabilizes the final decision. The predicted lesion class is then determined as:

$$\hat{y} = \arg\max_c \mathbf{p}^{fusion}(c). \tag{19}$$

This parallel fusion strategy enhances robustness by reducing uncertainty in visually ambiguous cases and reinforcing predictions with clinically relevant attributes. By combining high-level visual representations with structured patient information, the proposed multimodal framework achieves improved reliability and consistency, which are essential requirements for medical decision-support systems.

D. Implementation Details

Experiments were conducted using Python 3 on Google Colab with an NVIDIA A100 GPU.

For the image-based ViT-DFNN, the key hyperparameters were: Input images are resized to 224×224 pixels. Although preliminary experiments were conducted with 28×28 images, all reported results use 224×224 to better capture fine-grained dermoscopic features, which are critical for lesion classification and align with the standard input size for our method. Patch size $P = 6$, batch size 256, and learning rate 0.001. The images were augmented with random horizontal and vertical flips, as well as colour jitter (brightness, contrast, saturation ±0.2), and normalized using the HAM10000 dataset channel statistics (mean $[0.763, 0.546, 0.570]$, std $[0.141, 0.152, 0.169]$). The model was optimized using the AdamW optimizer with weight decay 10^{-4} and a cosine-annealing scheduler ($T_{\text{max}} = 350$). It comprise six self-attention heads, eight transformer encoder layers, and two DFNN layers. Training was performed for 350 epochs with early stopping (patience 20). L2 regularization [3] was applied with a weight coefficient of 0.01. A six-fold stratified cross-validation (6-FCV) strategy [4] was employed to ensure robust performance evaluation.

Algorithm 2 Parallel Multimodal Fusion Strategy

1: **Input:** Dermoscopic image x^{img}, tabular metadata x^{tab}
2: Compute $\mathbf{p}^{img}$ using ViT-DFNN
3: Compute $\mathbf{p}^{tab}$ using TabNet
4: Fuse predictions: $\mathbf{p}^{fusion} = \alpha \mathbf{p}^{img} + (1 - \alpha)\mathbf{p}^{tab}$
5: Predict class $\hat{y} = \arg\max \mathbf{p}^{fusion}$
6: **Return** $\hat{y}$

For the tabular-based TabNet model, the key hyperparameters were: feature and attention dimensions $n_d = 32$ and $n_a = 32$, number of decision steps $T = 6$, number of shared and independent layers $n_{\text{shared}} = 3$ and $n_{\text{independent}} = 3$, relaxation parameter $\gamma = 1.5$, and sparse regularization coefficient $\lambda_{\text{sparse}} = 10^{-4}$. The model was trained using the Adam optimizer with a learning rate of 2×10^{-3}, batch size 128, virtual batch size 64, and a maximum of 200 epochs with early stopping patience set to 30 epochs.

Weighted probabilistic fusion was employed to combine image- and tabular-based predictions. The fusion parameter α was optimized via grid search over $\{0.50, 0.55, \ldots, 0.90\}$ on the validation set, and the optimal value $\alpha = 0.70$ was used in the final model.

IV. Experimental Results and Discussion

In this section, we will present the HAM 10000 dataset of our method, the performance evaluation metrics, and the results obtained from our approach.

A. Dataset Description

1) HAM10000: HAM10000 dataset [16], [22] is a widely used and publicly available benchmark for skin lesion analysis. It contains 10,015 dermoscopic images collected from multiple clinical sources and annotated by expert dermatologists into seven diagnostic categories. The distribution of lesion classes in the image-based version of the dataset is illustrated in Figure 2. In addition to the HAM10000 provides structured CSV metadata. Each image is associated with a set of clinical

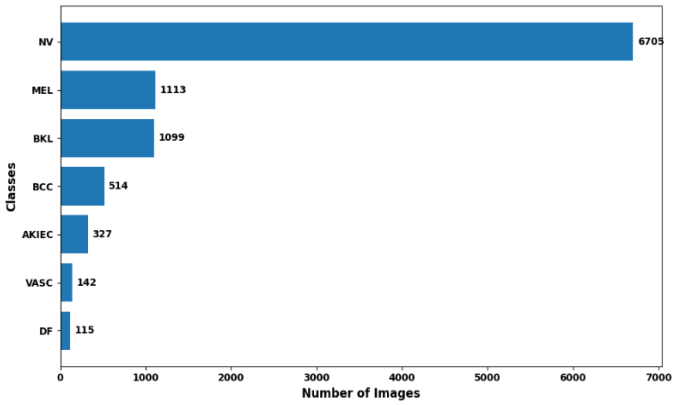

Fig. 2. Class distribution of dermoscopic images in the HAM10000 dataset.

attributes describing the lesion and the patient. The dataset includes seven variables, which are detailed in Table I.

TABLE I
DESCRIPTION OF FEATURES IN THE HAM10000 CSV METADATA.

Feature	Description
age	Age of the patient at the time of image acquisition
sex	Biological sex of the patient (male or female)
localization	Anatomical site of the lesion

2) ISIC 2019: ISIC 2019 [19], [21] dataset is a large-scale and publicly available benchmark for skin lesion classification, released as part of the International Skin Imaging Collaboration (ISIC) challenge. It contains over 25,331 dermoscopic images collected from multiple international clinical centers and annotated into eight diagnostic categories. The distribution of lesion classes in the dataset is illustrated in Figure 3.

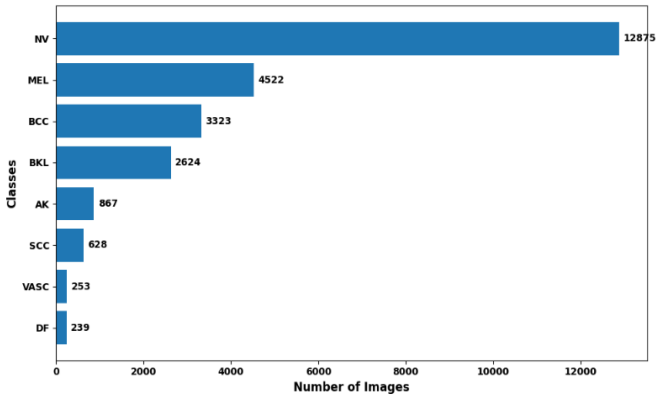

Fig. 3. Class distribution of dermoscopic images in the ISIC 2019 dataset.

In addition to image data, the ISIC 2019 dataset provides structured CSV metadata. Each image is associated with a set of basic clinical attributes describing the patient and lesion characteristics. Although the metadata is relatively limited compared to some medical datasets, it enables multimodal analysis. The available variables are summarized in Table II.

TABLE II
DESCRIPTION OF FEATURES IN THE ISIC 2019 CSV METADATA.

Feature	Description
image	Unique identifier of the dermoscopic image
age_approx	Approximate age of the patient
sex	Biological sex of the patient (male or female)
anatom_site_general	General anatomical location of the lesion

B. Performance Evaluation Metrics

We evaluated performance using accuracy, precision, specificity, sensitivity, Area Under the Curve (AUC), and the Geometric Mean (G-Mean). The G-Mean is particularly suitable for imbalanced classification problems, as it evaluates the balance between sensitivity and specificity:

$$\text{G-Mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}.$$

A higher G-Mean indicates balanced performance in identifying both positive and negative cases.

C. Results

The performance of our proposed multimodal framework is summarized in Table VI. By integrating dermoscopic image features with tabular clinical metadata, our method achieves high classification performance, reaching an accuracy of 93.15%, precision of 92.95%, sensitivity of 93.30%, specificity of 98.85%, G-mean of 96%, and AUC of 96.1%. These results demonstrate that the fusion of complementary information from images and clinical data effectively reduces uncertainty and enhances robustness, leading to improved lesion analysis.

TABLE III
PERFORMANCE COMPARISON OF UNIMODAL AND MULTIMODAL FUSION ON THE HAM10000 DATASET. THE PROPOSED FUSION STRATEGY ($\alpha = 0.7$) ACHIEVES THE BEST PERFORMANCE ACROSS ALL EVALUATION METRICS, DEMONSTRATING THE EFFECTIVENESS OF COMBINING DERMOSCOPIC IMAGES WITH CLINICAL METADATA.

Metric	ViT-DFNN	TabNet	Fusion ($\alpha = 0.7$)
Accuracy (%)	92.63	82.47	**93.15**
Precision (%)	92.49	82.70	**92.95**
Sensitivity (%)	92.74	82.20	**93.30**
Specificity (%)	98.77	84.10	**98.85**
G-Mean (%)	95.69	83.20	**96.00**
AUC (%)	95.90	86.50	**96.1**

D. Ablation Studies

This study investigates the performance contributions of components within our framework on the HAM10000 dataset, evaluating 1) the optimized ViT-DFNN and 2) TabNet, with Explainable AI techniques applied to both models.

1) ViT-DFNN Performance: As presented in Table IV using the same evaluation metrics (accuracy, precision, sensitivity, specificity, G-Mean, and AUC), the baseline ViT achieved 88.32%, 88.29%, 88.38%, 98.05%, 93.08%, and 89.33%, respectively. After integrating the DFNN, the ViT-DFNN improved these results to 91.30%, 91.29%, 91.39%, 98.55%, 94.90%, and 95.90%. A discrepancy greater than 5% between

training and test accuracy (97.85%(train), 91.30%(test)), is a strong indicator of overfitting. We addressed this problem by incorporating L2 regularization. In addition, we reinforced robustness through the k-FCV technique, where the dataset is split into k subsets and the model is iteratively trained and tested on different folds. Across multiple experiments with varying values of k, the most consistent and optimal results were achieved when $k \geq 6$.

An empirical study was also conducted to evaluate the impact of the number of fuzzy rules (M) in the DFNN. Various values of M were tested, and the configuration yielding the best classification performance on the validation set was selected. Specifically, $M = 5$ provided the most favorable balance between performance, and stability across folds, and was therefore adopted for all experiments. This empirical selection ensures that the fuzzy component effectively refines patch-level features without introducing unnecessary complexity.

Thanks to our optimization, the proposed ViT-DFNN achieves superior performance across all evaluation metrics, as detailed in Table V. The model attained an average test accuracy of 92.63%, precision of 92.49%, sensitivity of 92.74%, specificity of 98.77%, G-mean of 95.69%, and AUC of 95.90%.

TABLE IV
ABLATION STUDY COMPARING BASELINE VIT AND THE PROPOSED VIT-DFNN ON THE HAM10000 DATASET. THE INTEGRATION OF THE DFNN CONSISTENTLY IMPROVES ALL EVALUATION METRICS, DEMONSTRATING THE EFFECTIVENESS OF FUZZY-BASED FEATURE REFINEMENT.

Metric	ViT	ViT-DFNN
Accuracy (%)	88.32	91.30
Precision (%)	88.29	91.29
Sensitivity (%)	88.38	91.39
Specificity (%)	98.05	98.55
G-Mean (%)	93.08	94.90
AUC (%)	89.33	95.90

As shown in Figure 4, Score-CAM demonstrates strong performance in highlighting relevant image regions. It consistently and accurately emphasizes critical features such as asymmetric borders, color irregularities, and internal lesion texture. This precise and stable localization indicates that the model's predictions are grounded in clinically meaningful characteristics, confirming the reliability of the network.

2) TabNet Performance: As shown in Table III, the TabNet model achieved 82.47% accuracy, 82.70% precision, 82.20% sensitivity, 84.10% specificity, G-Mean of 83.20%, and AUC of 86.50%. Although its performance is lower than the image-based ViT-DFNN, TabNet effectively captures structured clinical features such as demographic information and lesion localization. These features provide complementary insights that are not directly observable from dermoscopic images, helping to stabilize predictions in ambiguous cases and contributing to the overall improvement observed in the multimodal fusion framework.

As shown in Figure 5, the visualization highlights the most influential clinical features for skin lesion classification.

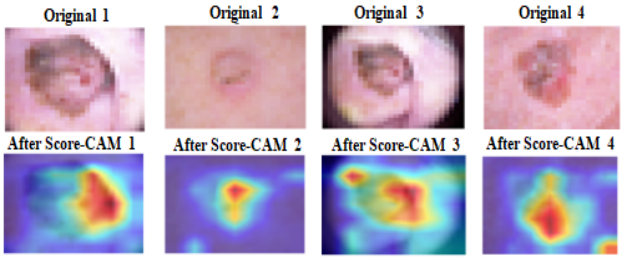

Fig. 4. Score-CAM visualization of relevant dermoscopic image regions for skin lesion classification using the ViT-DFNN on the HAM10000 dataset. The highlighted areas correspond to clinically meaningful features such as asymmetric borders, color irregularities, and internal lesion texture, confirming that the model's predictions are based on relevant visual cues and enhancing explainability and reliability.

Specifically, *age* and *localization* are identified as the most significant contributors, while *sex* plays a secondary role. This distribution confirms that TabNet effectively prioritizes clinically relevant features, thereby supporting reliable and interpretable decision-making from tabular patient data.

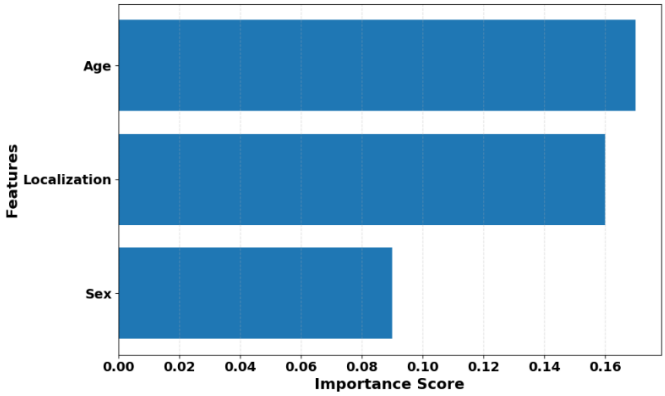

Fig. 5. GFI for the TabNet model on the HAM10000 dataset. The visualization highlights the most influential clinical features for skin lesion classification, with *age*, *localization*, and *sex* contributing the most.

E. Discussion

The findings highlight the benefit of a multimodal approach combining dermoscopic images and clinical metadata, which improves feature relevance and reduces classification uncertainty. XAI was applied to both models, ensuring transparency and helping clinicians understand the contribution of visual and clinical features.

Our experiments show a clear progressive improvement in classification performance across the evaluated models. Integrating the Deep Fuzzy Neural Network (DFNN) with the Vision Transformer (ViT) led to an approximate 3% improvement in overall performance metrics compared to the baseline ViT model. Further combining dermoscopic image features with tabular clinical metadata in our multimodal framework (ViT-DFNN + TabNet) produced an additional 2% improvement over the ViT-DFNN alone. These incremental gains highlight the effectiveness of feature refinement through

TABLE V

PERFORMANCE OF THE PROPOSED VIT-DFNN ACROSS 6-FCV ON THE HAM10000 DATASET. THE MODEL DEMONSTRATES STABLE AND CONSISTENT PERFORMANCE ACROSS FOLDS, CONFIRMING THE ROBUSTNESS AND GENERALIZATION CAPABILITY OF THE APPROACH.

Metric	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**Avg. ± CI**
Train Accuracy (%)	96.27	96.85	95.43	97.72	98.84	97.88	**97.16** ± 0.42
Test Accuracy (%)	92.45	92.74	92.51	92.57	92.83	92.71	**92.63** ± 0.38
Precision (%)	92.35	92.47	92.44	92.51	92.63	92.56	**92.49** ± 0.31
Sensitivity (%)	92.60	92.70	92.50	92.85	92.90	92.85	**92.74** ± 0.34
Specificity (%)	98.70	98.80	98.60	98.85	98.90	98.80	**98.77** ± 0.28
G-Mean (%)	95.53	95.73	95.30	95.80	95.90	95.90	**95.69** ± 0.35
AUC (%)	95.80	95.90	95.60	96	96.10	96.00	**95.90**± 0.14

TABLE VI

COMPARISON OF EXISTING MULTIMODAL FRAMEWORKS FOR SKIN LESION ANALYSIS. THE TABLE SUMMARIZES KEY INFORMATION INCLUDING YEAR, REFERENCE, METHOD, DATASET, AND REPORTED PERFORMANCE METRICS. OUR PROPOSED FRAMEWORK ACHIEVES COMPETITIVE PERFORMANCE ACROSS MULTIPLE DATASETS, DEMONSTRATING ITS EFFECTIVENESS IN COMBINING DERMOSCOPIC IMAGES WITH CLINICAL METADATA.

Year	**Ref.**	**Method**	Dataset	**Results**
2026	[13]	ViT+MLP	PAD-UFES-20	F1 score: 89%, G-mean: 90%
2025	[7]	CNN + FCN	HAM 10000	Accuracy: 94.5%, F1 Score: 94%
2025	[12]	ViT+MLP	ISIC 2019	Accuracy: 87%, F1 score: 92%
2025	[10]	CNNs	ISIC-DICM-17K	AUC: 88.51%
2024	[6]	ALBEF	HAM 10000	Accuracy: 94.11%, AUCROC: 94.26%
2022	[9]	MMF-Net	PAD-UPES-20	Accuracy: 76.8%, Balanced accuracy: 77.5%AUC: 94.7%
2022	[11]	ViT + SLE	Private, ISIC 2018	Accuracy: 81.6% (private), Accuracy: 93.81% and AUC: 99% (ISIC 2018)
2021	[8]	IM-CNN	HAM 10000	Accuracy: 95,1%, Sensitivity: 83,5%, Specificity: 93,2%, AUC: 97,8%, AUC_SEN_80: 96,0%.
Our Framework			HAM 10000	Accuracy: **93.15%**, Precision: **92.95%**, sensitivity: **93.30%**, Specificity: **98.85%**, G-mean: **96%**, AUC = **96.1%**.

DFNN and the complementary value of multimodal data.

To ensure the reliability of the reported metrics for ViT-DFNN, we computed the 95% confidence intervals (CI) across the six folds for each performance indicator. The narrow CI (±0.25–0.45) demonstrates the stability of the model and consistent generalization across different subsets. While full K-fold cross-validation was not applied to the multimodal framework, the consistent improvements over the ViT-DFNN baseline suggest robust performance gains.

To ensure the robustness and generalization capability of the proposed method, we further evaluated its performance on the ISIC 2019 dataset using a fixed image size of 224×224. This additional experiment aims to validate the model's ability to generalize across different datasets and conditions. Despite the inherent limitations of the ISIC 2019 dataset, particularly the lack of rich metadata, the proposed approach achieved strong performance. Specifically, it obtained an accuracy of 91.88%, a precision of 91.83%, a sensitivity of 91.02%, a specificity of 95.84%, and a G-mean of 93.46%. These results demonstrate that the proposed method maintains high effectiveness even under constrained data conditions, confirming its strong generalization capability and robustness in real-world clinical scenarios.

Table VI summarizes the performance of our framework alongside existing multimodal methods for skin lesion analysis. While the datasets and evaluation protocols vary across these studies, several trends can be observed. Our framework achieves high performance, demonstrating robust and generalizable classification performance.

Compared to models evaluated on HAM10000 [6]–[8], our method maintains competitive accuracy while providing better balance across sensitivity, specificity, and G-mean, highlighting the effectiveness of multimodal feature integration. When compared with methods evaluated on other datasets [9]–[13], our framework shows superior overall robustness and more consistent performance across lesion types, suggesting strong generalization capabilities.

Overall, this comparison indicates that integrating complementary image and tabular metadata in our framework improves both predictive reliability and adaptability to diverse lesion patterns, even when evaluated against methods on different datasets.

V. CONCLUSION

In this study, we presented a novel multimodal framework for reliable skin lesion classification that combines dermoscopic image analysis with structured clinical metadata. By leveraging a hybrid ViT–DFNN for image representation and a TabNet model for tabular clinical data, our approach effectively fuses heterogeneous information to enhance classification performance. The integration of XAI techniques, including Score-CAM and GFI, ensures transparency and interpretability of the model's predictions, which is critical for clinical applications.

Experimental results on the HAM10000 dataset demonstrate that the proposed framework achieves high performance, outperforming traditional image-only or tabular-only methods. These findings highlight the practical potential of the framework in supporting clinicians with reliable decision-making, reducing diagnostic errors, and improving patient outcomes. Moreover, the ability to combine visual and clinical information provides a template for developing trustworthy AI systems

in other medical imaging tasks.
Future work will focus on evaluating the proposed framework across multiple skin lesion datasets and real-world clinical environments to further validate its generalizability and robustness. Additionally, we plan to introduce a unified explainability method capable of simultaneously interpreting both the image and tabular metadata components, providing enhanced confidence and transparency in clinical decision-making.

VI. Acknowledgment

This work would not have been possible without the support of the Ministry of Higher Education and Scientific Research of Tunisia (LR11ES48). All these research topics align with the objectives of the joint Turkish-Tunisian project proposal entitled "REMEDIA," which was accepted for the 2024 Call for Proposals.

References

[1] Kumar, V., Jain, I., Dongla, D. S., Rai, S., Rastogi, V., Kaushik, E., & Kaushik, R. "Dermatological Diagnostics: A Unified Deep Learning Framework for Skin Lesion and Cancer Classification." 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N). IEEE, 2024.

[2] Bouzidi, S., Jdey, I., & Drira, F. "Towards Explainable Skin Cancer Diagnosis: A Vision Transformer Approach with Grad-CAM Visualization." 2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025.

[3] Bilgic, B., Chatnuntawech, I., Fan, A. P., Setsompop, K., Cauley, S. F., Wald, L. L., & Adalsteinsson, E. "Fast image reconstruction with L2-regularization." Journal of magnetic resonance imaging 40.1 (2014): 181-191.

[4] Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. "The'K'in K-fold Cross Validation." Esann. Vol. 102. 2012.

[5] Deng, Y., Ren, Z., Kong, Y., Bao, F., & Dai, Q. (2016). "A hierarchical fused fuzzy deep neural network for data classification." IEEE Transactions on Fuzzy Systems 25.4 (2016): 1006-1012.

[6] Adebiyi, A., Abdalnabi, N., Smith, E. H., Hirner, J., Simoes, E. J., Becevic, M., & Rao, P. "Accurate skin lesion classification using multimodal learning on the ham10000 dataset." MedRxiv (2024): 2024-05.

[7] Ahmad, M., Ahmed, I., Chehri, A., & Jeon, G. "Fusion of metadata and dermoscopic images for melanoma detection: Deep learning and feature importance analysis." Information Fusion (2025): 103304.

[8] Wang, S., Yin, Y., Wang, D., Wang, Y., & Jin, Y. "Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis." IEEE transactions on cybernetics 52.12 (2021): 12623-12637.

[9] Ou, C., Zhou, S., Yang, R., Jiang, W., He, H., Gan, W., ... & Li, J. "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata." Frontiers in Surgery 9 (2022): 1029991.

[10] Ahammed, S., Cui, X., Lu, W., & Yap, M. H. (2025). "Skin Lesion Classification Using Dermoscopic Images and Clinical Metadata: Insights from Multimodal Models." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

[11] Cai, G., Zhu, Y., Wu, Y., Jiang, X., Ye, J., & Yang, D. "A multimodal transformer to fuse images and metadata for skin disease classification." The Visual Computer 39.7 (2023): 2781-2793.

[12] Pavan, S., Bhanusree, T., Varma, A. A., Sucharita, S., & Jabbar, M. A. "Skin Lesion Classification Using Vision Transformers and Clinical Metadata." 2025 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). IEEE, 2025.

[13] Yasmin, I., Sultana, S., Akter, S., Cao, W., & Patwary, M. J. A. "Vision transformer-based fuzzy semi-supervised multimodal learning for skin lesion image classification." Soft Computing (2026): 1-19.

[14] Bouzidi, S., Hcini, G., Jdey, I., & Drira, F. "Synergizing CNNs and ViTs: A Survey on Hybrid Models for Fashion Image Classification." Revista de Informática Teórica e Aplicada 33.1 (2026): 11-27.

[15] Bouzidi, S., Jdey, I., & Drira, F. "Xg-vit: Explainable and generalizable vision transformer for benchmark image classification." in International Conference on Verification and Evaluation of Computer and Communication Systems (VECoS). sn, 2025.

[16] Hamdi, M., Bouzidi, S., Jdey, I., Drira, F., & Yanikoglu, B. "Auto-Optimized Parameter-Efficient Fine-Tuning for Skin Lesion Classification with Vision Transformers." 2025 IEEE 9th Forum on Research and Technologies for Society and Industry (RTSI). IEEE, 2025.

[17] Jdey, I., Toumi, A., Dhibi, M., & Khenchaf, A. "The contribution of fusion techniques in the recognition systems of radar targets." IET International Conference on Radar Systems (Radar 2012). Stevenage UK: IET, 2012.

[18] Hcini, G., Jdey, I., & Ltifi, H. "HSV-Net: a custom cnn for malaria detection with enhanced color representation." 2023 International Conference on Cyberworlds (CW). IEEE, 2023.

[19] Gayatri, E., & Aarthy, S. L. "Reduction of overfitting on the highly imbalanced ISIC-2019 skin dataset using deep learning frameworks." Journal of X-Ray Science and Technology 32.1 (2024): 53-68.

[20] Goyal, M., Knackstedt, T., Yan, S., & Hassanpour, S. "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities." Computers in biology and medicine 127 (2020): 104065.

[21] Aktan, A. T., Almouradi, A., Aptoula, E., & Yanikoglu, B. "Comparison of Leading Deep Learning Architectures on Skin Lesion Classification Using the ISIC 2019 Dataset." 2025 Medical Technologies Congress (TIPTEKNO). IEEE, 2025.

[22] Velaga, N., Vardineni, V. R., Tupakula, P., & Pamidimukkala, J. S. "Skin cancer detection using the HAM10000 dataset: A comparative study of machine learning models." 2023 Global conference on information technologies and communications (GCITC). IEEE, 2023.

[23] Yap, J., Yolland, W., & Tschandl, P. "Multimodal skin lesion classification using deep learning." Experimental dermatology 27.11 (2018): 1261-1267.

[24] Bouzidi, S., Jdey, I., & Alimi, A. "A vision transformer approach with L2 regularization for sustainable fashion classification." Available at SSRN 4686032.