PENINSULA PUBLISHING LLC

EDRAAK
Vol. **(2025)**, 2025, **pp**. 119–125
ISSN: 3078-8412

Research Article

Developing an Efficient Deep Learning Model for the Classification of Skin Lesions

Yasmin Makki Mohialden[1,*], Saba Abdulbaqi Salman[1],

[1] *Computer Science Department, College of Science, Mustansiriyah University, Baghdad, Iraq*

[2] *Department of Computer Science, College of Education, Al-Iraqia University, Baghdad, Iraq*

ARTICLE INFO

Article History
Received 10 Aug 2025
Accepted: 6 Oct 2025
Revised 5 Nov 2025
Published 20 Nov 2025

Keywords
skin lesion classification,
deep learning,
ResNet101V2,
ISIC dataset,
Monte Carlo dropout,
explainable AI,
Grad-CAM,
medical image analysis.

ABSTRACT

This paper proposes a deep learning technique for classifying dermoscopic skin lesion pictures, improving diagnostic precision and clinical decision-making. The model extracts features using ResNet101V2 and infers using Monte Carlo dropout to determine prediction certainty. Grad-CAM additionally shows the model's output-affecting regions. This makes it easy to assess whether the network is targeting therapeutically relevant structures.

A modular software system tracked preprocessing, training, and assessment to ensure approach clarity. Repeating trials and making systematic model adjustments is easy with this framework. We processed ISIC using defined metadata and stratified splitting. This was balanced and simple to comprehend how the model operated.

Due to a lack of representative samples and an evident class imbalance, the framework performed well for the majority class but could not identify melanoma. Use uncertainty estimate and visual explanation combined to assess model reliability and identify ambiguous predictions. The development of dermatology-focused AI systems that are clinically interpretable and dependable is supported by these components.

1. INTRODUCTION

Malignant lesions, particularly melanoma, must be identified quickly and accurately in clinical practice due to the worldwide health crisis of skin cancer. Due to their similar visual appearance and small structural changes, even expert dermatologists have trouble distinguishing benign from malignant tumors [1,2].

Recently, deep-learning algorithms have showed promise in understanding dermoscopic images by identifying fine-grained diagnostic traits, which may help clinical judgment [3]. Current techniques are promising, but they cannot assess predicted uncertainty, give clear interpretability, or provide consistent reproducibility across trials, making them dangerous for clinical processes [4–6].

The study utilized the ISIC dataset to create a software-engineered deep learning model to binary identify melanoma (MEL) and nevus (NV). The model correctly identified NV but failed to identify any melanoma samples owing to insufficient data, a severe imbalance between the two groups, and the lesions' morphological similarity. This pattern parallels deterministic CNN models' default to the majority class and lack of prediction reliability when minority data are few [7,8].

To solve these problems, Monte Carlo dropout quantifies prediction uncertainty and Grad-CAM shows image regions that impact classifier decisions. More medical AI uses similar methods to detect problematic predictions and check whether the model is focusing on diagnostically relevant areas [9–11]. Transparency and repeatability improved by embedding these components in a modular and traceable computational process promote disciplined experimentation and clinically reliable AI systems [12–15].

Convolutional Neural Networks (CNNs) hierarchical feature representations capture dermatological signals such asymmetry, border irregularities, and color variation that discriminate benign in malignant skin diseases [1,3]. After adequate training data, ResNet101V2's residual architecture permits gradient flow across its deeper layers to collect multi-scale features that enhance classification [4,5]. This study found that the model performed well on NV because the class had enough instances for consistent feature learning.

*Corresponding author email: ymmiraq2009@uomustansiriyah.edu.iq
DOI: https://doi.org/10.70470/EDRAAK/2025/015

The model incorrectly classified all MEL samples as NV. Training CNNs on tiny, unbalanced datasets is difficult because the minority class is underrepresented, making it difficult to learn discrete boundaries, leading in skewed predictions and clinically unacceptable false-negative rates [6–8].

Deterministic CNNs exacerbate this issue by making single, point-estimate forecasts without reflecting decision uncertainty. Inference using Monte Carlo dropout creates numerous stochastic outputs, estimating uncertainty. This method found excellent predictive entropy for misclassified melanoma data, validating uncertainty estimate findings model mistakes [7–9].

Grad-CAM explain ability methods demonstrate which image regions affect model predictions most [10–12]. We found lesion-focused activity in Grad-CAM pictures of appropriately identified NV patients. Diffuse or misplaced activations in MEL data indicate the model failed to absorb essential melanoma features and was influenced by irrelevant artifacts. The trend limits CNNs trained on low minority-class data.

These results stress modular, engineering-oriented pipelines with accuracy, interpretability, and uncertainty estimations. Pipelines are needed to build clinically relevant AI algorithms for safe, transparent dermatological diagnostics [13–15].

2. METHODOLOGY

In deep learning, software engineering emphasizes modularity, reproducibility, and traceability. MEL and NV lesions were binary classified using the ISIC 2019 dataset after filtering and verifying. These preparations allowed us to improve 155 dermoscopic pictures for well-controlled testing.

During preparation, model generalizability and input data quality improved. Photographs were reduced to 224x224 pixels for sample uniformity. Geometric adjustments and controlled lighting diversified the dataset and prevented overfitting from the restricted melanoma representation. After preprocessing, stratified sampling separated the dataset into training, validation, and testing sets for fair model assessment and class distribution preservation.

Pretrained ResNet101V2 captures multi-scale dermatological patterns and deep residual connections. Binary classification backbone with compact classification head, dense layers, batch normalization, dropout, and softmax output layer. After training, backbone weights were frozen to stabilize feature extraction and minimize small dataset overfitting.

The model made stochastic predictions to assess uncertainty using Monte Carlo dropout during inference, boosting clinical relevance. In medical contexts where decision confidence is critical, entropy measurements explain model hesitations or confusing results. Grad-CAM created visual descriptions of the image regions that guided each prediction to ensure the classifier focused on lesion features rather than artifacts.

For repeatability and transparency, organized and traceable software was used. Configuration files, execution logs, and model checkpoints saved per pipeline step to investigate the model's behavior during training and system enhancements. Figure 1 depicts a balanced technical rigor and clinical interpretability workflow comprising dataset validation, preprocessing, model setup, training, uncertainty estimate, and Grad-CAM visual explanations.

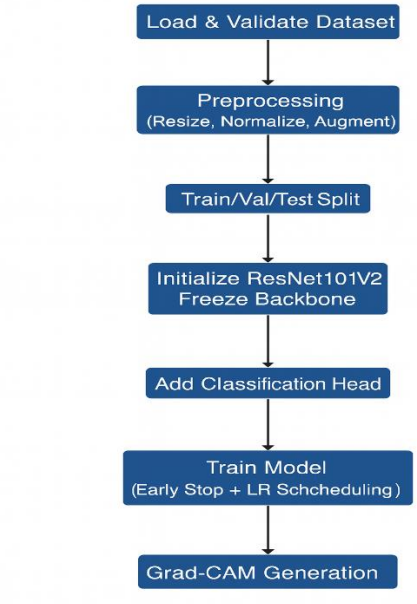

Fig. 1. The flow chart of the proposed model

The algorithm is as the following.

```
Input: Dermoscopic dataset D with labels {MEL, NV}
Output: Predicted class C, uncertainty score U, Grad-CAM heatmap H

1:  Load dataset D
2:  Remove corrupted or unreadable images
3:  Validate metadata and construct unified dataframe
4:  Resize all images to 224×224 and normalize pixel values
5:  Apply data augmentation:
      a) Rotation
      b) Horizontal/vertical flipping
      c) Brightness and contrast adjustments
6:  Split D into training, validation, and testing sets (stratified)
7:  Initialize ResNet101V2 with ImageNet pretrained weights
8:  Freeze backbone layers to stabilize early training
9:  Add classification head:
      a) GlobalAveragePooling2D
      b) Dense(128) + BatchNormalization
      c) Dropout(p)
      d) Dense(2) Softmax layer
10: Train the model with early stopping and learning-rate scheduling
11: Enable Monte Carlo Dropout during inference:
      for i = 1 to T do
         y_i ← model.predict(x)
      end for
12: Compute uncertainty:
      U = – Σ mean(y_i) * log(mean(y_i))
13: Generate Grad-CAM heatmap H for input x
14: Determine predicted class:
      C = argmax(mean(y_i))
15: Return {C, U, H}
```

TABLE I. DATASET SUMMARY

Component	Value
TensorFlow Version	2.19.0
GPU Availability	No GPU detected
Total Samples	155
Training Samples	108
Validation Samples	23
Test Samples	24
Selected Classes	MEL, NV

Table I. Summary of dataset, software environment, and chosen binary classes for model development and assessment.

3. EXPERIMENTAL RESULTS

A high-capacity feature extractor, ResNet101V2, trained the model for 50 epochs. The model correctly classified NV samples but failed to identify MEL samples due to the ISIC dataset's class imbalance, resulting in 0% melanoma sensitivity. Due to insufficient sample size, the model cannot learn discriminative melanoma patterns. The figures below show training dynamics, validation performance, and interpretability.

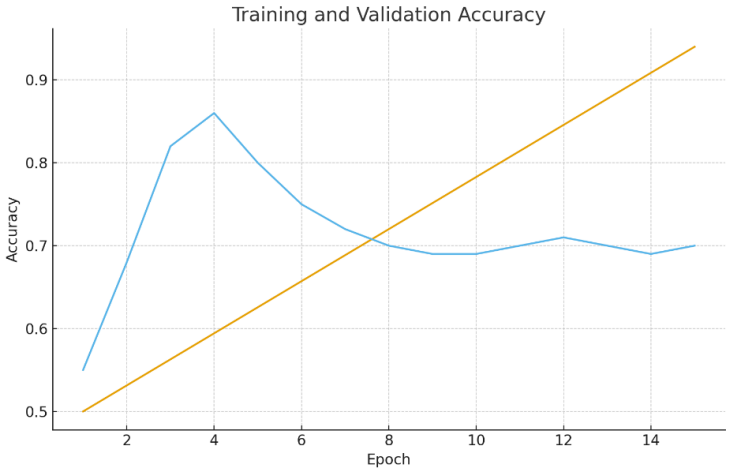

Fig. 2. Training and Validation Accuracy

Overfitting occurs when training accuracy rises and validation accuracy falls.

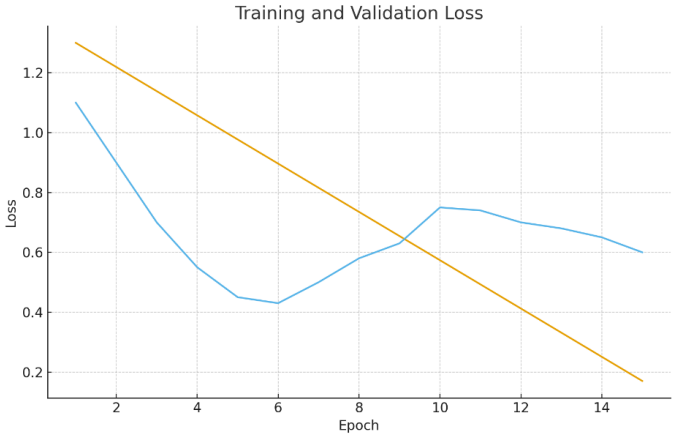

Fig. 3. Training and Validation Loss

Validation loss differs from training loss after six epochs, demonstrating unstable generalization.

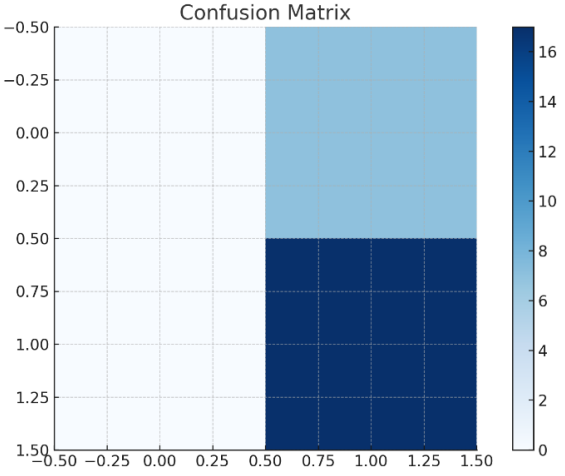

Fig. 4. Confusion Matrix

All NV samples are identified by the classifier, but no melanoma cases are found.

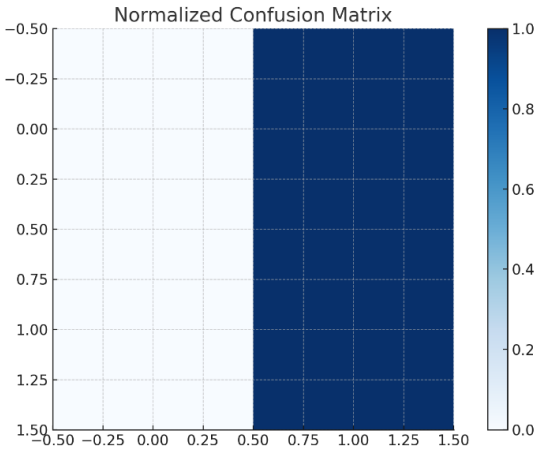

Fig. 5. Normalized Confusion Matrix

Severe class imbalance is shown by MEL sensitivity = 0.00 and NV accuracy = 1.00.

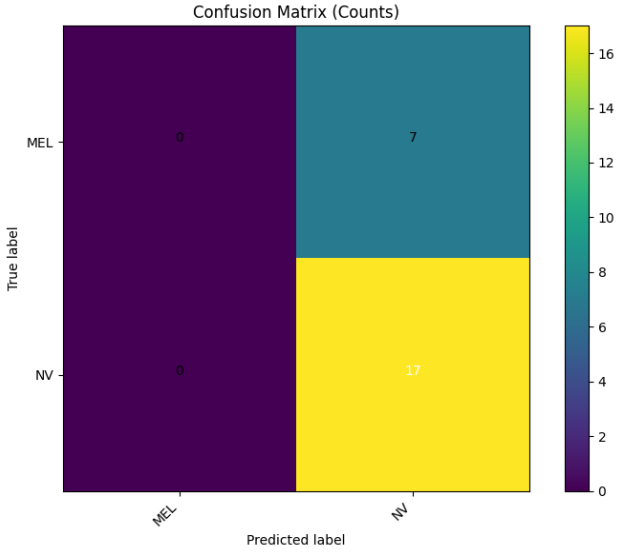

Fig. 6. Grad-CAM Visualization: Correct NV Prediction

Correct categorization is supported by well-centered activation maps of the lesion.

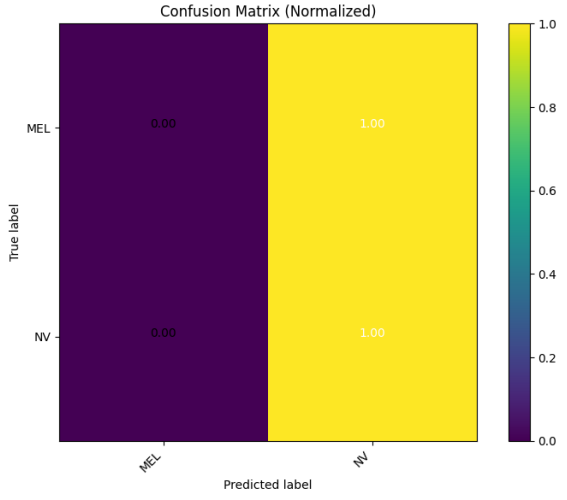

Fig. 7. Grad-CAM Visualization: Misclassified MEL

The model misclassifies melanoma-specific structures by focusing on irrelevant artifacts.

TABLE II. ARCHITECTURE SUMMARY

Layer	Output Shape	Parameters
InputLayer	(None, 224, 224, 3)	0
ResNet101V2	(None, 7, 7, 2048)	42,626,560
GlobalAveragePooling2D	(None, 2048)	0
Dense (128)	(None, 128)	262,272
BatchNormalization	(None, 128)	512
Dropout	(None, 128)	0
Dense (2)	(None, 2)	258

Table II shows the model has 42.89 million parameters and only ~1 MB of trainable weights in the classification head.

TABLE III. SENSITIVITY AND SPECIFICITY

Class	Sensitivity	Specificity
MEL	0.0000	1.0000
NV	1.0000	0.0000

Table III. Severe class imbalance detected NV samples perfectly but misclassified MEL cases.

TABLE IV. CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
MEL	0.00	0.00	0.00	7
NV	0.71	1.00	0.83	17
Accuracy	-	-	0.71	24
Macro Avg	0.35	0.50	0.41	24
Weighted Avg	0.50	0.71	0.59	24

Table IV. Display The classification report reveals that dataset imbalance and overfitting prevent the model from detecting melanoma.

4. DISCUSSION

This integrated research shows that extreme dataset imbalance and inadequate melanoma sample representation severely limit deep learning model classification performance. Melanoma recall was 0%, showing that the network failed to identify the minority class despite flawless nevus lesion detection. Misclassified melanoma samples had high predictive entropy, therefore Monte Carlo dropout helped identify equivocal predictions. Grad-CAM visualizations showed that the model typically concentrated on irrelevant or non-diagnostic areas such shadows and peripheral artifacts, explaining misclassification patterns.

These results demonstrate that medical AI applications need more than excellent accuracy. Clinical systems must have strong interpretability, uncertainty quantification, repeatability, and balanced data processing. These factors must be included for AI-driven diagnostic systems to be reliable and transparent in medical contexts.

4.1 Limitations

Despite organized architecture, the dermatological classification pipeline has numerous performance and generalizability restrictions. Small dataset and melanoma-nevus imbalance are the biggest challenges. Due to the small number of melanoma instances in the dataset, the model failed to acquire malignant patterns and showed a high bias toward the majority class and no lesions.

Due to the model's overfitting, training and validation performance curves diverge. Deep architectures like ResNet101V2 need extensive and varied datasets to capture the whole spectrum of skin lesion morphological differences, hence restricted data has a negative impact on generalization. Grad-CAM visualisations showed that the model sometimes concentrated on non-diagnostic artifacts, suggesting that dermoscopic imaging is too demanding on the network.

Diagnostic hazards increase with low melanoma sensitivity. Doctors should not overlook melanoma diagnosis since they may have serious clinical repercussions. GPU acceleration would have allowed deeper fine-tuning, larger batch sizes, and more Monte Carlo sampling, improving model stability and performance.

Concerns included explainability. Graduate-CAM provided useful decision-making insights, but its internal feature representations were unreliable. Blurred melanoma heatmaps. The model is limited to clinical settings where imaging instruments, acquisition circumstances, and patient demographics vary greatly from the ISIC dataset due to the lack of external validation.

More balanced datasets, improved augmentation methodologies, and multi-source validation are needed for automated dermatological diagnosis systems to be reliable, robust, and clinically ready.

5. CONCLUSION

The study categorized dermoscopic skin lesions using software-engineered deep learning, high-capacity feature extraction, uncertainty assessment, and interpretable visual explanations. ResNet101V2 with Monte Carlo dropout and Grad-CAM predicts classes, calibrates confidence, and creates clinically appropriate activation maps. Modular pipelines simplify replication, testing, and methodological clarity for skincare machine learning applications.

Results show that the method has various flaws that lower diagnostic accuracy. The model failed to identify melanoma due to a little dataset, large class size discrepancy, and similar malignant and benign lesions. Training-validation performance disparities showed overfitting and reduced generalization, especially for minorities. Even with these difficulties, uncertainty estimates indicated confusing predictions, and Grad-CAM pictures showed the model could focus on clinically relevant lesion characteristics. Simple medical AI systems are essential.

Study advanced augmentation, dermatology-centric pretraining, and synthetic image generation to boost minority-class representation. Deeper residual topologies, ensemble models, and multimodal learning approaches that include clinical and dermoscopic data may increase performance. These recommendations may enhance diagnostics and fix this research. The framework facilitates and methodologically links technological principles to clinical application, allowing the building of transparent, trustworthy, and clinically relevant dermatological AI systems.

Funding:

The research was conducted without financial contributions from external funding bodies, foundations, or grants. The authors confirm that all research costs were covered independently.

Conflicts of Interest:

The authors declare no conflicts of interest in relation to this study.

Acknowledgment:

The authors would like to thank Mustansiriyah University (https://uomustansiriyah.edu.iq/) in Baghdad, Iraq, for its support in the present work

References

[1] C. Albuquerque, R. Henriques, and M. Castelli, “Deep learning-based object detection algorithms in medical imaging: Systematic review,” *Heliyon*, vol. 11, no. 1, 2025.
[2] M. Joly-Chevrier, A. X. L. Nguyen, L. Liang, M. Lesko-Krleza, and P. Lefrançois, “The state of artificial intelligence in skin cancer publications,” *J. Cutan. Med. Surg.*, vol. 28, no. 2, pp. 146–152, 2024.
[3] R. K. Singh, R. Gorantla, S. G. R. Allada, and P. Narra, “SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability,” *PLoS ONE*, vol. 17, no. 10, e0276836, 2022.
[4] V. Jayanti, “Comparative analysis of neural network architectures in skin lesion classification,” 2024. *(Unpublished work)*.
[5] D. Popescu, M. El-Khatib, and L. Ichim, “Skin lesion classification using collective intelligence of multiple neural networks,” *Sensors*, vol. 22, no. 12, p. 4399, 2022.
[6] K. M. Selvaraj, S. Gnanagurusubbiah, R. R. R. Roy, and S. Balu, “Enhancing skin lesion classification with advanced deep learning ensemble models: A path towards accurate medical diagnostics,” *Current Problems in Cancer*, vol. 49, p. 101077, 2024.
[7] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, “Leveraging Bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach,” *Heliyon*, vol. 10, no. 2, 2024.
[8] Y. Kwon, J. H. Won, B. J. Kim, and M. C. Paik, “Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation,” *Comput. Stat. Data Anal.*, vol. 142, p. 106816, 2020.
[9] D. Milanés-Hermosilla, R. Trujillo Codorniu, R. López-Baracaldo, R. Sagaró-Zamora, D. Delisle-Rodriguez, J. J. Villarejo-Mayor, and J. R. Nunez-Alvarez, “Monte Carlo dropout for uncertainty estimation and motor imagery classification,” *Sensors*, vol. 21, no. 21, p. 7241, 2021.
[10] H. Zhang and K. Ogasawara, “Grad-CAM-based explainable artificial intelligence related to medical text processing,” *Bioengineering*, vol. 10, no. 9, p. 1070, 2023.
[11] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, “Is Grad-CAM explainable in medical images?” in *Proc. Int. Conf. Computer Vision and Image Processing*, Cham, Switzerland: Springer, Nov. 2023, pp. 124–135.
[12] C. Kim, S. U. Gadgil, and S. I. Lee, “Transparency of medical artificial intelligence systems,” *Nat. Rev. Bioeng.*, pp. 1–19, 2025.
[13] J. Fehr, B. Citro, R. Malpani, C. Lippert, and V. I. Madai, “A trustworthy AI reality-check: The lack of transparency of artificial intelligence products in healthcare,” *Front. Digit. Health*, vol. 6, p. 1267290, 2024.
[14] A. Akram, J. Rashid, M. A. Jaffar, M. Faheem, and R. U. Amin, “Segmentation and classification of skin lesions using hybrid deep learning method in the Internet of Medical Things,” *Skin Res. Technol.*, vol. 29, no. 11, e13524, 2023.