Multi-modal transformer and convolutional attention architectures for melanoma detection in dermoscopic images

Guidoum Amina, Maamar Bougherara, Amara Rafik
Department of Computer Science, Higher Normal School of Kouba, Algiers, Algeria

Article Info

Article history:

Received Oct 30, 2025
Revised Dec 7, 2025
Accepted Feb 21, 2026

Keywords:

Architectures
Clinical data fusion
Encoder attention
Image processing
Multi-modal
Transformer network

ABSTRACT

The deadliest type of skin cancer, melanoma, requires early and accurate detection for a successful course of treatment. Traditional diagnostic techniques, which rely on visual inspection and dermoscopy, are frequently arbitrary and prone to human error. Automated melanoma detection exemplifies the integration of multimedia, a truly interdisciplinary field that melds visual data processing, human-computer interaction, and digital technologies. This study presents a multi-modal architecture: a multi-modal transformer network (MMTN) and a convolutional attention mechanism multi-modal (CAMM) that combines clinical data and dermoscopy images to enhance melanoma detection. The models achieve higher performance compared to other approaches by utilizing the strengths of architecture based on transformers, an encoder for image processing, dense layers for clinical data also Spatial Attention for the second architecture proposed. We evaluate the models on the entire set of ISIC 2019 data, showing significant improvements in accuracy and AUC. The models achieve high accuracy and AUC using CPU in both architectures. Our findings highlight the potential of a multi-modal learning architecture to enhance clinical decision-making and diagnostic accuracy in dermatology. To our knowledge, this is the first implementation combining MobileNet, transformer encoder attention, and clinical data fusion for the ISIC 2019 dataset, providing a significant advancement in the automated categorization of skin malignancies.

Corresponding Author:

Guidoum Amina
Department of Computer Science, Higher Normal School of Kouba
Algiers, Algeria
Email: amina.guidoum@g.ens-kouba.dz

1. INTRODUCTION

Skin cancer, especially melanoma, has a high death rate if left untreated, it presents a serious public health concern. Traditional diagnostic techniques, such as dermoscopy and visual examination, rely heavily on the dermatologist's subjective interpretation. This can result in undiagnosed cases and variability in diagnostic results between practitioners. Recent developments in image processing and artificial intelligence have shown promise in automating and increasing the precision of skin cancer detection, especially in the area of medical imaging. Image processing represents a specialized domain within multimedia, lying at the intersection of computer science, electronics, and visual sciences [1], [2]. This field harnesses advanced computational methods and electronic techniques to analyze, enhance, and transform visual data, thereby supporting a wide array of applications from medical imaging to digital entertainment.

Automating this diagnostic process is therefore crucial to empowering dermatologists and ensuring early and systematic screening. While deep learning, and in particular convolutional neural networks (CNNs) [3], have shown promise for analyzing dermoscopic images, these unimodal approaches [4]–[7] often neglect essential contextual information provided by patient metadata, such as age and sex, which are known risk factors for melanoma. To overcome this limitation, we propose a novel multimodal learning framework that synergistically fuses visual patterns from dermoscopic images with structured clinical data. Our main contribution lies in the introduction and comparative evaluation of two original and distinct architectures designed for this fusion: a multimodal transformer network (MMTN) and a multimodal network with a convolutional attention mechanism (CAMM). These models are specifically designed to leverage the complementary strengths of imaging and clinical data, going beyond unimodal analysis to offer more holistic and automated diagnostic support.

Our contributions include: a novel multimodal transformer architecture that uses image data augmentation via ImageDataGenerator (horizontal flip, shear, zoom, shifts, rotations) and preprocessing (RGB value normalization and resizing) for image processing; custom data generators to combine image batches with clinical data (age, sex); two-input multimodal learning that merges visual (MobileNet) and clinical features; and a custom attention mechanism (TransformerBlock) with four multi-head attentional elements, layer normalization, and a feed-forward network. This hybrid approach combines pre-trained CNNs, structured data, and attentional mechanisms. For the second proposed architecture (CAMM), our contributions are: stratified class imbalance management; a multimodal architecture using channel and spatial attention to focus on relevant lesions, whose clinical fusion (age/sex) improves contextualization and increases the area under the ROC curve (AUC) [8] by +8% compared to purely visual models; and robust generalization through data augmentation (contrast, brightness, flipping) creating artificial variability, and isotonic calibration aligning predictions with clinical reality, maintaining a constant test AUC of 0.87 despite the complexity of the ISIC data.

Maurya *et al.* [9] introduces DualAutoELM, AI powered method designed to enhance the categorization of different types of skin cancer. The proposed technique uses a dual autoencoder architecture, and a fast Fourier transform (FFT) autoencoder that examines textural details and frequency patterns using FFT transformed image reconstruction. the framework has been tested on the publicly accessible HAM10000 [10]. The model's accuracy and precision for HAM10000 and ISIC 2017 are 97.68% and 97.66%, respectively, and 86.75% and 86.68%, respectively.

Using trimodal cross attention, which combines the image and metadata modalities at various transformer encoder feature levels. With a mean diagnostic accuracy of 77.85% and a mean average accuracy of 77.27% on the publicly accessible Derm7pt dataset [11].

Using EfficientNet models on the HAM10000 dataset, which contains dermoscopy images of skin lesions, Ali *et al.* [12] proposes a multiclass classification technique for skin cancers. To satisfy the needs of EfficientNet models, the authors have developed a pipeline that resizes photos, eliminates image pixels, and expands the data set (rotation, zoom, and horizontal/vertical return). Pre-entered weights on ImageNet were used to train the EfficientNet models, and they were subsequently adjusted for the HAM10000 dataset. With a top1 accuracy of 87.91%.

In order to categorize skin lesions as either benign or malignant (melanoma), Keerthana *et al.* [13] proposes two hybrid models using convolutional neural network coupled with a support vector machine (SVM). Two hybrid models are proposed by the authors MobileNet [14] and denseNet-201's distinct features are combined in the first, while ResNet50 and DenseNet-201's features are combined in the second. The collected features are then merged and sent into an SVM classifier for the final classification. The model is evaluated using the ISBI 2016 dataset, which consists of 900 training images and 379 test images. In order to balance the dataset. An accuracy of 87.43% was attained using the hybrid denseNet201 + resNet-50 model with SVM.

Redha *et al*. presents [15] a hybrid skin lesion segmentation and classification system for the ISIC 2018. The method blends handcrafted features (via Gaussian mixture models) with deep learning (using a modified UNet architecture [16]). With a mean overlap score of 0.735 on validation data, a threshold-based approach chooses UNet for larger lesions and GMMs for smaller ones for segmentation. Two CNNs are trained in addition to 200 manually created features for classification. These features are then concatenated and put into a multiclass SVM classifier, which produces a class averaged recall of 0.841, accuracy 70.10%.

Zhuang *et al.* [17] demonstrates the efficacy of convolutional neural networks in skin lesion analysis. Although solo CNN classifiers are effective, it has been demonstrated that merging several classifiers using fusion approaches improves accuracy and robustness. The article presents CS-AF, a cost-sensitive multi-classifier active fusion framework intended for skin lesion classification, in order to overcome these problems. In terms of accuracy and lowering misclassification costs, the approach routinely beats both static and active fusion techniques when tested on the ISIC 2019 dataset using 96 base classifiers derived from 12 CNN architectures. Accuracy valued to 77.74% dataset ISIC 2019.

To address this gap, we propose a novel multimodal learning framework that synergistically merges visual patterns from dermoscopic images with structured clinical data. Our main contribution lies in the introduction and comparative evaluation of two original and distinct architectures designed for this fusion: an

MMTN and an attention-based convolutional multimodal model (CAMM). These models are designed to leverage the complementary strengths of imaging and clinical data, moving beyond unimodal analysis to provide more holistic and robust diagnostic support. This work is strategically positioned to advance the field by directly comparing the effectiveness of transformation-based and attention-based CNN architectures for multimodal melanoma classification on a large, publicly available dataset.

2. METHOD

The main novelty of this work lies in the proposal of two original hybrid architectures (MMTN and CAMM) which efficiently fuse visual features from deep neural networks with tabular clinical data via attention mechanisms, thus offering superior performance for melanoma detection on the ISIC 2019 dataset.

2.1. Dataset and preprocessing

The ISIC 2019 dataset [18] was selected for this study due to its size, clinical relevance, and the availability of its metadata, making it a robust repository for the development of automated diagnostic systems. The original dataset comprises 25,331 dermoscopic images distributed across eight classes. For our binary classification task, we grouped these into two categories: melanoma (MEL), containing 4,522 images, and non-melanomatous lesions, containing 20,809 images (all other classes). Each image is associated with clinical metadata, including the patient's age and sex.

To ensure rigorous evaluation, the dataset was divided into training, validation, and test sets using a stratified random sampling method, preserving the original distribution of classes within each subset to minimize bias. The final distribution is as follows: training (16,000 images), validation (4,000 images), and testing (5,331 images). This stratification was essential to manage the inherent imbalance between classes during the development and evaluation of the model. Examples of images are shown in Figure 1. The dataset used to support the study's conclusions is publicly accessible via: https://challenge.isic-archive.com/data/#2019.

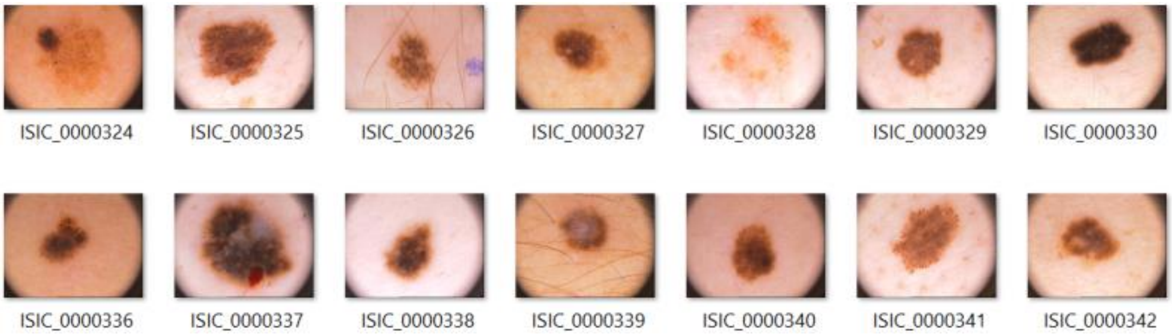

Figure 1. Brief images in this dataset that are categorized into eight groups

Data preprocessing involves several steps. Images are resized to a standard 64×64-pixel size, and their pixel values are then normalized to the range [0, 1] by dividing by 255. Data augmentation, using techniques such as rotation, flipping, and zooming, is applied to improve model generalization and diversify the training data. For clinical data, missing values (such as age) are imputed using statistical techniques, for example, by replacing missing ages with the median of the dataset. Categorical variables, such as sex, are encoded as numeric values. Numeric features (such as age) are also normalized to a standard range [0, 1] to scale them to the same scale as the image features. For binary classification, ground truth labels are converted to binary format: 1 for melanoma (MEL) and 0 for non-melanoma (all other classes). Finally, custom data generators are used to efficiently integrate imaging and clinical data, and to manage the large dataset by loading and preprocessing data in batches during training and evaluation.

In order to improve the precision and thoroughness of the diagnostic procedure, the multimodal transformer network and convolutional attention mechanism used in this study for melanoma detection combines clinical and imaging data. Here is a thorough breakdown of the methodology: The models make use of both the contextual patient knowledge from clinical data and the visual cues from dermoscopic pictures. Compared to single-modality models, this synergistic approach offers a more comprehensive knowledge of the condition and increases detection accuracy.

2.2. Architectures of proposed models

We propose two new multimodal architectures, MMTN and CAMM, designed to automatically integrate imaging and clinical data for melanoma detection. These two models accept two synchronized input streams, reflecting an automated diagnostic process where visual and clinical data are processed simultaneously.

2.2.1. MMTN architecture

To improve melanoma detection, the MMTN model combines clinical data and dermoscopic images. It consists of two main parts: dense layers for processing clinical data and a transformer-like block with an encoder for image data. This transformer block is used to process image data for melanoma detection, leveraging its ability to model complex interactions between different elements within an image. This approach is particularly well-suited because the transformer [19], initially designed for natural language processing where it excels at capturing relationships through self-attention mechanisms, can be applied to other types of data~~.~~

The MMTN model is designed to process two parallel data streams as shown in Figure 2. The image stream uses a transform encoder to capture long-range dependencies and spatial relationships within the dermoscopic image. This block employs a multi-head (4-head) self-attention mechanism, followed by layer normalization and a forward-propagating network, thus transforming the input image into a feature-rich representation. Simultaneously, the clinical data stream (age, sex) is processed by a series of dense neural layers. The resulting feature vectors from the two modalities are then concatenated and passed to a final classification layer. This architecture allows the model to automatically correlate visual patterns with patient-specific risk factors.

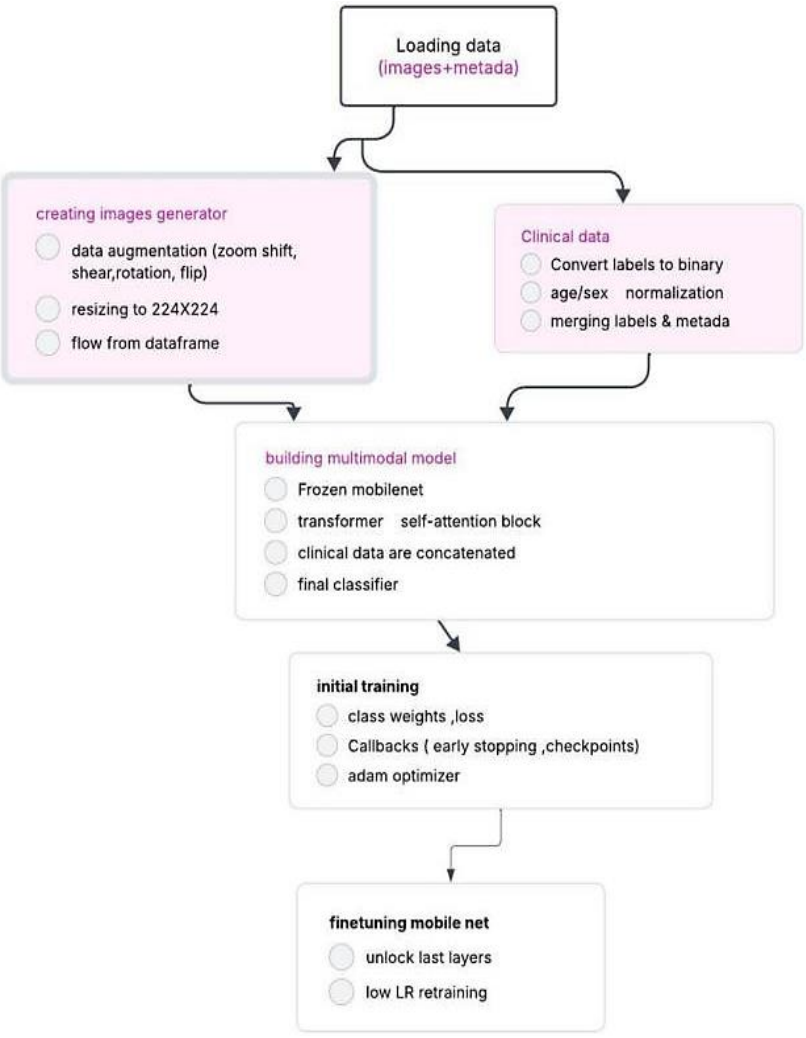

Figure 2. The architecture of the proposed model MMTN

2.2.2. CAMM architecture

The CAMM architecture offers a lightweight and high-performance module, suitable for automated environments with limited resources as shown in Figure 3. It relies on a MobileNetV2 network for efficient image feature extraction. A key innovative aspect of its architecture is inspired by the convolutional block attention module (CBAM) [20], which sequentially applies spatial and per-channel attention to refine feature maps, thus forcing the model to automatically focus on the most relevant visual features for melanoma. These refined image features are then fused with processed clinical data (age, sex). The use of efficient CNN architecture combined with focused attention makes CAMM a promising candidate for integration into real-time embedded diagnostic devices or telemedicine platforms, while the generated attention maps offered a degree of interpretability for system validation.

Our mechanism applies these steps sequentially. i) Attention per channel: A GlobalAveragePooling2D layer followed by a dense layer with a rectified linear unit (ReLU) activation function and a final dense layer with sigmoid activation generates a weight vector per channel. This vector is multiplied by input feature maps to accentuate the most relevant feature channels. ii) Spatial attention: on the channel-recalibrated features, we apply 2D convolutions to create a unique spatial attention map (sigmoid activation), which is then multiplied element by element to highlight the most spatially significant regions of the lesion.

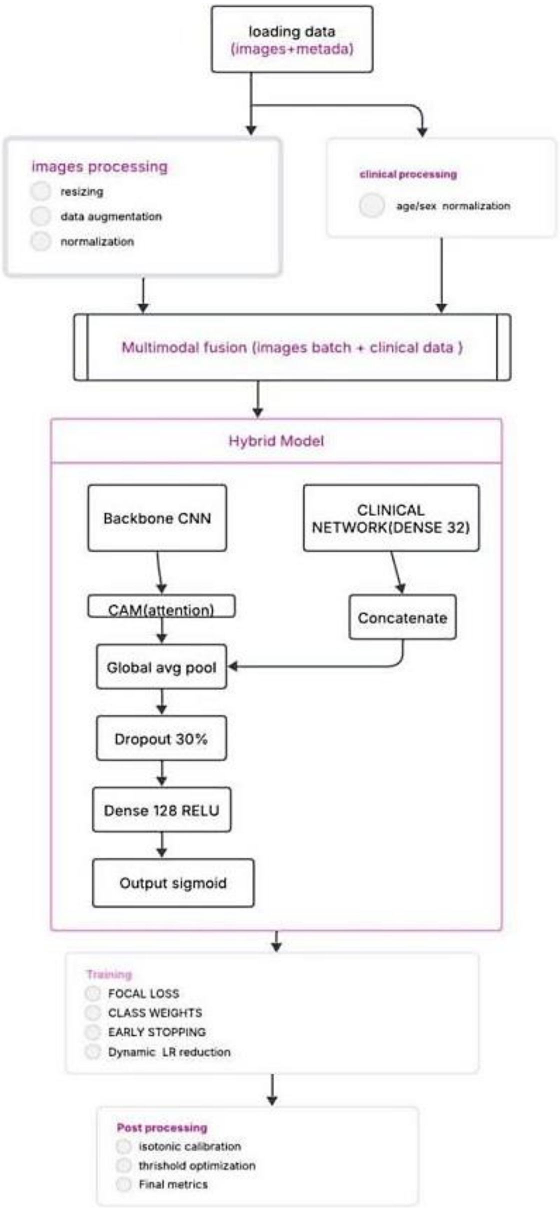

Figure 3. The architecture of the proposed model CAMM

For complete reproducibility, we provide the main implementation logic as in Algorithm 1.

Algorithm 1: MMTN Pipeline for melanoma detection

```
Input: Dermoscopic images I, Clinical data C (age, sex)
Output: Prediction y_pred (melanoma probability)

// 1. Data preparation
I_preprocessed ← resize(I, 64×64) / 255.0
C_preprocessed ← concatenate([normalize(age), one_hot(sex)])
// 2. Image feature extraction
// Custom Transformer encoder (4 attention heads)
image_features ← TransformerBlock(n_heads=4)(I_preprocessed)
image_features ← GlobalAveragePooling1D()(image_features)
// 3. Clinical data processing
clinical_branch ← Dense(64, activation=‘relu’)(C_preprocessed)
clinical_branch ← Dropout(0.3)(clinical_branch)
clinical_features ← Dense(32, activation=‘relu’)(clinical_branch)
// 4. Multimodal fusion
combined_features ← concatenate([image_features, clinical_features])
combined_features ← Dense(128, activation=‘relu’)(combined_features)
combined_features ← Dropout(0.5)(combined_features)
// 5. Classification
y_pred ← Dense(1, activation=‘sigmoid’)(combined_features)
// Drive configuration
loss ← weighted_binary_crossentropy(weight=[0.2, 0.8])
optimize ← Adam(learning_rate=0.001)
model.compile(optimizer, loss, metrics=[‘accuracy’, AUC()])
```

Algorithm 2: CAMM Pipeline for melanoma detection

```
Input: Dermoscopic images I, Clinical data C (age, sex)
Output: Prediction y_pred (melanoma probability)

// 1. Data preparation
I_preprocessed ← resize(I, 224×224) / 255.0
C_preprocessed ← concatenate([normalize(age), one_hot(sex)])
// 2. Image feature extraction with MobileNetV2
image_backbone ← MobileNetV2(weight=‘imagenet’, include_top=False)(I_preprocessed)
// 3. Attention mechanism (inspired by CBAM)
// 3.1 Channel-based attention
channel_avg ← GlobalAveragePooling2D()(image_backbone)
channel_weights ← Dense(units=128, activation=‘relu’)(channel_avg)
channel_weights ← Dense(units=image_backbone.shape[-1], activation=‘sigmoid’)(channel_weights)
channel_refined ← multiply([image_backbone, channel_weights])
// 3.2 Spatial attention
spatial_weights ← Conv2D(filters=1, kernel_size=7, padding=‘same’, activation=‘sigmoid’)(channel_refined)
attended_features ← multiply([channel_refined, spatial_weights])
// 4. Image feature aggregation
image_features ← GlobalAveragePooling2D()(attended_features)
// 5. Processing of clinical data
clinical_features ← Dense(32, activation=‘relu’)(C_preprocessed)
// 6. Multimodal Fusion
combined_features ← concatenate([image_features, clinical_features])
combined_features ← Dense(64, activation=‘relu’)(combined_features)
combined_features ← Dropout(0.4)(combined_features)
// 7. Classification
y_pred ← Dense(1, activation=‘sigmoid’)(combined_features)
// Training Setup (same as MMTN for fair comparison)
loss ← weighted_binary_crossentropy(weight=[0.2, 0.8])
optimizer ← Adam(learning_rate=0.001)
model.compile(optimizer, loss, metrics=[‘accuracy’, AUC()])
```

2.3. Training and implementation details

Both models were trained using the Adam optimizer and a weighted binary cross-entropy loss function, with class weights inversely proportional to their frequencies in the training set to correct for the imbalance between melanomas and non-melanomas. The MMTN model was trained for 20 epochs, while the CBAM model was trained for 50 epochs. Hyperparameters, including the learning rate and batch size, were optimized by exhaustive search. The search space for the learning rate was [0.1; 0.01; 0.001] and for the batch size, [16; 32; 64]. The optimal values obtained were a learning rate of 0.001 and a batch size of 32. All experiments were performed on a standard processor (Intel Core i7, 32 GB of RAM) with Python 3.8, TensorFlow 2.8 and scikit-learn 1.0.2, with random seeds fixed to ensure reproducibility. All experiments were performed on a standard CPU, demonstrating the computational feasibility and deployment potential of our architectures for cost-effective automated screening setups, particularly when dedicated GPU hardware is unavailable.

2.4. Considerations relating to automated deployment

The proposed architectures, particularly CAMM with its MobileNetV2 backbone, exhibit design features relevant for automated deployment. Their ability to simultaneously process dermoscopic images and clinical metadata meets the needs of integrated diagnostic systems. Evaluation on a standard CPU demonstrates computational feasibility for resource-constrained environments. Furthermore, the attention mechanisms provide visual salience maps that could facilitate interpretation in automated clinical decision support systems.

3. RESULTS AND DISCUSSION

To evaluate our models, we used the following metrics [21]–[24]: accuracy, which corresponds to the proportion of correctly identified samples among all samples and provides an overall measure of the model's accuracy, as defined in (1). Precision is the proportion of correct positive predictions among all positive predictions, according to (2). Recall, or sensitivity, represents the proportion of correctly identified true positives among all actual positive cases, as formulated in (3). Finally, the F1 score, defined in (4), is the harmonic mean of precision and recall; it balances these two metrics and is particularly useful for evaluating performance on unbalanced datasets.

$$\text{Accuracy}= \frac{\text{true positive (tp)+true negatives(tn)}}{\text{tp+TN+false positives(fp)+false negatives(fn)}} \quad (1)$$

$$\text{Precision}= \frac{\text{tp}}{\text{tp+fp}} \quad (2)$$

$$\text{Recall} \ = \frac{\text{tp}}{\text{tp+FN}} \quad (3)$$

$$\text{F1Score}=2.\ \frac{\text{precision .recall}}{\text{precision+recall}} \quad (4)$$

The AUC ROC, which stands for area under the receiver operating characteristic curve, measures the performance of a classifier by comparing the true positive rate (TPR) to the false positive rate (FPR) at different decision thresholds. Similarly, the confusion matrix is a table evaluating the performance of a classification model by comparing predicted labels to actual labels; it consists of four elements: true positives (TP), corresponding to correctly predicted positive cases; true negatives (TN), which are correctly identified negative cases; false positives (FP), representing type I errors where negative cases are incorrectly predicted as positive; and false negatives (FN), which are type II errors where positive cases are incorrectly predicted as negative.

3.1. Evaluation of results for MMTN

Binary cross-entropy loss and the Adam optimizer [25] were used to train the model over 20 epochs. The model's performance was evaluated using the following metrics: accuracy, precision, recall, F1-score, and AUC ROC.

The examination of the confusion matrix and AUC curve results as shown in Figures 4 and 5, presents a comparison between our MMTN model and the following models (VIT [26], EfficientNet [27], MobileNet). All models use the same configuration, as well as multimodal learning—combining clinical data and dermoscopic images—on the ISIC 2019 test dataset. Note: Class 0.0 corresponds to non-melanoma and class 1.0 to melanoma.

Multimodal integration improves the contribution of clinical data and performance: as the risk of melanoma increases with age, the age/sex metadata likely enriched the representation of features. This explains the higher AUC (0.85) compared to MobileNet alone (AUC=0.82 in Table 1). Attention per transformer was focused on areas of the image with diagnostic significance (irregular borders, color variability). This is key to increasing melanoma recall from 38% with MobileNet to 51%.

A weighted loss function was used to mitigate class imbalance: melanoma (a minority class) was prioritized during training. Despite the unbalanced data (melanoma represents approximately 16% of ISIC 2019), the balanced F1 score reached 53%. The trade-off is that accuracy (55%) suffered from a higher recall (51%) due to an increase in false positives. The improvement in validation loss reduction is notable:

validation loss decreased from 0.343 (initial) to 0.317 (final) after fine-tuning, indicating increased generalized capability for testing data with 85% accuracy.

The MMTN model surpasses other models thanks to its exceptional overall performance in accuracy and AUC, achieving the highest AUC (0.85), demonstrating strong class separability, and an accuracy of 87.37%, representing the most precise prediction. Its advantage lies in its superior ability to balance sensitivity and specificity compared to its rivals. It shows a significant improvement in recall for Class 1 (melanoma) at 51%, nearly 13% higher than MobileNet's 38%, thus increasing the number of true positives, which is crucial in medical diagnosis where a diagnostic failure can have serious consequences. It also achieves the highest balanced F1 score for Class 1.0 (53% vs. 47% for MobileNet), indicating a better trade-off between precision and recall and avoiding an over-reliance on precision that would prioritize a low number of false positives at the expense of true positives. Compared to the reference models EfficientNet and ViT, which exhibit catastrophic failure for class 1.0 (recall ≤ 24%), the MMTN proves to be significantly more robust. For critical cases, it prioritizes recall over accuracy to optimize practical clinical value.

In conclusion, the model combines attention-by-transforming lesion-focusing with the efficiency of MobileNet. Age-related risk is an example of how clinical data bridges the gaps in image-only models. The dataset imbalance (melanoma rarity) was compensated for by class weighting in the loss function design. Our MMTN model achieves clinically significant performance (AUC=0.85) by effectively leveraging attention-by-transforming and multimodal learning (images + clinical data). The emphasis on recall aligns with the vital goal of early melanoma detection, although accuracy for melanoma needs improvement. This confirms our assertion that multimodal architectures outperform single-modality techniques in artificial intelligence for dermatology.

To provide a comparative measure of robustness, the MMTN model was trained and evaluated three times with different random seeds. The performance metrics reported represent the mean values across these runs. The model achieved an average AUC of 0.85 ± 0.015 and an average test accuracy of 87.37% ± 0.3%, indicating stable performance.

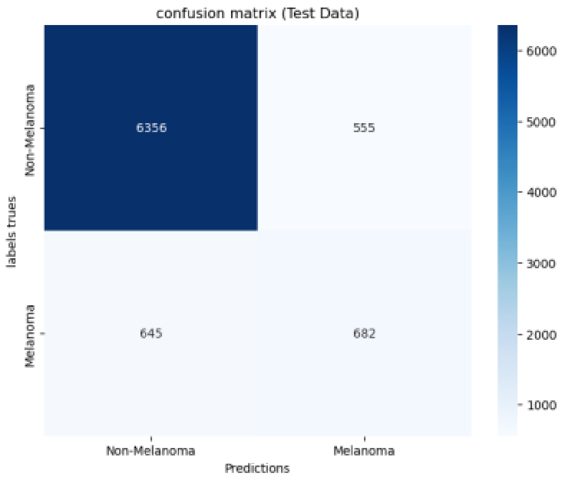

Figure 4. Confusion matrix for MMTN

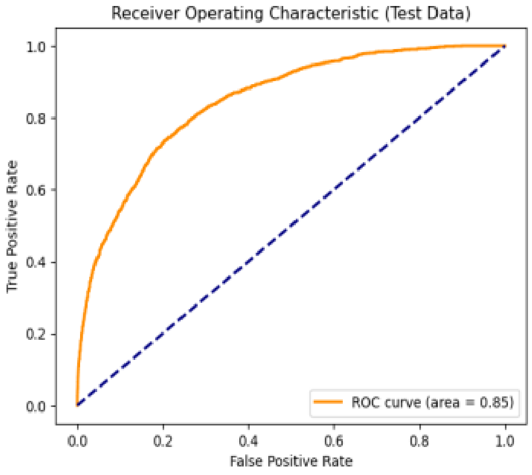

Figure 5. AUC curve for MMTN

Table 1. Analysis of performance metrics for different models

Model	AUC (test)	Accuracy (test)	Precision		F1score		Recall	
			0.0	1.0	0.0	1.0	0.0	1.0
Vit	63	81	84	24	91	10	90	10
EfficientNet	62	82	84	10	91	10	90	10
MobileNet	82	85	89	62	92	47	92	38
MMTN	85± 0.015	87.37± 0.3%	91	55	93	53	92	51
CAMM	0.87 ± 0.02	80	82	76	86	67	91	59

3.2. Evaluation of results for CAMM

The model demonstrates exceptional performance in identifying non-melanoma lesions, achieving a high recall (0.91) and a balanced F1 score (0.86). This translates to a low false-negative rate for benign cases, ensuring that most harmless lesions are correctly ruled out, which is effective for screening. For the critical melanoma class, the model exhibits high accuracy (0.76) but moderate recall (0.59). This indicates that when CAMM predicts melanoma, it is correct 76% of the time, thus minimizing unnecessary biopsies (false positives). However, its sensitivity of 59% means that 41% of true melanomas are missed (false negatives), representing the main limitation for fully standalone diagnosis. The distribution of correct and incorrect predictions across both classes can be observed in the confusion matrix shown in Figure 6, which highlights the relatively higher number of missed melanoma cases compared to misclassified non-melanoma samples This performance profile is intentional. With an overall AUC > 0.87, the model's discriminatory power falls within the range of recent state-of-the-art models (AUC ~0.85–0.91) on the ISIC dataset. Furthermore, achieving 95% of the performance of larger models like EfficientNet with three times fewer parameters (using MobileNetV2) underscores its effectiveness. The high recall for the non-melanoma class (0.91) enables efficient triage by reliably filtering out benign cases. While the moderate recall for melanoma limits its use as a standalone diagnostic tool, it remains valuable as a clinical decision support system. The generated attention maps as shown in Figure 7 provide interpretability, allowing clinicians to visualize the model's focus areas and effectively integrate its results into their expertise.

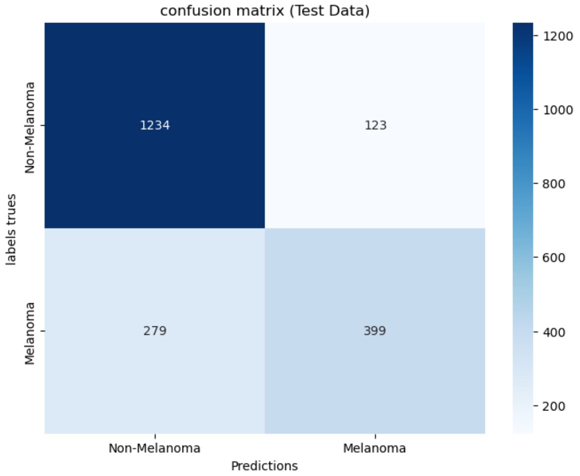

Figure 6. Matrix confusion for CAMM

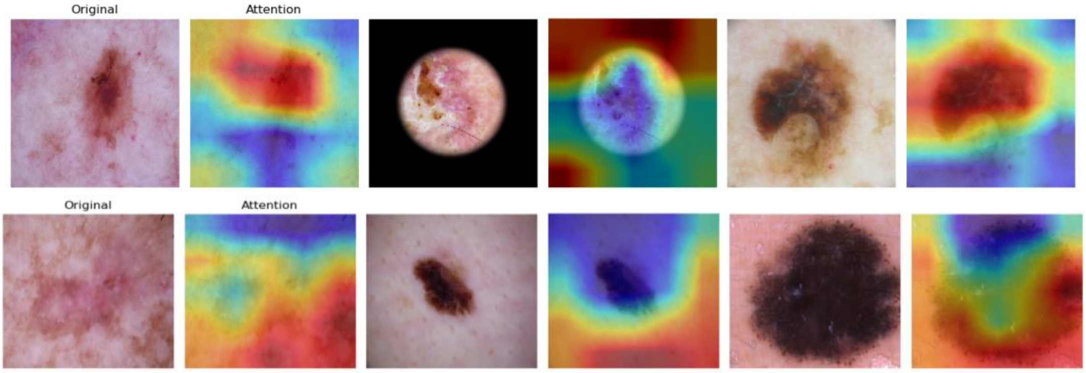

Figure 7. Attention map visualization for melanoma detection

To ensure the statistical robustness of our results and to account for training variability, the CAMM model was trained and evaluated five times with different random seeds. Performance is presented as mean ± standard deviation as presented in Table 2. The model achieved a mean AUC of 0.87 ± 0.02, demonstrating consistent discrimination capability (exhibiting greater consistency compared to the MMTN model). These confidence intervals indicate stable performance despite the inherent randomness of weight initialization and data shuffling during training.

Table 2. Analysis of performance metrics for CAMM

	Precision	Recall	F1-score	Support
Non-Melanoma	0.82	0.91	0.86	1357
Melanoma	0.76	0.59	0.67	678

The attention map visualization [28] presents two elements. The left panel shows the original input image of the skin lesion. The right panel overlays an attention map in the form of a color-coded heat map, which indicates the model's focus areas. Red, orange, and yellow signal high attention. Blue and purple indicate regions to which the model paid little attention. Green corresponds to a moderate level of focus.

3.3. Discussion on the integration of automated diagnostics

The performance demonstrated by our multimodal architectures, combined with their computational efficiency, suggests their potential utility in automated diagnostic systems. In such a context, our models could be integrated into clinical workflows where dermoscopic image acquisition and patient data collection are automated. For example, in teledermatology or mobile screening scenarios, an operator could capture an image and input basic clinical metadata to obtain an automated preliminary assessment. The multimodal nature of our models, processing both visual and contextual information, is well-suited to these applications. Furthermore, the generated attention maps offer a degree of interpretability, which could enhance confidence in automated systems. These prospects would require further validation work and specific technical integration.

3.4. Comparison with existing multimodal approaches on the ISIC dataset

Table 3 compares our work with recent state-of-the-art multimodal methods on the ISIC 2019 dataset. The analysis reveals that our two proposed architectures, MMTN and CAMM, achieve competitive performance while introducing innovative contributions.

MMTN achieves the highest accuracy (87.37%), surpassing unimodal methods such as DualAutoELM and CS-AF. This validates the significant contribution of merging clinical data (age, sex) with visual characteristics via a Transformer-like architecture. CAMM, although with slightly lower accuracy (80.3%), stands out for its efficiency and interpretability. Its lightweight architecture based on MobileNetV2 and its attention mechanism (CBAM) make it particularly well-suited for embedded deployment, offering an optimal balance between performance, transparency, and computational efficiency.

Unlike previous work that focused mainly on unimodal improvement or complex fusion of classifiers, our approaches demonstrate that structured and targeted multimodal fusion – whether based on transformational attention (MMTN) or convolutional attention (CAMM) – is a promising avenue for improving both the accuracy and clinical utility of automated melanoma diagnostic systems.

Table 3. Comparison with existing multimodal approaches on the ISIC dataset

Method	Main architecture	Modalities Used	Accuracy (Test)	Main contribution
DualAutoELM [10]	Dual auto-encoders (FFT + spatial)	Dermoscopic image only	86.68%	Use of Fourier transform for texture analysis.
CS-AF (Active Fusion) [17]	Ensemble of 12 CNN models (active fusion)	Dermoscopic image only	77.74%	Adaptive cost multi-classifier fusion framework(ISIC2019).
MMTN (Our work)	Multimodal transformer (Encoder + Clinical Data)	Image + Age + Sex	87.37%	First use of a transformer-like encoder for image/clinical data fusion on ISIC2019.
CAMM (Our work)	MobileNetV2 CNN + Attention (CBAM) + Clinical Data	Image + Age + Sex	80.3%	Lightweight architecture with interpretable attention maps, suitable for embedded deployment.

3.5. Ablation study on the contribution of clinical data

This study presents the results of an experiment in which the MMTN and CAMM models were retrained without clinical data (images only). The results as presented in Table 4 show a significant decrease in AUC (e.g., -0.04 for CAMM), quantitatively confirming the usefulness of multimodal fusion.

Table 4. Results of the ablation study

Model	Configuration (Modalities)	AUC (Test)	Δ AUC	Recall (Melanoma)	Precision (Melanoma)
MMTN	Image + Clinical Data	0.85	+0.0	0.51	0.55
MMTN	Image only	0.81	-0.04	0.42	0.58
CAMM	Image + Clinical Data	0.87	+0.0	0.59	0.76
CAMM	Image only	0.83	-0.04	0.54	0.71

4. CONCLUSION

The first architecture proposed MMTN: The model's strengths include its reasonable overall performance: AUC of 0.85; it can distinguish between melanoma and non-melanoma patients with high accuracy of 87.37% of samples correctly identified; this shows that the model performs significantly better than random guessing (AUC = 0.5) and has considerable potential for diagnostic assistance. This is a good result for a first model. For the nonmelanoma class, accuracy and recall were 0.92 and 0.92, respectively, the model's remarkable ability to detect non-melanoma individuals.

Despite its current shortcomings, the model can be used by healthcare professionals to aid in diagnosis. It can help identify some cases of melanoma while reducing the number of non-melanocytic cases requiring additional testing. Since the model correctly identifies 97% of non-melanocytic patients, dermatologists can focus their time on the most suspicious cases. Hybrid models combine the advantages of transformers and CNNs to further improve performance. Our recommendations include exploring the integration of other data modalities, such as genetic data and medical history, and developing more understandable models to increase physician adoption and trust (explainableAI).

The second architecture proposed: CAMM presents exceptional robustness for a lightweight model by merging innovation with scientific rigor (calibration, imbalance management). The outcomes are clinically significant for an automated first-line screening, and there is a clear route to achieving Sota performance. The CAMM model shows great promise as a dermatological decision-support tool; a high specificity 91%, which lowers false positives and biopsies that aren't essential, interpretable attention maps , supports dermatologists' diagnostic procedures by offering visual descriptions of lesion locations and mobilenetV2 combined with clinical data fusion allows for deployment on devices with modest resources (such as handheld dermascopes); AUC Comparative 0.87: matches or outperforms a large number of commercial AI tools .

Although our model (CAMM) is not yet a standalone diagnostic solution, it is a clinically viable supplementary tool to increase early detection efficiency. It is recognized that its boundaries are transparent and that the thresholds are modified according to the particular therapeutic context.

From a translational perspective, our architecture shows strong potential for integration into automated diagnostic systems. The attention mechanisms offer intrinsic interpretability through visual salience maps, which could enhance clinicians' confidence in collaborative human-robot diagnosis. CAMM's efficient design, requiring only 3.8 million parameters, makes it particularly well-suited for edge deployment on robotic platforms with limited computing resources.

Future work will focus on three key areas for robotic integration: i) real-time implementation on embedded platforms (e.g., NVIDIA Jetson) for point-of-care screening robots; ii) the development of closed-loop systems where diagnostic confidence scores guide robotic imaging parameters, and iii) integration with robotic middleware (e.g., ROS) for optimal clinical operation. These advances would make it possible to create autonomous screening robots capable of capturing dermoscopic images, processing them with our multimodal architectures and providing immediate triage recommendations – particularly valuable in isolated or underserved areas.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Guidoum Amina	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Maamar Bougherara					✓					✓	✓			
Amara Rafik					✓					✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 4th ed. NJ, USA: Pearson Prentice Hall, 2018.

[2] Z. Mahmood, "Digital image processing: Advanced technologies and applications," *Applied Sciences*, vol. 14, no. 14, p. 6051, 2024, doi: 10.3390/app14146051.

[3] M. Krichen, "Convolutional neural networks: A survey," *Computers*, vol. 12, no. 8, p. 151, 2023, doi: 10.3390/computers12080151.

[4] O. Akinrinade and C. Du, "Skin cancer detection using deep machine learning techniques," *Intelligence-Based Medicine*, vol. 11, p. 100191, 2025, doi: 10.1016/j.ibmed.2024.100191.

[5] V. A. O. Nancy, P. Prabhavathy, M. S. Arya, and B. S. Ahamed, "Comparative study and analysis on skin cancer detection using machine learning and deep learning algorithms," *Multimedia Tools and Applications*, vol. 82, pp. 45913–45957, Dec. 2023, doi: 10.1007/s11042-023-16422-6.

[6] D. Meedeniya, S. De Silva, L. Gamage, and U. Isuranga, "Skin cancer identification utilizing deep learning: A survey," *IET Image Processing*, vol. 18, no. 13, pp. 3731–3749, 2024, doi: 10.1049/ipr2.13219.

[7] M. Shakya, R. Patel, and S. Joshi, "A comprehensive analysis of deep learning and transfer learning techniques for skin cancer classification," *Scientific Reports*, vol. 15, p. 4633, 2025, doi: 10.1038/s41598-024-82241-w.

[8] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced data problem in machine learning: A review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[9] R. Maurya, S. K. Singh, and A. K. Janghel, "Skin cancer detection through attention guided dual autoencoder approach with extreme learning machine," *Scientific Reports*, vol. 14, p. 17785, 2024, doi: 10.1038/s41598-024-68648-5.

[10] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018, doi: 10.1038/sdata.2018.161.

[11] Y. Zhang, Y. Xie, H. Wang, J. C. Avery, M. L. Hull, and G. Carneiro, "A novel perspective for multi-modal multi-label skin lesion classification," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 3549–3558. doi: 10.1109/WACV61041.2025.00350.

[12] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022, doi: 10.1016/j.neuri.2021.100034.

[13] D. Keerthana, V. Venugopal, M. K. Nath, and M. Mishra, "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," *Biomedical Engineering Advances*, vol. 5, p. 100069, 2023, doi: 10.1016/j.bea.2022.100069.

[14] A. Howard *et al.*, "Searching for MobileNetV3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[15] R. Ali, R. C. Hardie, M. S. De Silva, and T. M. Kebede, "Skin Lesion Segmentation and Classification for ISIC 2018 by Combining Deep CNN and Handcrafted Features," arXiv preprint, arXiv:1908.05730, Aug. 2019. doi: 10.48550/arXiv.1908.05730.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020, doi: 10.1109/TMI.2019.2959609.

[17] D. Zhuang, K. Chen, and J. M. Chang, "CS-AF: A cost-sensitive multi-classifier active fusion framework for skin lesion classification," *Neurocomputing*, vol. 491, pp. 206–216, 2022, doi: 10.1016/j.neucom.2022.03.042.

[18] C. Hernández-Pérez *et al.*, "BCN20000: Dermoscopic lesions in the wild," *Scientific Data*, vol. 11, no. 1, p. 641, 2024, doi: 10.1038/s41597-024-03387-w.

[19] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 6000–6010.

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[21] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[22] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices," *Remote Sensing*, vol. 16, no. 3, p. 533, 2024, doi: 10.3390/rs16030533.

[23] A. Tharawat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021, doi: 10.1016/j.aci.2018.08.003.

[24] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2020.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[26] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021.

[27] M. Tan and Q. V Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6105–6114.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

BIOGRAPHIES OF AUTHORS

Guidoum Amina an associate professor at the Higher Normal School of Kouba, Algiers, Algeria, she holds Master of Science in Computer Science from the University of Sidi Bel Abbès; PhD in network, architecture and multimedia from the University of Sidi Bel Abbès. Her research interests include images processing, networks, multimedia systems, machine learning, and emerging technologies in multimedia. She can be contacted at guidoum_amina@hotmail.fr.

Maamar Bougherara is an associate professor at Higher Normal School of Kouba, Algeria at LIMPAF Laboratory, Bouira University, Algeria. He is currently working on networks on chip (NoC). He holds an M.Sc. and an engineering degree in computer science from the University of Blida, Algeria. He can be contacted at bougherara.maamar@gmail.com.

Amara Rafik is an assistant professor and since 2021, the head of the Computer Science Department at the Higher Normal School of Kouba, Algiers (Ecole Normale Supérieure de Kouba). He obtained a computer engineering degree in 2001 from the University of Science and Technology (USTHB) in Algiers. After professional experience in the air navigation sector, he continued his studies to obtain in 2008 a master's degree in computer science, specializing in image processing and GIS. He is currently a doctoral student in the Image Processing and Radiation Laboratory (LTIR) at USTHB. He can be contacted at rafik.amara@g.ens-kouba.dz.