Natural Language Processing Journal 16 (2026) 100232

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.sciencedirect.com/journal/natural-language-processing-journal

BNDrop-XNet: A lightweight multi-modal deep learning framework with weighted fusion for melanoma detection using dermoscopic and clinical data

Samad Najjar-Ghabel *, Shamim Yousefi

Department of Computer Engineering, University of Mohaghegh Ardabili, Ardabil, Iran

ARTICLE INFO

Keywords:
Clinical metadata fusion
Deep learning
Dermoscopic imaging
Ensemble learning
Melanoma detection

ABSTRACT

Melanoma is one of the most life-threatening forms of skin cancer, where early detection improves survival outcomes. Despite recent advances in Deep Learning (DL) for dermoscopic image analysis, existing approaches face critical limitations, including reliance on image-only inputs, inability to incorporate clinical metadata, and unstable performance on imbalanced datasets observed in real-world screening. Furthermore, many architectures are computationally heavy, limiting their suitability for clinical deployment. To address these gaps, in this paper, we propose a lightweight multi-modal diagnostic framework (BNDrop-Xnet) that integrates dermoscopic image features with clinical metadata (age, sex, and lesion location) through a weighted fusion strategy. Our framework consists of four main stages: (1) data preprocessing, including normalization, stratified sampling, and SMOTE-based class balancing; (2) image model creation using a four-block Convolutional Neural Network (CNN) architecture with Conv2D, Batch Normalization, MaxPooling, and Dropout layers; (3) metadata model construction with impact-factor computation and Multi-Layer Perceptron (MLP)-based learning of diagnostic feature contributions; and (4) ensemble integration through a weighted fusion and decision threshold optimization. It was evaluated on the ISIC 2020 dataset and achieved 90% accuracy, 80% precision, and a 73% F1-Score, outperforming several state-of-the-art models. These results demonstrate the effectiveness of combining visual and clinical cues for robust melanoma screening. BNDrop-XNet offers an efficient, interpretable, and meaningful solution for real-world deployment in resource-limited environments.

1. Introduction

Melanoma is the most life-threatening type of skin cancer that causes almost 75% of all skin cancer deaths worldwide, even though it represents less than 5% of total skin cancer cases (Virgens et al., 2025, Nasir et al., 2025). According to the World Health Organization (WHO) and the U.S. Centers for Disease Control and Prevention, over 325000 new cases are detected each year, and the number continues to rise among fair-skinned people in areas with strong sunlight, such as Australia, Europe, and North America (Ferreirinha et al., 2025, Javed et al., 2020). However, early detection of melanoma is the most critical determinant of survival. Indeed, when it is found at Stage I, the five-year survival rate is above 95%, but drops below 20% once the disease spreads. Despite this, early diagnosis remains difficult, even for skilled dermatologists, because benign and malignant skin lesions share similar visual characteristics under dermoscopy (Kakish et al., 2025, Yousefi et al., 2024). Variations in lighting, imaging devices, and lesion morphology complicate the interpretation process (Oniga et al., 2025, Karimi et al., 2025). These challenges have increased the need for Computer-Aided Diagnostic (CAD) systems using Artificial Intelligence (AI) and Deep Learning (DL) to support doctors in detecting subtle malignant features (Ahmad et al., 2025, Kong et al., 2025). Therefore, developing robust and reliable AI models for melanoma detection is a critical health goal rather than a computational vision challenge.

Over the last decade, advances in DL have improved the automation of the melanoma detection workflow. Convolutional Neural Networks (CNNs) such as ResNet, DenseNet, and EfficientNet demonstrated superior performance in learning complex visual patterns from dermoscopic images compared to traditional Machine Learning (ML) methods (Ye et al., 2024, Yousefi et al., 2025, Montoya et al., 2025). Using large-scale public datasets like ISIC 2018, ISIC 2019, and ISIC 2020, many models achieved area under the ROC curve (AUC) scores ranging

* Corresponding author.
E-mail addresses: S.Najjar@uma.ac.ir (S. Najjar-Ghabel), Sh.yousefi@uma.ac.ir (S. Yousefi).

https://doi.org/10.1016/j.nlp.2026.100232
Received 21 May 2026; Received in revised form 17 August 2026; Accepted 2 September 2026
Available online 6 September 2026
2949-7191/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

between 0.88 and 0.92, surpassing expert-level detection accuracy (Yousefi et al., 2024, Zhang and Chaudhary, 2025, Yousaf et al., 2019). The models extract hierarchical features such as color variation, border irregularity, and asymmetry for differentiating malignant from benign lesions. More recent studies extended this progress through multi-modal learning, where dermoscopic images are combined with clinical metadata such as age, gender, and lesion location to provide richer contextual understanding (Virgens et al., 2025, Yousefi et al., 2024, Najjar-Ghabel et al., 2025). Overall, DL has established a strong foundation for melanoma detection and opened the door to new models that are more adaptive, lightweight, and reliable.

Recent studies on melanoma detection achieved remarkable progress using DL architectures such as CNNs, U-Nets, and hybrid segmentation–classification frameworks, trained on large public datasets like ISIC 2019–2020 and HAM10000 (Yousefi et al., 2024, Najjar-Ghabel et al., 2025, Khan et al., 2025, Kaur et al., 2025). Although these approaches improved lesion localization and classification accuracy, most rely on dermoscopic images, overlooking valuable clinical metadata such as age, gender, and lesion location. Moreover, the majority of recent models employ static fusion mechanisms or computationally demanding multi-stage pipelines, making them unsuitable for low-resource clinical settings. They also lack a predefined feature integration and robust class imbalance handling, which are essential for reliable medical applications (Cai et al., 2026). These limitations underscore the need for an adaptive multi-modal framework, which fuses dermoscopic and clinical data at both feature and decision levels to achieve higher detection accuracy, robustness, and clinical applicability.

In this paper, we propose a multi-step framework named BNDrop-XNet to enhance melanoma detection by integrating dermoscopic images with clinical metadata. Our model follows a structured workflow that begins with data preprocessing, including image normalization, stratified sampling, and class-imbalance correction. The core of the framework is a lightweight CNN composed of four sequential blocks that extract hierarchical lesion features while preventing overfitting. In parallel, clinical metadata (age, sex, and lesion location) are processed through a metadata model that quantifies their diagnostic influence using impact-factor analysis. The outputs of both models are then fused through an adaptive weighted ensemble, followed by threshold optimization to refine diagnostic decisions. Together, these steps create an end-to-end, interpretable, and efficient system for producing reliable melanoma classification even in heterogeneous clinical settings. The main contributions of this paper are as follows:

- Introduction of BNDrop-XNet, a lightweight four-block CNN designed for medical imaging tasks with limited computational resources.
- Predefined global weighted fusion of image and metadata representations to enable integration of visual and clinical cues for accurate diagnosis.
- Metadata impact-factor analysis, which quantifies the diagnostic contribution of age, sex, and lesion location.
- Optimized decision thresholding that maximizes sensitivity–specificity balance for enhanced real-world clinical applicability.

Experimental results demonstrate that the proposed BNDrop-XNet framework delivers strong and balanced diagnostic performance on the ISIC 2020 dataset. The model achieved 90% accuracy, 80% precision, and a 73% F1-Score, outperforming several recent state-of-the-art approaches. These improvements highlight the effectiveness of combining dermoscopic image features with clinical metadata through our weighted fusion strategy. Threshold optimization further enhanced the trade-off between sensitivity and specificity, enabling stable decision-making across diverse lesion types. Overall, the results demonstrate that BNDrop-XNet offers a reliable, efficient, and clinically meaningful solution for supporting real-world melanoma screening in settings where diagnostic accuracy and model interpretability are critical.

The rest of the paper is organized as follows. Section 2 reviews recent related works on melanoma detection using DL and multi-modal fusion approaches. The details of the proposed framework, including the preprocessing steps, network architecture, and weighted fusion strategy, are presented in Section 3. Section 4 reports the experimental setup, evaluation metrics, and comparative results on the ISIC 2020 dataset. Section 5 discusses the results and clinical applicability of our framework. Finally, the paper is concluded in Section 6.

2. Related works

Recent advances in ML have improved the accuracy of the melanoma detection process from dermoscopic images. It has led to the development of numerous frameworks integrating segmentation, classification, and data augmentation techniques. In this context, our team (Yousefi et al., 2024) developed a hybrid melanoma detection framework that combines Discrete Cosine Transform (DCT)–based features from dermoscopic images with clinical metadata for classification. It extracts DCT coefficients to capture texture and frequency details, then fuses them with metadata to improve diagnostic accuracy compared to image-only baselines. The study utilized the ISIC 2020 dataset and implemented traditional ML classifiers, such as Support Vector Machine (SVM) and Random Forest (RF), trained on DCT-extracted image features combined with clinical metadata for melanoma classification. Its main advantage is the use of multi-modal information, enhancing robustness and interpretability. However, the model depends on handcrafted DCT features with static fusion between modalities, lacks adaptive deep feature learning and dual-level fusion, and shows limited scalability.

Zhang and Chaudhary (2025) introduced a hybrid DL approach that combines a U-Net segmentation network with an EfficientNet classifier to enhance melanoma detection based on the HAM10000 and ISIC 2020 datasets. The model achieved a high accuracy on ISIC 2020. It demonstrates strong performance by segmenting lesion regions before classification to minimize background interference. It improves lesion localization and classification precision through sequential learning. Despite its high accuracy, the model relies on dermoscopic images, lacks integration of clinical metadata, and requires substantial computational resources. Kaur et al. (2025) presented a DL-based CAD system for automated melanoma detection using the ISIC 2020 dataset. The pipeline includes several stages, including image preprocessing, lesion segmentation using a context-aggregation convolutional network, and classification. It achieved reasonable accuracy and improved lesion localization by removing background noise. Its main advantage is enhancing the visual quality of dermoscopic inputs before classification. However, the model relies on dermoscopic images without integrating clinical metadata, lacks a predefined global fusion, and introduces higher computational complexity that limits applicability in low-resource clinical environments.

Marie et al. (2025) proposed a hybrid detection framework that combines the high-speed object detection model YOLOv9 with the region-precise classifier Faster R-CNN to detect melanoma in dermoscopic images. Their algorithm demonstrates solid performance on benchmark datasets (e.g., ISIC 2019/2020) by using the strengths of both detection and classification networks to improve localisation. However, it focuses on dermoscopic images and lacks integration of clinical metadata. Moreover, its two-stage structure is computationally heavy and lacks predefined global fusion. Nazari and Garcia (2025) designed an attention-based Convolutional Neural Network (CNN) for automatic melanoma detection from dermoscopic images. The model uses an attention mechanism to highlight important lesion areas while suppressing background noise, improving classification accuracy and reducing false positives. Trained on a large dermoscopy dataset, it achieves better sensitivity and performance than baseline CNN models. Although it increases interpretability, the method relies on image data,

lacking clinical metadata integration and introducing high computational cost, which limits its robustness and real-world deployability.

Thaku et al. (2025) planned a CAD system that combines a Double U-Net for lesion segmentation with several deep classifiers, including EfficientNetB7, DenseNet201, and ResNet152V2, for melanoma detection. The method, tested on HAM10000 and ISIC datasets, achieved high accuracy and showed strong performance across multiple architectures. The main advantage is its high precision and effective lesion localization through segmentation and classification. However, it depends on image data and uses large networks, making the model resource-intensive and less suitable for low-resource deployment. Alzamel et al. (2025) developed four CNN models (ResNet v2, VGG16, EfficientNet-B5, and EfficientNet-B7) for melanoma detection using the ISIC 2020 archive. It also emphasized the importance of handling class imbalance by applying Generative Adversarial Networks (GANs) for data augmentation, which improved F1-Scores. The main advantage of this work is comparing multiple CNN backbones and showing that image-only pipelines with augmentation can achieve competitive performance. However, the approach remains confined to image data, does not integrate clinical metadata. It also lacks the fusion of clinical metadata with image features. Moreover, the model uses conventional CNN backbones and augmentation, but does not implement a robust multi-level fusion strategy.

Khan et al. (2025) introduced a DL model for early-stage melanoma detection that addressed dataset imbalance using a CNN enhanced with DenseNet blocks, batch normalization, and data augmentation. The model achieved a high accuracy on the ISIC-2019 and HAM10000 datasets, showing strong results in imbalanced conditions. However, it relies on image data without clinical metadata fusion, lacks dual-level adaptive fusion, and remains computationally intensive. Our team, in another paper (2025), presented a multi-modal DL framework for melanoma detection that integrates dermoscopic images and clinical metadata. The model combines EfficientNet for extracting complex visual patterns from images and a Multi-Layer Perceptron (MLP) for processing structured clinical data. These two modalities are fused using an attention-based integration layer that prioritizes relevant features. It achieves high accuracy on the ISIC 2020 dataset. The key advantage of this method is its balanced performance and ability to use complementary information from both visual and contextual sources, reducing false positives and negatives. Yet, the model uses a static fusion mechanism without adaptive weighting or decision-level integration and does not address class imbalance and lightweight deployment.

A summary of the methodologies, datasets, advantages, and limitations of the reviewed papers is presented in Table 1. It shows that most existing melanoma detection frameworks rely on image-based DL models, such as CNNs, U-Nets, and hybrid segmentation–classification pipelines, which have achieved high accuracies on datasets like ISIC 2019–2020 and HAM10000. While the methods have advanced lesion localization and classification, they lack integration of clinical metadata, key contextual information such as age, gender, and lesion location, which has been demonstrated to enhance detection reliability. Furthermore, most frameworks depend on static fusion mechanisms, handcrafted features, or computationally intensive multi-stage architectures, making them less suitable for real-world low-resource clinical deployment. Only a few models address class imbalance or provide interpretable outputs, limiting their clinical trustworthiness. These gaps highlight the need for a lightweight and adaptive multi-modal framework, which fuses dermoscopic and clinical data at both feature and decision levels, balances data distribution, and ensures high accuracy with improved practical deployability in healthcare environments.

3. BNDrop-Xnet framework

In this section, we propose a comprehensive multi-stage model for automated melanoma detection that integrates both dermoscopic images and clinical metadata to improve diagnostic reliability. The overall workflow of the proposed framework, BNDrop-XNet, is illustrated in Fig. 1. The process begins with data collection and sampling, followed

Table 1
A summary of the reviewed papers.

Authors & Year	Objective	Methods	Dataset(s)	Advantages	Gaps
(Yousefi et al., 2024)	Develop a hybrid framework combining image features with clinical metadata	DCT feature extraction + SVM, RF	ISIC 2020	Multi-modal design using metadata; improved interpretability and accuracy	Uses handcrafted features; static fusion; no deep feature learning; limited scalability
(Zhang and Chaudhary, 2025)	Improve detection via lesion segmentation and classification	Double-stage DL: U-Net + EfficientNet	HAM10000, ISIC 2020	High accuracy; strong lesion localization and background noise reduction	Image-only model; lacks metadata integration; high computational cost; no adaptive fusion
(Kaur et al., 2025)	Build a CAD system with preprocessing, segmentation, and classification	Context-aggregation CNN + classifier pipeline	ISIC 2020	Improved lesion localization and visual quality; enhanced image clarity	Image-only; lacks metadata fusion; computationally demanding for real-time deployment
(Marie et al., 2025)	Combine object detection and classification for accurate lesion localization	YOLOv9 + Faster R-CNN hybrid	ISIC 2019, ISIC 2020	Precise detection and refined classification; improved lesion region focus	Image-only; no metadata use; computationally demanding; lacks a fusion mechanism
(Nazari and Garcia, 2025)	Enhance melanoma detection using attention-based CNNs	Attention-based CNN	Large dermoscopy dataset	Improves model focus on salient lesion areas; interpretable attention maps; high sensitivity	Image-only; no metadata fusion; high computational cost; lacks decision-level fusion
(Thakur et al., 2025)	Combine segmentation and classification to improve precision and localization	Double U-Net + EfficientNetB7, DenseNet201, ResNet152V2	HAM10000, ISIC	High precision and effective lesion localization; consistent accuracy	Image-only; lacks metadata; resource-intensive; no dual-level fusion
(Alzamel et al., 2025)	Compare CNN backbones and address class imbalance	ResNet v2, VGG16, EfficientNet-B5/B7 + GAN-based augmentation	ISIC 2020	Addresses class imbalance; improved F1-Score; model comparison	Image-only; lacks metadata fusion; no adaptive fusion; conventional CNN design
(Khan et al., 2025)	Detect early-stage melanoma and handle class imbalance	CNN + DenseNet blocks + augmentation	ISIC-2019, HAM10000	High accuracy in imbalanced data; robust CNN	Image-only; no metadata fusion; lacks dual-level fusion; computationally intensive
(Najjar-Ghabel et al., 2025)	Combine image and metadata for melanoma detection	EfficientNet + MLP + attention-based fusion	ISIC 2020	Balanced multi-modal learning; reduced false positives; improved contextual understanding	Fusion is static; lacks adaptive weighting and decision-level fusion; no imbalance handling

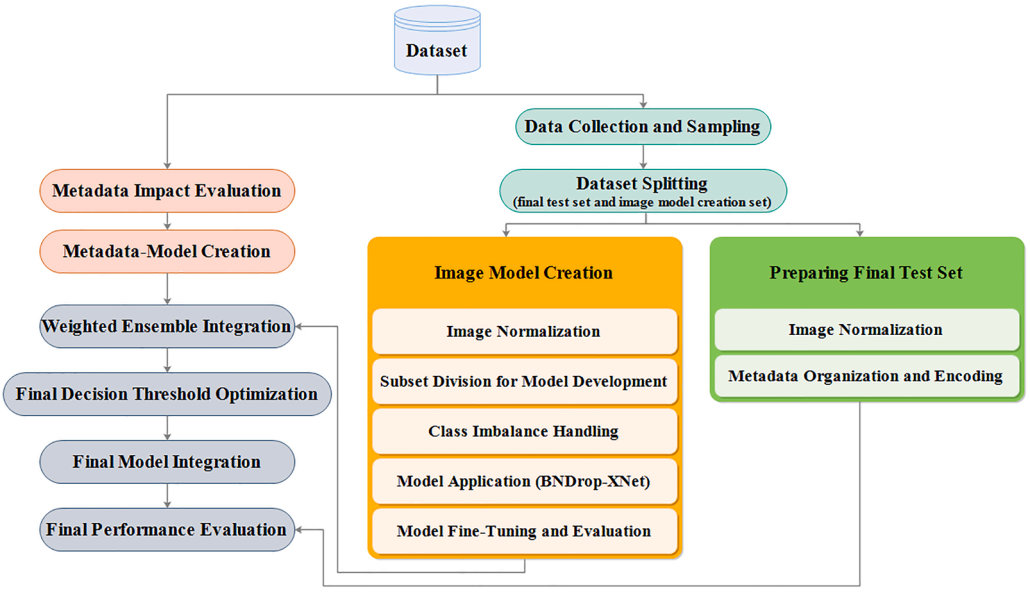


Fig. 1. Overall workflow of the proposed BNDrop-XNet framework for melanoma detection.

by dataset splitting into an image model creation set and a final test set to ensure unbiased evaluation. The image model creation stage involves preprocessing steps such as image normalization, data partitioning, and class imbalance handling before training the DL model. In parallel, the final test set is prepared through normalization and metadata encoding to provide a consistent evaluation environment. These stages ensure standardized, balanced, and high-quality data preparation for subsequent learning and validation.

The core of our proposed framework is the BNDrop-XNet model, a lightweight convolutional neural network composed of four sequential convolutional blocks. This configuration enables hierarchical feature extraction while maintaining model stability and preventing overfitting. After completing image feature extraction, metadata impact evaluation is performed to analyze the detection relevance of patient data such as age, sex, and lesion location. The metadata-model creation stage then develops a model based on clinical metadata, generating predictive outputs that reflect non-visual diagnostic cues. These image- and metadata-based outputs are then combined through a weighted ensemble integration strategy, enabling adaptive fusion of visual and contextual features. Finally, the framework performs decision threshold optimization, final model integration, and comprehensive performance evaluation, resulting in an end-to-end interpretable system that enhances diagnostic accuracy, robustness, and clinical applicability by merging image-based and patient-specific data.

The symbols used in our model are summarized in Table 2. These parameters define the variables employed during model development, training, and integration of image- and metadata-based components. They include the dataset split parameters n_{train}, $n_{validation}$, and n_{test}, as well as training-related parameters such as batch size B and the maximum number of epochs E_{max}. It also defines the metadata-analysis variables, including the correlation coefficient r_j, information gain IG_j, normalized correlation contribution C_j, normalized information-gain contribution G_j, and the resulting metadata impact factor IF_j for age, sex, and lesion location. In addition, the predefined global fusion weights $a_{M_{Im}}$, a_{age}, a_{sex}, and a_{site}, with the final diagnostic score *FinalS* and the optimized decision threshold W_{Best}, are defined to describe the weighted multimodal decision-fusion process.

Table 2
Description of symbols used in the formulation of our model.

Symbol	Description
n_{train}	Number of training samples
n_{test}	Number of test samples
$n_{validation}$	Number of validation samples
KS	kernel_size of Conv1D
B	Batch size
E_{Max}	Maximum number of epochs
M_{IM}	Image Model (BNDrop-XNet)
$a_{M_{Im}}$	Coefficient of Image Model
a_{age}	Coefficient of age
a_{sex}	Coefficient of sex
a_{site}	Coefficient of lesion location
$FinalS$	Final Score
$FinalD$	Final Decision
r_j	Correlation coefficient between metadata variable x_j and melanoma target y
IG_j	Information gain of metadata variable x_j with respect to melanoma target y
C_j	Normalized correlation contribution of metadata variable x_j
G_j	Normalized information-gain contribution of metadata variable x_j
IF_j	Final standardized impact factor of metadata variable x_j
IF_{age}	Impact factor of age
IF_{sex}	Impact factor of sex
IF_{site}	Impact factor of the lesion location
λ	Parameter controlling the relative contribution of correlation and information gain
W	Total threshold
W_{Best}	Best Threshold

3.1. Data collection and sampling

The proposed framework exploits the ISIC 2020 dataset, which contains dermoscopic images labelled as melanoma and non-melanoma, along with corresponding clinical metadata including age, sex, and lesion location (Cassidy et al., 2022). Let n_{train}, n_{test}, and $n_{validation}$ represent the number of samples in the training, test, and validation subsets, respectively. The dataset exhibits a strong class imbalance, where the number of non-melanoma samples exceeds that of melanoma cases. To ensure balanced model learning, a representative subset of non-melanoma samples is selected to maintain a proportional distribution between the two classes. The sampling step prevents the model from biasing toward the majority class and enhances its sensitivity to melanoma features. Only high-quality images with complete metadata are included, and samples with missing clinical information or poor image quality are excluded to guarantee data consistency throughout the

training and evaluation process.

3.2. Dataset splitting

In this step, the dataset samples are divided into two main sets: (1) an image model creation set used for training, validation, and fine-tuning, and (2) a final test set reserved for unbiased performance evaluation. This separation ensures that the final assessment of the BNDrop-XNet model is performed on unseen data, preventing information leakage and overfitting. It enhances the generalization capability of our framework in real-world melanoma detection.

3.3. Image model creation

Before model training, all dermoscopic images are processed through pixel value normalization to standardize the intensity range across samples. Each image is scaled to a range of [0, 1], ensuring consistent contrast and brightness while preserving lesion characteristics essential for feature learning. It minimizes illumination variance and device-related inconsistencies and improves the convergence stability of the model. This preprocessing phase is crucial in medical imaging tasks, where heterogeneous acquisition conditions can lead to biased feature extraction and low diagnostic accuracy.

The normalized dataset is then divided into training, validation, and testing subsets. Stratified sampling is applied to maintain a consistent class distribution across subsets, ensuring fair evaluation and preventing overfitting to the majority class. To further mitigate the imbalance between melanoma and non-melanoma cases, the Synthetic Minority Oversampling Technique (SMOTE) is used on the training set (Najjar-Ghabel et al., 2024, Idwan et al., 2025). SMOTE generates synthetic minority samples by interpolating between neighboring melanoma examples, which not only enhances the representation of rare lesions but also prevents overfitting caused by simple data duplication. It improves the sensitivity of the framework to malignant features, promoting a more balanced decision boundary.

The core of our proposed framework is the BNDrop-XNet architecture, illustrated in Fig. 2, which consists of four sequential convolutional blocks, each containing a Conv2D, Batch Normalization, MaxPooling, and Dropout layer. These components work together: Conv2D extracts hierarchical lesion patterns, Batch Normalization stabilizes training and accelerates convergence, MaxPooling reduces spatial redundancy, and Dropout mitigates overfitting by randomly deactivating neurons. The fourth block flattens the extracted features and connects them to dense layers with ReLU and sigmoid activations for nonlinear mapping and binary classification. The number of filters in each convolutional block (32, 64, and 128) is increased to enable hierarchical feature extraction. This allows the model to capture both low-level textures and high-level lesion structures as depth increases. The Dropout rates of 0.25 and 0.4 are chosen to balance model regularization and learning capacity; lower rates in early layers preserve essential features, while higher rates in deeper layers prevent overfitting during complex pattern learning.

After creating the architecture of the image model, its parameters, including learning rate, batch size, and dropout ratio, are fine-tuned using the validation set to minimize validation loss. The final model is then evaluated on the independent test set, demonstrating that the lightweight BNDrop-XNet balances computational efficiency, robustness, and detection accuracy. It makes the model suitable for resource-constrained applications.

3.4. Metadata model creation

The metadata model creation step focuses on quantifying the diagnostic relevance of clinical information, including age, sex, and lesion location, and incorporating these non-visual cues into the final prediction. The metadata are first encoded and normalized to ensure a consistent numerical representation and compatibility with the model-

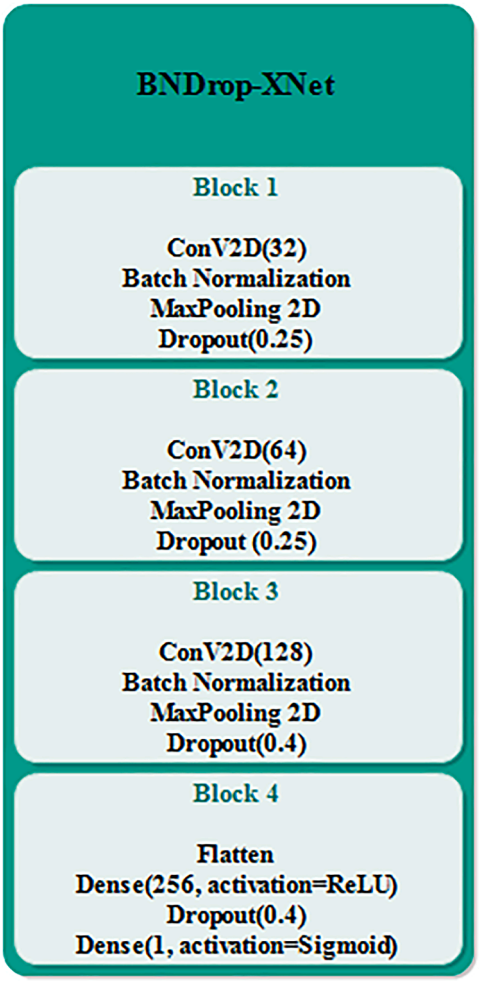


Fig. 2. Architecture of the BNDrop-XNet model.

ling pipeline (Yousefi et al., 2024). For each metadata variable x_j, where j ∈ {age, sex, site}, an impact factor (IF_j) is calculated by combining correlation- and information-theoretic measures. The correlation component quantifies the strength of association between each metadata variable and the binary melanoma target, whereas information gain quantifies the reduction in target uncertainty contributed by that variable. To ensure comparability, the absolute correlation and information-gain values are normalized as in Equations (1) and (2):

$$C_j = \frac{|r_j|}{\sum_{k=1}^{m} |r_k|} \tag{1}$$

$$G_j = \frac{IG_j}{\sum_{k=1}^{m} IG_k} \tag{2}$$

where r_j denotes the correlation coefficient between the metadata variable x_j and the melanoma target, G_j is its information gain, and $m = 3$ represents the three metadata variables. The final impact factor is then defined as Equation (3):

$$IF_j = \lambda C_j + (1 - \lambda) G_j \tag{3}$$

where λ controls the relative contribution of the correlation- and information-theoretic components (In this paper, $\lambda = 0.5$ was used to assign equal importance to the two components). Thus, IF_{age}, IF_{sex}, and IF_{site} provide standardized measures of the relative diagnostic relevance of the three clinical variables. The encoded metadata are then processed using an MLP to capture nonlinear interactions among age, sex, and

lesion location. The resulting metadata information and impact factors are incorporated into the final diagnostic score through the predefined global weighted-fusion strategy described in the next step. So, the coefficients a_{age}, a_{sex}, and a_{site} are treated as predefined global fusion weights rather than adaptive sample-specific coefficients learned through backpropagation. This formulation makes the contribution of clinical metadata explicit and reproducible while ensuring that complementary demographic and lesion-related information contributes to the final decision.

3.5. Final model integration

The final decision-making process is implemented through a predefined global weighted ensemble strategy that integrates the output of the image-based model with the contributions of the clinical metadata. The final diagnostic score (FinalS) is calculated as in Equation (4):

$$FinalS = M_{Im} \times a_{M_{Im}} + a_{age} \times IF_{age} + a_{sex} \times IF_{sex} + a_{site} \times IF_{site} \quad (4)$$

$$s.t.\ 0 \leq FinalS \leq 1,\ a_{M_{Im}} + a_{age} + a_{sex} + a_{site} = 1$$

where M_{Im} represents the image model output, while IF_{age}, IF_{sex}, and IF_{site} denote the impact factors of the metadata model. The coefficients $a_{M_{Im}}$, a_{age}, a_{sex}, and a_{site} are global fusion weights assigned to the image prediction and the three metadata components. These coefficients determine the relative contribution of each information source to the final diagnostic score and remain fixed for all samples during inference. Therefore, the fusion weights are not sample-dependent and are not dynamically generated from intermediate feature representations. Instead, they constitute a predefined global weighting scheme designed to combine visual and clinical information within a simple ensemble formulation. If $> W_{Best}$, the case is classified as melanoma ($FinalD = 1$); otherwise, it is classified as non-melanoma ($FinalD = 0$). This formulation provides an explicit fusion mechanism that allows the model to combine visual and clinical information rather than relying on image-based cues.

The decision threshold W_{Best} is optimized using the validation set. Candidate threshold values are evaluated using the ROC curve and the corresponding sensitivity–specificity trade-off. The threshold providing the most suitable diagnostic balance is selected and fixed before evaluation on the independent test set. The predefined fusion weights remain fixed throughout inference, and the optimized decision threshold is not adjusted using the final test samples. It provides a transparent multimodal fusion mechanism in which the contribution of the dermoscopic image prediction and each clinical metadata component is represented in the final score. Unlike image-only diagnostics, our ensemble allows clinically relevant contextual information to contribute to the classification decision.

The contribution of the proposed integration strategy lies in combining heterogeneous visual and clinical information through an interpretable weighted formulation while maintaining low computational complexity. Although the current fusion weights are global rather than input-adaptive, the framework provides a practical mechanism for incorporating metadata that may contain complementary diagnostic information not fully represented by dermoscopic appearance alone.

3.6. Final model evaluation

For the final evaluation step, the independent test set is preprocessed in a manner consistent with the training pipeline to ensure unbiased performance assessment. All dermoscopic images are first normalized to a fixed pixel intensity range of [0, 1] to eliminate illumination inconsistencies and guarantee compatibility with the trained BNDrop-XNet model. The clinical metadata are then organized, encoded, and standardized to align with the metadata model input structure. Finally, the integrated framework is applied to this normalized test set to generate final predictions based on the weighted ensemble strategy. Model performance is evaluated using accuracy, precision, recall, and F1-Score to provide an evaluation of model reliability. This consistent preprocessing and unified evaluation strategy ensures that the reported results reflect the generalization capability, robustness, and clinical applicability of our framework.

Algorithm 1 illustrates the workflow of the BNDrop-XNet framework. The first step focuses on data preprocessing, where dermoscopic images are normalized, samples with missing data are excluded, and the dataset is divided into two main parts: an image-model creation set and a final test set. Within the creation set, a stratified split maintains class balance, and SMOTE is also applied to the training subset to address the strong class imbalance. The second and third steps correspond to image model creation and metadata model construction. The image model, BNDrop-XNet, is a lightweight CNN consisting of four sequential convolutional blocks, each including Conv2D, Batch Normalization, MaxPooling, and Dropout layers. These layers work together to extract hierarchical features while minimizing overfitting. After fine-tuning its hyperparameters using the validation set, the trained image model is evaluated on an internal test subset. In parallel, metadata variables (age, sex, and lesion site) are encoded and normalized, and their impact factors are computed to quantify diagnostic relevance. An MLP is then trained to learn adaptive coefficients.

The fourth and fifth steps describe weighted ensemble integration and final evaluation. Using the outputs of the image and metadata models, a final score (*FinalS*) is computed as a weighted sum of both modalities. The adaptive weights are optimized on the validation set to satisfy the constraints. The decision threshold (W_{Best}) is then fine-tuned to maximize diagnostic performance. Finally, the integrated model is tested on the independent final test set. At last, the performance of the final model is evaluated through standard metrics. Together, these steps form an end-to-end interpretable and robust framework that integrates visual and clinical information to improve melanoma detection reliability.

4. Findings and results

This section presents the experimental findings of the BNDrop-XNet framework. We begin by describing the dataset, followed by a detailed presentation of the implementation results. Finally, our framework is compared with several recent state-of-the-art melanoma detection approaches to demonstrate the performance advantages of our model. All implementations were conducted in *Python* on an *Apple M2* system with *8 GB RAM*, running *macOS Sonoma 14.4.1*, providing a reproducible, efficient, and lightweight computational environment suitable for medical imaging research.

Table 3 summarizes the hyperparameters, configuration settings, and implementation details employed throughout the experiments. The designated subset of the dataset was divided into 1432 training, 128 test, and 32 validation samples, ensuring a balanced and stratified evaluation protocol. A batch size of 64 and a maximum of 200 epochs were used to train BNDrop-XNet, supported by SMOTE-based resampling to address class imbalance. The CNN architecture incorporated Conv2D filters of 64 and 128, ReLU activations, and dense layers with 256 and 1 unit for nonlinear representation and final classification. The model was optimized using the Adam optimizer with a learning rate of 1×10^{-4}, momentum parameters ($\beta_1 = 0.9,\ \beta_2 = 0.999$), and training stabilizers such as Early Stopping and ReduceLROnPlateau. For the ensemble phase, the learned fusion weights were set to $a_{M_{Im}} = 0.2$, $a_{age} = 0.2$, $a_{sex} = 0.2$, and $a_{site} = 0.4$, with an optimized decision threshold $W_{Best} = 0.5$. Finally, the performance was evaluated using standard metrics, including Accuracy, Precision, Recall, and F1-Score, providing a comprehensive assessment of diagnostic reliability.

Algorithm 1
Pseudo-code of the Proposed BNDrop-XNet Framework.

Input: Dermoscopic image set X with labels y and metadata Z = {age, sex, site};
parameters n_{train}, $n_{validation}$, n_{test}, B, and E_{Max}
Output: Final integrated model and evaluation metrics
1. Data preparation
Normalize pixel values of all images to [0,1]
Remove samples with missing or poor-quality metadata
Split the dataset into:
(a) Image-model creation set → for training, validation, and test
(b) Final test set → for unbiased evaluation
Within the image-model creation set:
Stratified split into train, validation, and test
Apply SMOTE on training data to balance melanoma / non-melanoma classes
2. Image model creation (BNDrop-XNet)
Define CNN with four sequential convolutional blocks:
For blocks 1–3:
Conv2D(filters = 32→64→128, kernel_size = (3,3), activation = ReLU)
BatchNormalization()
MaxPooling2D(2,2)
Dropout(rate = 0.25→0.25→0.40)
Block 4:
Flatten → Dense(256, ReLU) → Dropout(0.50) → Dense(1, Sigmoid)
Train the model using the Adam optimizer,
loss = binary_crossentropy, metrics = {Accuracy, ROC-AUC, PR-AUC}
Fine-tune hyperparameters (learning rate, batch size, dropout) on the validation set
Evaluate the trained image model $\boldsymbol{M_{IM}}$ on the internal test subset
3. Metadata model creation and impact analysis
Encode and normalize metadata variables (age, sex, and lesion site)
For each metadata variable x_j:
Compute the correlation coefficient r_j between x_j and target y
Compute information gain IG_j between x_j and target y
Normalize correlation contribution as (1)
Normalize information-gain contribution as (2)
Compute metadata impact factor as (3)
Obtain IF_{age}, IF_{sex}, and IF_{site}
Train the metadata MLP to capture nonlinear interactions among age, sex, and lesion location.
4. Weighted ensemble integration
Compute final score for each sample:
$FinalS = M_{Im} \times a_{M_{Im}} + a_{age} \times IF_{age} + a_{sex} \times IF_{sex} + a_{site} \times IF_{site}$
Subject to:
$0 \le FinalS \le 1,\ a_{M_{Im}} + a_{age} + a_{sex} + a_{site} = 1$
Set predefined global fusion weights
Determine the best decision threshold W_{Best} for sensitivity–specificity balance:
if FinalS > W_{Best} → FinalD = 1 (Melanoma)
else → FinalD = 0 (Non-melanoma)
5. Final evaluation
Normalize pixel values and encode metadata of the independent final test set
Apply integrated model and decision rule
Compute evaluation metrics:
Accuracy, Precision, Recall, and F1-Score
Optionally perform subgroup analyses (age, sex, lesion site)
Return final integrated BNDrop-XNet model, predefined global fusion weights, optimized decision threshold, and performance metrics

4.1. Dataset

The proposed framework was evaluated using the ISIC 2020 Challenge Dataset, a large-scale dermoscopic image collection provided for melanoma classification (Cassidy et al., 2022). It contains high-resolution dermoscopic images labelled as melanoma or non-melanoma, accompanied by relevant metadata such as patient age, sex, and anatomical location of the lesion. The clinical variables play an essential role in melanoma risk assessment and are integrated into our metadata model. Before analysis, all samples were screened for completeness, and entries with missing critical information or low-quality images were excluded to ensure reliability. The dataset was then divided into designated subsets for model training, validation, and final testing, following a stratified strategy to preserve the natural class distribution. This structured dataset preparation ensures that both visual and contextual patient information are represented throughout the modelling process.

Table 4 summarizes the metadata fields of the dataset and indicates whether each feature was included or excluded in our modelling pipeline. The identifiers image_name, patient_id, and lesion_id, although non-null, were dropped because they do not provide diagnostic information and may introduce unwanted bias or data leakage. In contrast, sex, age, and anatom_site were retained, as these clinical features have documented relevance to melanoma risk. The diagnosis field provides textual detection labels but was not directly used in modelling, while the target field (binary 0/1 label for benign vs. melanoma) was retained as the ground-truth output variable. This selection process ensures that only meaningful metadata contributes to the prediction framework, while redundant or biased identifiers are excluded.

Fig. 3 shows the distribution of key metadata features in the ISIC 2020 dataset. Fig. 3(a) illustrates an uneven distribution across anatomical sites, with the torso representing the largest portion of the dataset (approximately 47–50% of all samples). This is followed by the lower extremity (~25–27%) and upper extremity (~15–18%), while regions such as head/neck, palms/soles, and oral/genital areas constitute a small minority (collectively <10%). This skew highlights a real-world data challenge: melanoma risk varies across body sites, but the frequency of available dermoscopic images does not align uniformly with clinical prevalence patterns. Such an imbalance can bias image-only DL models toward more common anatomical regions and reduce

Table 3
Hyperparameter configuration, training settings, and ensemble coefficients.

Hyperparameter	Value/Method
n_{train}	1432
n_{test}	128
$n_{validation}$	32
B	64
E_{Max}	200
random_state	42
Dropout (rate)	{0.25, 0.4, 0.5}
Resampling	SMOTE
Conv2D	Filter = {64, 128}, Activation = ReLU
Dense	Unit={256,1}, Activation = {ReLU, Sigmoid}
Optimizer	Adam(learning_rate = 1e-4), Adaptive learning with momentum ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
callbacks	[early_stop, reduce_lr]
$a_{M_{Im}}$	0.2
a_{age}	0.2
a_{sex}	0.2
a_{site}	0.4
W_{Best}	0.5
Metrics	Accuracy, Precision, Recall, F1-Score

Table 4
Description of the metadata fields in the ISIC 2020 dataset.

Name	Data Type	Null	Used/Drop
image_name	object	non-null	Drop
patient_id	object	non-null	Drop
lesion_id	object	non-null	Drop
sex	object	non-null	Used
Age	float64	non-null	Used
anatom_site	object	non-null	Used
diagnosis	object	non-null	
target	**int64**	**non-null**	**Used (Label)**

generalizability. The proposed framework overcomes this limitation by incorporating metadata-based impact factors, enabling the model to account for differences in melanoma likelihood associated with anatomical site when visual data from rare sites are underrepresented.

Fig. 3(b) shows that the dataset contains near-equal representation of male and female patients, with females making up 49–50% and males around 50–51%. Although the distribution is balanced, sex remains an influential clinical feature because melanoma incidence, anatomical site distribution, and lesion morphology differ between sexes in epidemiological studies. For example, males are more prone to melanoma on the trunk, while females more often present lesions on the lower extremities. Image-only frameworks cannot capture these meaningful clinical patterns. In contrast, BNDrop-XNet integrates the sex metadata coefficient (a_{sex}) and impact factor to adjust decision-making.

Fig. 3(c) demonstrates that the age distribution follows a near-normal pattern, with the majority of samples falling between 35 and 70 years, and a peak frequency around 45–55 years. Younger (<20) and older (>80) patients constitute a small fraction of the dataset (<5%), which may lead traditional classifiers to underperform on age groups that deviate from the central tendency of the dataset. Age, however, is a significant predictor of melanoma severity. BNDrop-XNet addresses this limitation by using age-based impact values (IF_{age}) and adaptive weighting, allowing our model to compensate for demographic imbalance and incorporate non-visual risk factors into its decision process. So, the framework reduces bias toward the dominant middle-aged population and improves fairness across the full age spectrum.

The correlation analysis presented in Table 5 shows that the relationships between the clinical metadata (sex, age, and anatomical site) and the melanoma target label are weak (still meaningful within the context of a multi-modal diagnostic framework). Age shows the strongest positive correlation with the target (r = 0.086), indicating that increasing age is associated with a higher likelihood of melanoma, consistent with known epidemiological trends. Sex exhibits a small positive correlation with melanoma (r = 0.028), suggesting a minor difference in melanoma prevalence between males and females. Anatomical site shows a slight negative correlation with the target (r = –0.01), implying that lesion location has limited linear predictive power but may contribute to non-linear patterns that our metadata model can capture. Although all correlations fall below 0.10 (i.e., <10% linear association), their combined weak–moderate interactions justify their inclusion in the BNDrop-XNet framework, where impact factors and adaptive weighting allow these subtle signals to be used within a multi-modal ensemble for improved melanoma detection.

Fig. 4 illustrates the distribution of melanoma and non-melanoma samples across the training, validation, and test sets used for image- and metadata-based model development. In the training set, which contains 1432 samples, non-melanoma cases account for 894 samples (62.43%), while melanoma cases represent 538 samples (37.57%), providing an imbalanced but representative learning environment after controlled sampling. The validation set maintains a similar proportion, with 21 non-melanoma (65.62%) and 11 melanoma samples (34.38%), ensuring stable hyperparameter tuning without class-driven bias. The test set includes 93 non-melanoma (72.66%) and 35 melanoma samples (27.34%), creating a challenging evaluation scenario that reflects real-world class prevalence. Overall, the complete dataset used for model development comprises 1592 samples, with 1008 non-melanoma and 584 melanoma cases, preserving the natural imbalance of melanoma screening tasks. These well-structured splits highlight the strength of our framework. Indeed, the model can handle class imbalance while maintaining generalizability across subsets by combining stratified partitioning, controlled resampling, and metadata-driven learning. This balanced design ensures that our multi-modal fusion strategy uses both visual features and clinical metadata, enabling the final integrated model to perform reliably on unseen and imbalanced real-world data.

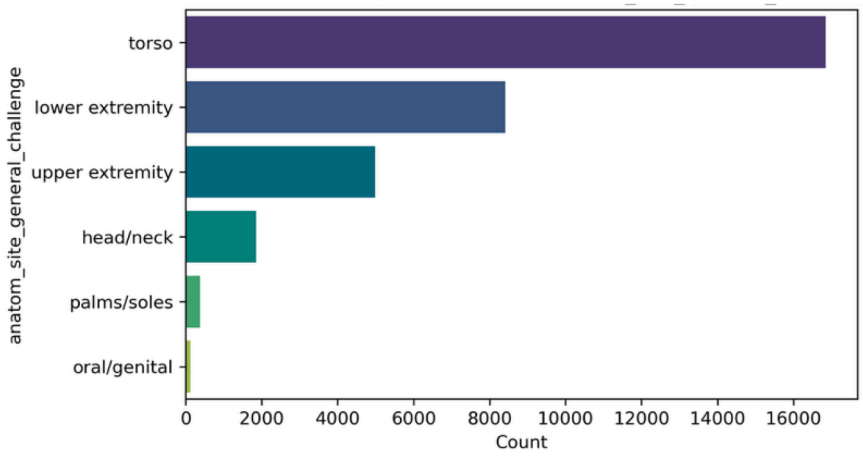
4.2. Implementation results

To evaluate the effectiveness of the BNDrop-XNet framework, we first optimized the decision threshold W used in the weighted ensemble model, and then analyzed the resulting confusion matrix to assess diagnostic behaviour.

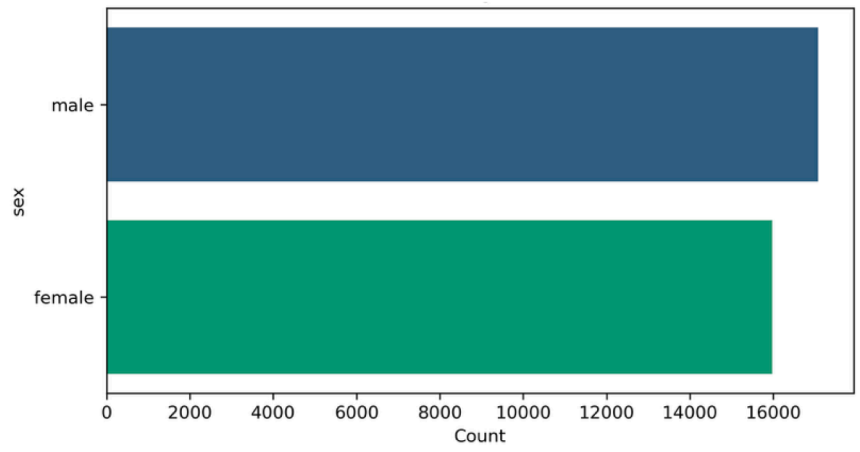
Fig. 5 illustrates the threshold-tuning process and its influence on performance metrics. Fig. 5(a) shows how precision, recall, F1-Score, and accuracy change as a function of W, highlighting the threshold region where the model achieves its optimal predictive balance. In this figure, precision increases sharply when $W > 0.40$, reaching its peak around 0.50, while recall decreases beyond this region, illustrating the classic trade-off between the two metrics. The F1-Score and accuracy curves converge around the same threshold, confirming that $W = 0.50$ provides the most stable diagnostic performance. This alignment indicates that the weighted fusion of image and metadata outputs is balanced in a way that maximizes melanoma detection without compromising the false-positive rate. The smooth transitions of the curves also demonstrate the stability of the ensemble and its capacity to maintain discriminative power across varying decision thresholds.

Fig. 5(b) presents the relationship between accuracy and Youden's J statistic, used to identify the threshold that maximizes diagnostic usefulness by balancing sensitivity and specificity. It further confirms that the optimal threshold corresponds to the maximum of both the accuracy curve and Youden's J, with the peak located at $W = 0.50$. This threshold ensures a trade-off between sensitivity and specificity, which is critical in melanoma screening scenarios where both missed diagnoses and unnecessary alarms must be minimized.

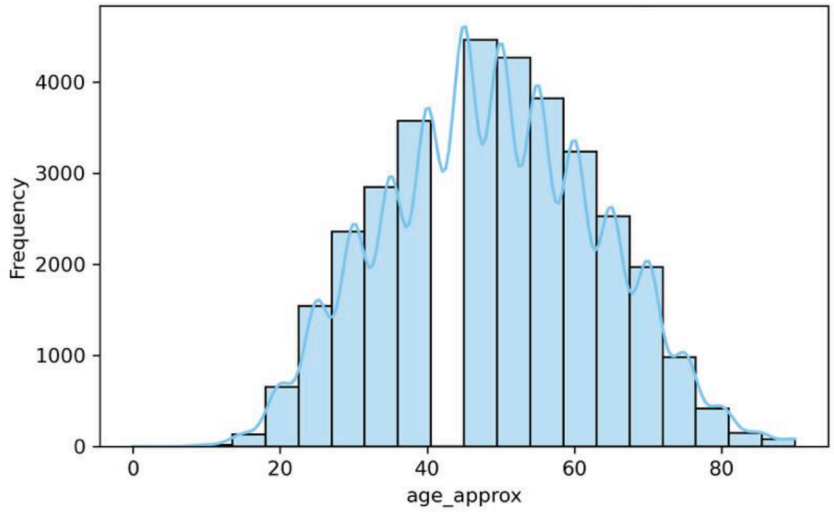
Fig. 5(c) displays how the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) evolve with changes in the threshold, providing insight into the error patterns and robustness of the model (Yousefi et al., 2020, Babu Mupparaju and Reddy, 2024). It shows that at the optimized threshold, FP decrease



(a)



(b)



(c)

Fig. 3. Distribution of key metadata features in the ISIC 2020 dataset: (a) anatomical site, (b) sex, and (c) age.

sharply while FN remain controlled, and TN rise substantially. This behaviour confirms that the integrated model is capable of distinguishing melanoma from benign lesions more effectively than image-only or metadata-only baselines, benefiting from the strengths of the two modalities.

The confusion matrix of our final integrated model, shown in Fig. 6, demonstrates reliable diagnostic performance across both melanoma and non-melanoma classes (Yousefi et al., 2024, Siarov et al., 2024). It provides a class-specific assessment of performance under the remaining class imbalance. For the non-melanoma class, 194 of 202 cases were classified, corresponding to a precision of 91.94%, recall of 96.04%, and F1-score of 93.95%. For the melanoma class, 33 of 50 cases were

Table 5
Pearson correlation matrix of clinical metadata.

	Sex	Age	Anatom Site	Target
Sex	1	0.11	0.022	0.028
Age	0.11	1	-0.041	0.086
Anatom Site	0.022	-0.041	1	-0.01
Target	0.028	0.086	-0.01	1

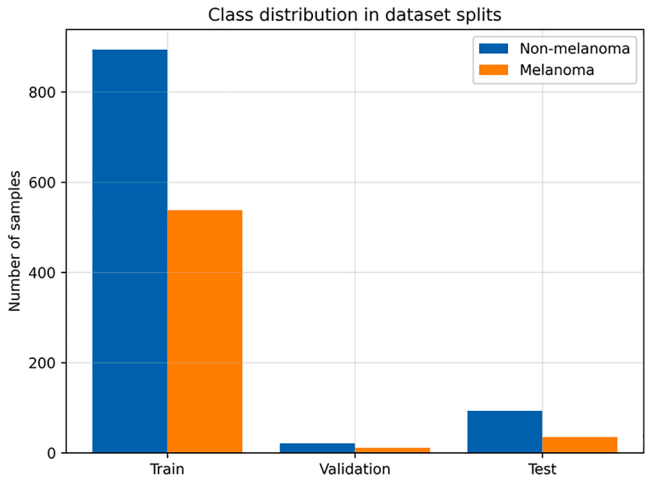


Fig. 4. Class distribution across the training, validation, and test subsets for melanoma and non-melanoma samples.

identified, yielding a precision of 80.49%, recall (sensitivity) of 66.00%, and F1-score of 72.53%, while 17 melanoma cases were misclassified as non-melanoma. Although the overall accuracy reached 90.08%, the balanced accuracy was 81.02%, and the lower melanoma recall and F1-score demonstrate that residual class imbalance remains a challenge. The findings indicate that SMOTE improved minority-class representation during training but did not eliminate the performance disparity between melanoma and non-melanoma cases. Thus, class-specific metrics provide a more informative assessment of the framework than overall accuracy alone.

Table 6 provides a class-wise evaluation of BNDrop-XNet and highlights the effect of class imbalance on model performance. For the non-melanoma class, the model achieves high precision (91.94%), recall (96.04%), and F1-score (93.95%), indicating that benign cases are identified consistently with relatively few FP and FN errors. In contrast, the melanoma class shows lower performance, with a precision of 80.49%, recall of 66.00%, and F1-score of 72.53%. The lower recall indicates that a notable proportion of melanoma cases are still missed, which explains why the melanoma F1-score remains modest despite the high overall accuracy reported for the model. This difference between the two classes suggests that the class-imbalance problem is not resolved by SMOTE alone and that the model is more reliable for identifying non-melanoma cases than melanoma cases. Nevertheless, the melanoma precision of 80.49% shows that when the model predicts melanoma, its predictions are reliable. Overall, the class-wise results provide a more informative assessment than accuracy and indicate that future improvements should focus on increasing melanoma sensitivity and reducing FN cases.

4.3. Comparison results

To demonstrate the effectiveness of the proposed framework, we compared its performance with four recent state-of-the-art melanoma detection approaches. Table 7 summarizes the results across four standard metrics. These models include MD-MML (Najjar-Ghabel et al., 2025), AMD (Yousefi et al., 2024), VGG 16 (Alzamel et al., 2025), and ResNet152V2 (Thakur et al., 2025), which represent widely used CNN baselines and advanced deep architectures.

As shown in Table 7, our BNDrop-Xnet model outperforms all competing methods in overall accuracy, achieving 90%, which is 2% higher than MD-MML (88%), 20% higher than AMD (70%), and exceeds traditional CNN baselines such as VGG 16 (+34%). This substantial gain demonstrates the effectiveness of integrating dermoscopic features with clinical metadata using our weighted fusion mechanism. Similarly, our model achieves the highest precision at 80%, outperforming MD-MML by 14 percentage points, AMD by 9 points, and VGG 16 and ResNet152V2 by more than 59 points, indicating a much lower FP rate and stronger diagnostic reliability.

In terms of F1-Score, our model also ranks first with 73%, slightly ahead of MD-MML (72%) and higher than AMD (68%). Compared with VGG 16 (26%) and ResNet152V2 (33%), BNDrop-XNet achieves 180–280% relative improvement, demonstrating that metadata-aware feature fusion enhances classification consistency. Although our recall (66%) is slightly lower than MD-MML (80%), it remains competitive and higher than AMD (65%), VGG 16 (41%), and close to ResNet152V2 (86%) while maintaining the best balance between sensitivity and specificity. This balanced performance is crucial for clinical settings, where models must avoid excessive FP without compromising melanoma detection.

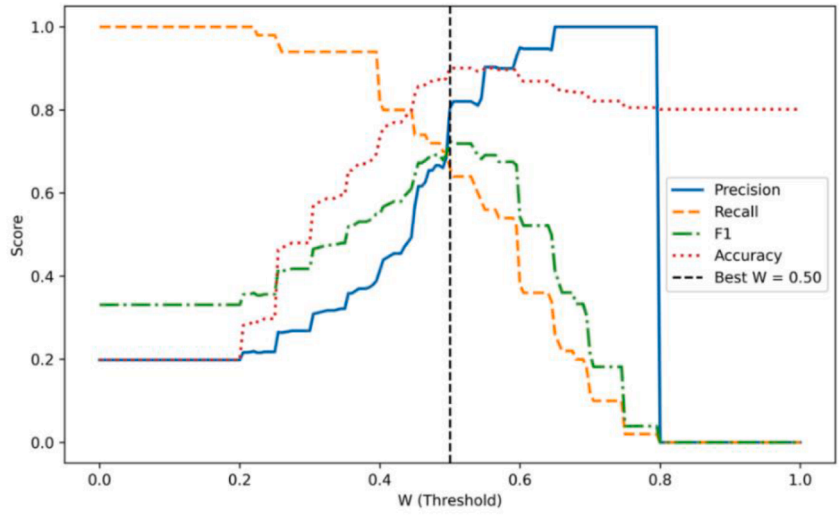
Overall, the comparison shows that BNDrop-XNet provides the most stable performance across all major metrics, driven by its lightweight CNN design, metadata-aware MLP modelling, and predefined global weighted fusion strategy. While some competing models may achieve higher recall alone, they suffer from low precision and accuracy, making them impractical for real-world applications. In contrast, our framework maintains a meaningful balance, achieving high accuracy, high precision, and a strong F1-Score while avoiding the diagnostic instability present in other DL architectures.

5. Discussion

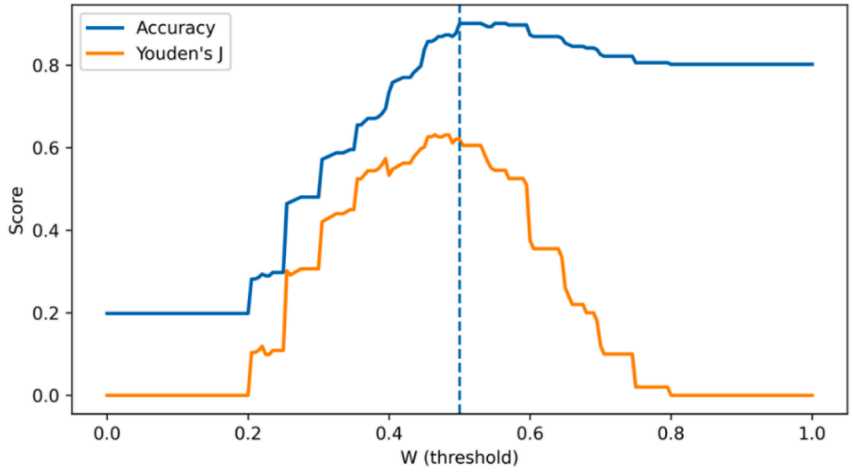
The results of this paper demonstrate that BNDrop-XNet offers a powerful improvement over existing melanoma detection approaches. By integrating dermoscopic image features with patient-level metadata through a predefined global weighted fusion strategy, our model outperforms recent state-of-the-art models across accuracy, precision, and F1-score. The precision of the model of 80% and accuracy of 90% indicate a substantial reduction in FP, an essential requirement for minimizing unnecessary biopsies and patient anxiety in dermatology practice. Moreover, the balanced F1-Score of 73% highlights the ability of our framework to maintain diagnostic stability across imbalanced classes when melanoma cases are limited (a common challenge in real-world skin cancer datasets).

Another key strength of BNDrop-XNet is its multi-modal design, which uses information beyond dermoscopic image content. Age, sex, and lesion location each contribute additional diagnostic cues that pure image-based CNNs often overlook. The metadata impact–factor mechanism enables the model to incorporate epidemiological patterns, such as age-dependent risk and site-specific variability, into its decision-making process. Threshold optimization further enhances clinical utility by fine-tuning the sensitivity–specificity balance, ensuring the model can be tailored to different screening priorities. These characteristics demonstrate how our model can support clinicians in early melanoma recognition in resource-limited or high-volume screening environments where reliable decision support is needed.

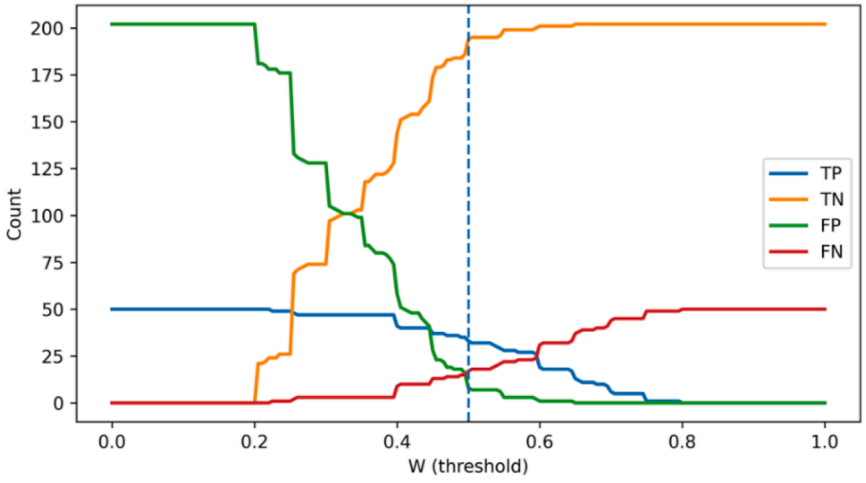
Although SMOTE was used to improve the representation of the minority class during training, the class-specific results show that the imbalance was not fully addressed. Melanoma recall remained at 66.00%, with 17 false-negative cases, despite an accuracy of 90.08%, indicating that oversampling was not sufficient to close the performance gap between the two classes. Future work should consider additional imbalance-handling strategies, such as cost-sensitive learning, focal loss,



(a)



(b)



(c)

Fig. 5. Effect of decision threshold *W* on the performance of BNDrop-Xnet: (a) Variation of Precision, Recall, F1-Score, and Accuracy across different threshold values, (b) Threshold optimization using Accuracy and Youden's J statistic, and (c) Impact of threshold variation on confusion matrix components.

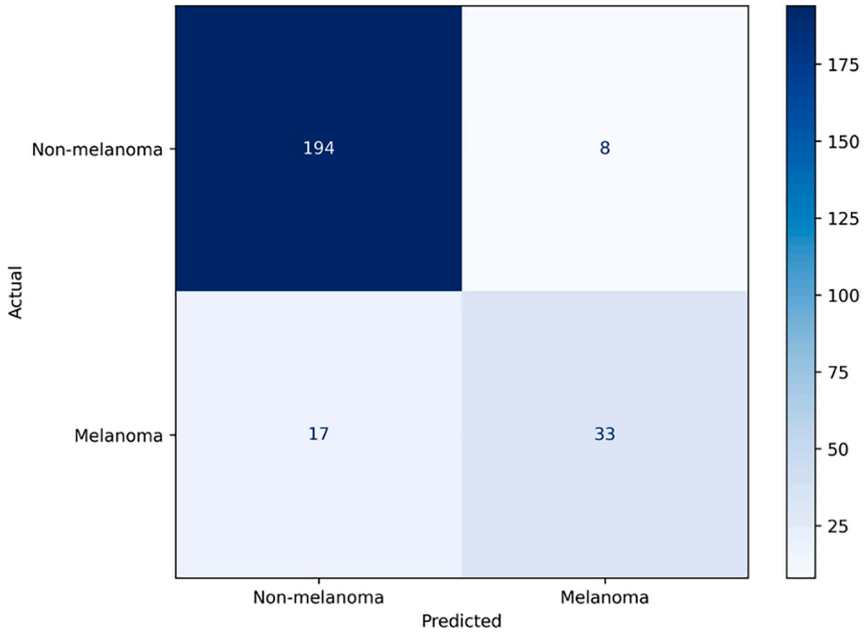


Fig. 6. Confusion matrix of BNDrop-XNet.

Table 6
Class-wise performance of BNDrop-XNet based on the confusion matrix.

Class	Precision	Recall	F1-score
Non-melanoma	91.94%	96.04%	93.95%
Melanoma	80.49%	66.00%	72.53%

Table 7
Performance comparison of BNDrop-XNet with recent state-of-the-art melanoma detection models.

Approach	Accuracy	F1-Score	Precision	Recall
Our Model (BNDrop-XNet)	90%	73%	80%	66%
MD-MML (Najjar-Ghabel et al., 2025)	88%	72%	66%	80%
AMD (Yousefi et al., 2024)	70%	68%	71%	65%
VGG 16 (Alzamel et al., 2025)	56%	26%	20%	41%
ResNet152V2 (Thakur et al., 2025)	32%	33%	21%	86%

class-weighted optimization, and targeted augmentation. The final integrated model was evaluated on an independent test set that was kept separate from model development; however, the small test set and the absence of cross-validation limit the strength of broader generalization claims. Future studies should incorporate stratified cross-validation and external validation on larger and more heterogeneous datasets. The ISIC 2020 dataset also has limited diversity in terms of imaging devices and patient demographics, which affects model generalizability across different clinical populations. In addition, the current framework uses only three metadata variables—age, sex, and lesion site—whereas richer clinical information, such as skin type, lesion evolution, and patient history, may further improve predictive performance. Finally, although BNDrop-XNet was designed as a lightweight architecture, deployment on embedded or resource-constrained devices may still require further optimization, such as model quantization. Future research should therefore focus on improving minority-class sensitivity, incorporating richer clinical metadata, validating the model across diverse datasets, and applying more advanced explainability and deployment-oriented techniques to support reliable integration into dermatological workflows.

6. Conclusion

In this paper, we introduced BNDrop-XNet, a lightweight and multimodal melanoma detection framework that integrates dermoscopic image features with clinical metadata through a predefined global weighted fusion strategy. The proposed model demonstrated superior performance compared to recent DL approaches, achieving 90% accuracy, 80% precision, and a 73% F1-Score, underscoring the value of combining visual and contextual patient information for robust melanoma screening. The results confirm that metadata-aware fusion enhances diagnostic stability, reduces FP, and offers improved clinical relevance over image-only DL methods.

Future work will focus on addressing the limitations identified in this paper. Improved imbalance-handling strategies, including cost-sensitive learning, focal loss, class-weighted optimization, and targeted augmentation, will be investigated to enhance melanoma sensitivity. The metadata space will also be expanded to incorporate additional clinically relevant factors. To strengthen model robustness, future studies will employ stratified cross-validation and external validation using larger, multi-center, and more heterogeneous datasets. Further research will also explore advanced explainable AI to improve model interpretability, together with deployment-oriented optimization methods such as model quantization for real-time use. These developments are expected to improve the clinical applicability of BNDrop-XNet for early melanoma detection in diverse healthcare environments.

Data availability

The data used in this study are publicly available from the ISIC 2020 Challenge dataset (Link: https://challenge2020.isic-archive.com/).

CRediT authorship contribution statement

Samad Najjar-Ghabel: Writing – review & editing, Validation, Software, Project administration, Formal analysis, Conceptualization. **Shamim Yousefi:** Writing – review & editing, Writing – original draft, Resources, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the University of Mohaghegh Ardabili (Research grant number 1404/D/9/21636).

References

M. Ahmad, I. Ahmed, A. Chehri, G. Jeon, Fusion of metadata and dermoscopic images for melanoma detection: Deep learning and feature importance analysis, Information Fusion. 124 (2025) 103304. doi:10.1016/j.inffus.2025.103304.

Alzamel, M., Iliopoulos, C., Lim, Z., 2025. Deep learning approaches and data augmentation for melanoma detection. Neural Comput. Appl. 37, 10591–10604. https://doi.org/10.1007/s00521-024-10590-8.

Babu Mupparaju, S., Reddy, R., 2024. A Comprehensive Analysis of Melanoma Skin Cancer Detection Using Machine Learning and Deep Learning Algorithms. In: 2024 International Conference on Data Science and Network Security (ICDSNS). IEEE, pp. 1–5. https://doi.org/10.1109/ICDSNS62112.2024.10691081.

Cai, Z., Chen, Y., Wang, J., He, X., Pei, Z., Lei, X., Lu, C., 2026. DAFNet: A novel Dynamic Adaptive Fusion Network for medical image classification. Inf. Fusion 126, 103507. https://doi.org/10.1016/j.inffus.2025.103507.

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H., 2022. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Med. Image Anal. 75, 102305. https://doi.org/10.1016/j.media.2021.102305.

A. Ferreirinha, V. Farricha, A.L. João, Melanoma Diagnosis With 3D Total-Body Photography, Actas Dermo-Sifiliográficas. (2025). doi:10.1016/j.ad.2024.09.030.

Idwan, S., Etaiwi, W., Rafayia, H., Matar, I., 2025. A comprehensive review of statistical variants and enhancements of SMOTE oversampling method. Int. J. Data Sci. Anal. 20, 6887–6904. https://doi.org/10.1007/s41060-025-00878-w.

Javed, R., Rahim, M.S.M., Saba, T., Rehman, A., 2020. A comparative study of features selection for skin lesion detection from dermoscopic images. Netw. Model. Anal. Health Inform. Bioinform. 9, 4. https://doi.org/10.1007/s13721-019-0209-1.

Kakish, D.R.K., AlSamhori, J.F., Ayman, A., Al-Sawalha, M., Hijazeen, T., Clementina, R., Ahmed, R., Masadeh, N., Al-Zuriqat, M., Mohammed, A.Y., Al kurdi, W., Nashwan, A.J., 2025. Amelanotic melanoma: Diagnostic challenges, treatment innovations, and the emerging role of in early detection. J. Med. Surg. Public Health 6, 100189. https://doi.org/10.1016/j.glmedi.2025.100189.

Karimi, M., Karimi, Z., Khosravi, M., Delaram, Z., Dehsheikhim, M.H., Najafabadi, S.A., Alizadeh Aliabadi, M., Tavakoli, N., 2025. Feature Selection Methods in Big Medical Databases:A Comprehensive Survey. Int. J. Theor. Appl. Comput. Intell. 181–209. https://doi.org/10.65278/IJTACI.2025.21.

Kaur, R., GholamHosseini, H., Lindén, M., 2025. Advanced Deep Learning Models for Melanoma Diagnosis in Computer-Aided Skin Cancer Detection. Sensors 25, 594. https://doi.org/10.3390/s25030594.

Khan, A.R., Mujahid, M., Alamri, F.S., Saba, T., Ayesha, N., 2025. Early-Stage Melanoma Cancer Diagnosis Framework for Imbalanced Data From Dermoscopic Images. Microsc. Res. Tech. 88, 797–809. https://doi.org/10.1002/jemt.24736.

L. Kong, J.D. Velasquez, V. Snášel, M. Pant, J.-S. Pan, J. Nowakova, Enhancing skin cancer detection through category representation and fusion of pre-trained models, Information Fusion. 124 (2025) 103369. doi:10.1016/j.inffus.2025.103369.

Marie, M.I., Elredeny, M.S., Essayed Yakoub, A., 2025. A Hybrid Model for Early Melanoma Detection: Integrating YOLOv9 and Faster R-CNN for Enhanced Diagnostic Accuracy. IEEE Access 13, 120327–120344. https://doi.org/10.1109/ACCESS.2025.3587625.

L.N. Montoya, J.S. Roberts, B.S. Hidalgo, Towards Fairness in AI for Melanoma Detection: Systemic Review and Recommendations, in: 2025: pp. 320–341. doi:10.1007/978-3-031-84460-7_21.

Najjar-Ghabel, S., Yousefi, S., Habibi, P., 2024. Comparative Analysis and Practical Implementation of Machine Learning Algorithms for Phishing Website Detection. In: 2024 9th International Conference on Computer Science and Engineering (UBMK). IEEE, pp. 1–6. https://doi.org/10.1109/UBMK63289.2024.10773479.

Najjar-Ghabel, S., Yousefi, S., Ismael, E.H., 2025. Enhancing Melanoma Detection through Multi-Modal Learning: Integration of Dermoscopic Images and Clinical Data. In: 2025 IEEE 7th Symposium on Computers & Informatics (ISCI). IEEE, pp. 70–75. https://doi.org/10.1109/ISCI65687.2025.11167604.

Nasir, S., Bilal, M., Khalidi, H., 2025. Detection and Classification of Skin Cancer by using CNN-enabled Cloud Storage Data Access Control Algorithm based on Blockchain Technology. Int. J. Theor. Appl. Comput. Intell. 2025, 145–169. https://doi.org/10.65278/IJTACI.2025.31.

Nazari, S., Garcia, R., 2025. Going Smaller: Attention-based models for automated melanoma diagnosis. Comput. Biol. Med. 185, 109492. https://doi.org/10.1016/j.compbiomed.2024.109492.

Oniga, M., Sultana, A., Alexandrescu, B., Orzan, O., 2025. Towards an integrated imaging for melanoma diagnosis: A review of multispectral, hyperspectral, and thermal technologies with preliminary system development. Comput. Biol. Med. 185, 109570. https://doi.org/10.1016/j.compbiomed.2024.109570.

Siarov, J., Siarov, A., Kumar, D., Paoli, J., Mölne, J., Neittaanmäki, N., 2024. Deep learning model shows pathologist-level detection of sentinel node metastasis of melanoma and intra-nodal nevi on whole slide images. Front. Med. 11. https://doi.org/10.3389/fmed.2024.1418013.

Thakur, A., Mishra, S.K., Bhutani, M., Dash, P.P., Upadhyay, P.K., Sahu, S.S., 2025. Deep learning approaches for detecting malignant melanoma in dermoscopic imagery. Int. J. Inf. Technol. https://doi.org/10.1007/s41870-025-02651-5.

Virgens, G.S., Teodoro, J.A., Iarussi, E., Rodrigues, T., Amaral, D.T., 2025. Enhancing and advancements in deep learning for melanoma detection: A comprehensive review. Comput. Biol. Med. 194, 110533. https://doi.org/10.1016/j.compbiomed.2025.110533.

Ye, Z., Zhang, D., Zhao, Y., Chen, M., Wang, H., Seery, S., Qu, Y., Xue, P., Jiang, Y., 2024. Deep learning algorithms for melanoma detection using dermoscopic images: A systematic review and meta-analysis. Artif. Intell. Med. 155, 102934. https://doi.org/10.1016/j.artmed.2024.102934.

Yousaf, K., Mehmood, Z., Saba, T., Rehman, A., Munshi, A.M., Alharbey, R., Rashid, M., 2019. Mobile-Health Applications for the Efficient Delivery of Health Care Facility to People with Dementia (PwD) and Support to Their Carers: A Survey. BioMed Res. Int. 2019, 1–26. https://doi.org/10.1155/2019/7151475.

Yousefi, S., Derakhshan, F., Karimipour, H., 2020. Applications of Big Data Analytics and Machine Learning in the Internet of Things. Handbook of Big Data Privacy. Springer International Publishing, Cham, pp. 77–108. https://doi.org/10.1007/978-3-030-38557-6_5.

Yousefi, S., Najjar-Ghabel, S., Danehchin, R., Band, S.S., Hsu, C.-C., Mosavi, A., 2024. Automatic melanoma detection using discrete cosine transform features and metadata on dermoscopic images. J. King Saud Univ. - Comput. Inf. Sci. 36, 101944. https://doi.org/10.1016/j.jksuci.2024.101944.

Yousefi, S., Najjar-Ghabel, S., Owaid, Z.S., 2025. A Supervised Learning-based Framework for Failure Detection in Toy Cars Using Acoustic Signal Analysis. In: 2025 IEEE 7th Symposium on Computers & Informatics (ISCI). IEEE, pp. 76–81. https://doi.org/10.1109/ISCI65687.2025.11167791.

Yousefi, S., Najjar-Ghabel, S., Shafaei, H., 2024. A Technical Analysis and Practical Implementation of Machine Learning Algorithms for Predicting Survival in Breast Cancer Patients. In: 2024 9th International Conference on Computer Science and Engineering (UBMK). IEEE, pp. 1168–1173. https://doi.org/10.1109/UBMK63289.2024.10773521.

Zhang, P., Chaudhary, D., 2025. Hybrid Deep Learning Framework for Enhanced Melanoma Detection. IEEE Trans. Comput. Biol. Bioinform. 1–11. https://doi.org/10.1109/TCBBIO.2025.3596589.