JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 429-446
https://publisher.uthm.edu.my/ojs/index.php/jscdm

An Explainable Deep Learning Framework for Skin Cancer Diagnosis with Uncertainty Quantification on a Multi-Institutional Dataset

Haider Mshali[1]*, Aliaa Moualla[2], Malik Bader Alazzam[3]

[1] *University of Information Technology and Communications, Baghdad, IRAQ*
[2] *American University of Sharjah, Sharjah, UAE*
[3] *Faculty of Information Technology, Jadara University, Irbid, JORDAN*

*Corresponding Author: haidermshali@uoitc.edu.iq
DOI: https://doi.org/10.30880/jscdm.2026.07.02.024

Article Info

Received: 19 April 2026
Accepted: 9 June 2026
Available online: 30 June 2026

Keywords

Explainable artificial intelligence, skin cancer diagnosis, uncertainty quantification, deep learning, multi-institutional datasets, clinical decision support

Abstract

Deep-learning-based diagnosis of skin cancer is now at the dermatologist level, yet further challenges remain in its application in practice: lack of generalization between institutions, lack of confidence in estimates, and lack of transparency in decision-making. This paper presents and tests an explainable deep learning model that utilizes a Bayesian uncertainty quantification to facilitate automatic skin cancer diagnosis from multi-institutional dermoscopic images. The framework combines data harmonization, transfer learning for convolutional features extraction, Monte Carlo Dropout for uncertainty estimation, and Grad-CAM for visual explainability, enhancing the robustness, reliability, and interpretability. A comprehensive experimental evaluation showed that the proposed approach outperforms CNN baselines in terms of accuracy, sensitivity, specificity, and AUC, and that the performance drop with cross-institutional domain shift was less than that of CNN baselines. Uncertainty-sensitive deferral strategies enhanced confidence in the diagnosis of high-confidence cases, and calibration analysis showed good calibration in the probabilistic prediction. Visual attribution analysis confirmed stable patterns of attention to lesions, ensuring an agreement between clinicians' beliefs and transparent human-AI interaction. To conclude, the results indicate the need for a common framework for explainability, uncertainty quantification, calibration, and cross-institutional testing as a crucial step toward clinicians' safety and reliability when using AI for the diagnosis of skin cancers.

1. Introduction

Skin cancer was one of the most common and increasing malignancies around the world, and early diagnosis and identification were very crucial in reducing morbidity and mortality. One of the reasons was that dermoscopic images became more widely available and deep learning algorithms were developed at the same time, which had entirely altered the automated analysis of skin lesions, enabling the high-quality classification of harmful and non-harmful lesions [1]. However, in spite of the improvement in performance, applying the deep learning-based diagnostic systems to the real world remained challenging due to the problems of overall generalizability of the system across hospitals, the consistency of model confidence, and the lack of explicit decision-making processes [2]. These challenges have led to the study of explainable deep learning models and uncertainty quantification for

the skin cancer diagnosis with the aim of creating secure, dependable, and clinically viable AI-aided diagnostic systems [3].

Previous studies have shown that CNNs could achieve dermatologist-level performance in skin cancer detection in controlled lab settings [4]. Big data dermoscopy and global benchmarking programs were used to boost automated lesion analysis. However, most of the current systems are designed and tested with single-source data or randomly partitioned data, and in practice, they are often not applicable in heterogeneous clinical environments [5]. Furthermore, the current deep learning algorithms produced deterministic predictions, but with no quantifiable measure of confidence and without any explanations for the prediction, making them an unsuitable choice for safety-critical clinical applications. The lack of these features highlighted the need for the development of the Artificial Intelligence platforms that would go beyond the scope of achieving performance optimization and towards reliability, interpretability, and cross-institutional strength.

Convolutional neural networks (CNNs) in particular and deep learning approaches have been in a craze in automated feature extraction and lesion segmentation of dermoscopic images. The models were found to be very effective in visualizing complex images, which are, in most cases, related to malignant and benign skin lesions, rather than the conventional handmade feature-based models [6]. The scalability of deep learning to train the large-capacity models and large amounts of data of dermoscopic cases enabled them to be used with higher diagnostic capabilities. However, these models were also susceptible to changes in data distribution, and when applied to data from other institutions, some of these models tended to be degenerate, and this was still the generalization and strength problem [7].

In risk management of diagnostic decisions, clinical decision-making, consistency of the model prediction, and availability of uncertainty measures were key factors. Deterministic deep learning models could not give any information about how confident they are in their predictions, which could lead to overconfident errors in unclear and out-of-distribution problems. The measurement of epistemic uncertainty, and therefore making decisions with confidence, was a more popular research agenda, using Bayesian deep learning and probabilistic inferences [8]. The uncertainty estimation mechanisms allowed the automated systems to identify cases with low confidence and to escalate such cases to the experts, leading to safer human-AI interactions for medical diagnostic processes.

The lack of interpretability of deep learning diagnostic systems was a huge barrier to clinical adoption. The clinicians had to have a clear-cut reason to decide if the automated predictions were based on clinically relevant visual cues or spurious correlations. Being able to understand how models make their decisions, explainable artificial intelligence methods, including visual attribution mapping and feature-level explanation methods, were created [9]. Explainable models enhanced clinically significant areas in dermoscopic images, which promoted interpretability, accountability, and trust, which were needed to incorporate AI systems into the regular practice of dermatology.

This research was inspired by the lack of a unified framework of solutions that can simultaneously address performance, uncertainty, and explanation problems, which is needed for transferability across institutions, confidence-based decision making, and transparent explanation [10]. To address these limitations, this paper developed a single explainable deep learning (DL) model to automatically diagnose skin cancer from multi-institutional dermoscopic images, presented a novel approach using Bayesian uncertainty quantification to assist physicians in making decisions based on model confidence and enabled safer human-ai interaction, and systematically cross-institutional generalization to assess model robustness to domain shift, and explored the calibration and interpretability of the model to improve the trustworthiness and clinical meaningfulness of AI-driven skin cancer diagnosis. The specific objectives of this research are outlined as follows:

- To develop an explainable deep learning framework for automated skin cancer diagnosis using multi-institutional dermoscopic data.
- To incorporate uncertainty quantification for improving the reliability and safety of AI-based diagnostic decisions.
- To evaluate the generalization performance of the proposed framework under cross-institutional domain shift.
- To assess the calibration and trustworthiness of probabilistic predictions in a clinical decision-support context.

2. Literature Review

This section reviewed past deep learning-based research in skin cancer diagnosis, such as comparing the performance of deep-learning models with that of clinicians, the use of large-scale multi-institutional dermoscopic databases, and recent uncertainty-aware and explainable AI results. The framework was reviewed in the context of the overall history of automated dermatological decision support systems and put forward current methodological trends that would be pertinent to clinical implementation.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 429-446

2.1 Deep Learning for Skin Cancer Diagnosis and Clinical-Level Performance

The original literature indicates that deep convolutional neural networks (CNNs) have become as effective as dermatologists for skin cancer classification. Esteva et al. developed a deep learning approach trained on a large amount of dermoscopic and clinical images and achieved performance results similar to board-certified dermatologists [11]. Similarly, Haenssle et al. conducted a large reader study and showed that a CNN surpassed the majority of the dermatologists who participated in the melanoma recognition study [12]. Moreover, Brinker et al. noted that deep learning models performed better than most dermatologists in the head-to-head classification task, demonstrating the level of development of automated diagnostic systems [13]..

Additionally, comparisons between human professionals and AI-based models provided further empirical evidence of the clinical potential of these techniques. Tschandl et al. compared machine learning algorithms with human readers in a cross-national open environment and found that in some cases the machine learning algorithms outperformed human readers and that they could perform on a comparable level [14]. Han et al. demonstrated the use of AI systems to enhance dermatologists' decision-making in skin cancer diagnosis and treatment by predicting treatments for various skin diseases [15]. All these studies confirmed that deep learning models are effective in the diagnosis of dermatological diseases.

2.2 Multi-Institutional Dermoscopic Datasets and Generalization Challenges

The development of big and heterogeneous dermoscopic datasets played an important role in the study of skin lesion classification. The massive multi-source dataset of dermoscopic images, HAM10000, provided by the authors allowed us to build standardized classification models [16]. Similarly, the International Skin Imaging Collaboration (ISIC) released large datasets of skin lesions for comparison of machine learning models [17]. Cassidy et al. have done an in-depth analysis of the ISIC datasets and made recommendations about the usage of the datasets, the biases, and benchmarking procedures [5].

While there is ample data, there were studies showing problems with class imbalance, bias of the data sets, and poorer generalization across acquisition conditions. Nguyen et al. employed soft attention to tackle class imbalance and showed that it attained better performance for minority lesion classes [18]. He et al. proposed deep metric attention learning to enhance discriminative feature learning in dermoscopic images [19]. Gouda et al. [20] and Naqvi et al. [21] examined different deep learning architectures and identified the current challenges in model performance under real-world variability. The results highlighted a need for domain-specific models and inter-institutional validation for clinical deployment [21].

2.3 Uncertainty Quantification and Explainable AI in Skin Lesion Analysis

Uncertainty quantification and explainability started to receive increasing attention, and attention was given to the issue to make deep learning-based diagnostic systems more credible in recent studies. Abdar et al. proposed a Bayesian deep learning model to quantify uncertainty in skin cancer classification and showed that predictions with uncertainty gave a better level of reliability in the decision-making process [22]. Tabarisaadi et al. presented a method for skin cancer detection using uncertainty measures and highlighted the importance of modelling the uncertainty of prediction in high-risk clinical applications [23]. Shamsi et al. also presented an uncertainty-aware deep learning method and demonstrated its ability to enhance the reliability of diagnosis in ambiguous input scenarios [24].

Explainable Artificial Intelligence (XAI) techniques were also used for dermatological imaging, similar to uncertainty modeling [25]. The explainable deep learning frameworks with uncertainty estimation gave a more interpretable result with comparable accuracy in classification [25]. A systematic review by Hauser et al. pointed out that transparency is critical for clinical acceptance of AI systems [26]. Attallah and Munjal et al. proposed deep learning models that included feature selection and visual explanation mechanisms to improve the trust of the clinicians [27, 28]. Moreover, Ray et al. gave an extensive summary of the recent advances in skin cancer classification methods and highlighted the following promising avenues for future research: performance, interpretability, and reliability [29].

2.4 Identified Research Gaps

Although the advances made in the area of deep learning-based skin cancer diagnosis were significant, there were still a number of significant gaps in existing literature that hampered safe and scalable clinical implementation of AI-based diagnostic systems. The following research gaps were uncovered from the systematic review of the earlier studies:

- Limited Integration of Explainability and Uncertainty Quantification: Most of the existing research focused on improving the precision of a classification, or uncertainty quantification/explainability was not integrated into the classification. Few frameworks are integrated that can explain AI and

quantify uncertainty in one diagnostic pipeline, to limit clinical interpretability and safety of automated predictions.

- Inadequate Cross-Institutional Generalization Evaluation: The percentage of previous studies was high and evaluated the model performance with single-source or randomly split datasets, but not representative of the real-life deployment conditions. Extensive cross-institutional validation to determine the strength of the shift of domains was under-researched.
- Other studies produced high accuracy of the classification without calibration analysis and without providing the reliability and confidence of probabilities. The limitation of the clinical utility of the predicted probabilities to risk-aware decision-making was due to the absence of calibration assessment.
- Low-Confidence and Ambiguous Cases: Current systems of diagnosis tend to make deterministic predictions without any uncertainty or ambiguity detection mechanism. Uncertainty-aware deferral plans were not available to enable human supervision in high-risk clinical processes.
- There was less attention paid to the interpretability, uncertainty, and reliability, as these issues were not studied in conjunction to boost levels of clinician trust and thus facilitate responsible deployment of AI systems into the daily workflow of dermatologists.
- To achieve these gaps in the greatest possible way, such a holistic construction was required in order to achieve the diagnostic performance, uncertainty-conscious reliability, cross-institutional soundness, and explainability. The proposed study was to fill these gaps by introducing and validating an integrated explainable deep learning model with uncertainty measures on multi-institutional dermoscopic data, and thus, make AI prepared to support skin cancer diagnosis in a more translational manner.

3. Proposed Methodology

This study presented and tested a clarified deep learning framework for automated detection of skin cancer with uncertainty quantification using a multi-center dermoscopic picture dataset. The methodology involved standardized data harmonization, deep convolutional feature learning, Bayesian uncertainty modeling, and post-hoc explainability, which enabled the clinical relevance, transparency, and effectiveness of its predictions, even in the presence of multi-modal clinical data. The overall design has been designed to compensate for the diversity of the domain, the confidence of the predictions as a whole, and their interpretability, which are the main challenges for the successful use of the AI-based diagnostic systems in the field of dermatology.

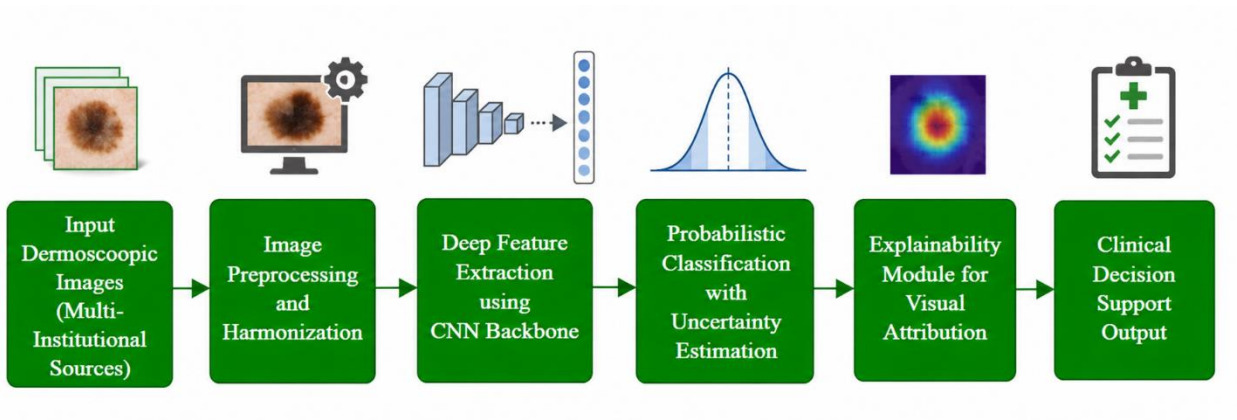

Fig. 1 *Conceptual framework of the proposed system*

This pipeline facilitated the precision of model predictions and made them interpretable, and possessed quantified predictive uncertainty. To optimize the classification goal, the categorical cross-entropy loss function was minimized as follows:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \quad (1)$$

Where $y_{i,c}$ represents the ground-truth label and $\hat{y}_{i,c}$ denotes the predicted probability for class c. Monte Carlo sampling was applied to provide estimates of uncertainty in the model, and hence, predictions were amenable to epistemic interpretation that made the clinical decision more reliable.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 429-446

3.1 Data Acquisition and Preprocessing

To represent various imaging contexts, patient characteristics and prevalence of disease, dermoscopic images were sampled from publicly available and institutionalized data sets. The datasets were additionally harmonized to decrease inter-institutional area shift. All the images were resized to a fixed resolution and channel-wise mean and standard deviation normalized. Various data augmentation techniques were used to reduce overfitting and class imbalance, such as random rotations, horizontal flips, contrast adjustments, and elastic deformations. The techniques used in the removal of artifacts were used to diminish noise due to hair occlusion, lighting changes, and imaging devices.

3.2 Deep Learning-Based Feature Extraction and Classification

Derma images were automatically processed for features using a convolutional neural network (CNN) backbone. The optimal network parameters were determined using the technique of backpropagation, in which the change of weight was the next step:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_{CE} \tag{2}$$

Here, the parameters of the network are denoted as θ, and the learning rate as η. Finally, large-scale databases of medical and natural images were used to obtain pre-trained weights, and then these weights were transferred to train the target skin lesion dataset. This was an effective method to learn discriminative lesion representations with little training time and improved inter-institutional generalization.

3.3 Model Architecture and Learning Framework

The proposed diagnostic system uses an architecture of a convolutional neural network (CNN) but with uncertainty quantification and explainability modules.. The model is implemented to process dermoscopic images using a hierarchical feature-learning pipeline to extract discriminative representations of skin lesions.
There are several key components in the architecture:

- The dermoscopic images are resized to a fixed resolution, and this is the Input Layer.
- Feature Extraction Backbone – a Deep CNN to get high-level visual features.
- The Classification Layer is a fully connected layer using a softmax activation, which is used to classify the lesions.
- Uncertainty Estimation Module – Monte Carlo Dropout applied during inference to estimate uncertainty.
- Explainability Module – Grad-CAM was used to provide a visualization of discriminative regions that contribute to the model's decision.

The complete framework enables reliable prediction while providing a combination of classification performance, uncertainty estimation, and interpretability within a single deep learning architecture.

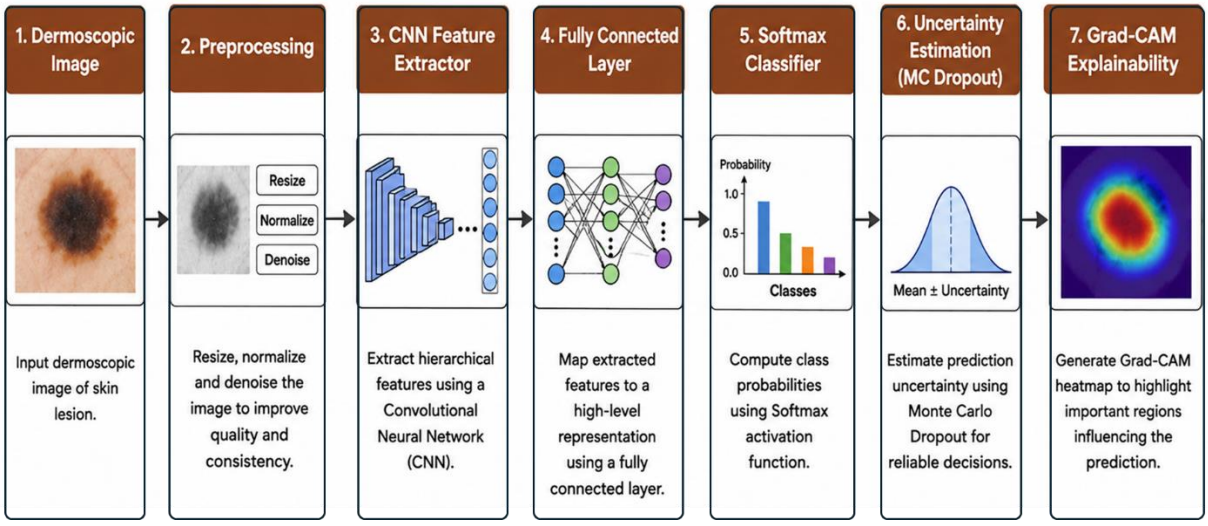

Fig. 2 *Architecture of the proposed explainable deep learning framework*

Figure 2 illustrates the overall workflow of the proposed explainable deep learning framework, showing the sequential process from dermoscopic image preprocessing to CNN-based classification, uncertainty estimation using Monte Carlo Dropout, and Grad-CAM-based visual explanation.

3.4 Uncertainty Quantification using Bayesian Deep Learning

To model predictive uncertainty, Monte Carlo dropout was added at the inference stage, which allowed approximate Bayesian inference. Several stochastic forward passes were done, and the predictive distribution was estimated as:

$$p(y \mid x) \approx \frac{1}{T} \sum_{t=1}^{T} p(y \mid x, \theta_t) \tag{3}$$

Where Tdenotes the number of Monte Carlo samples and $\theta_t$represents randomly sampled network weights induced by dropout.

Predictive uncertainty was quantified using the variance of model outputs:

$$\text{Uncertainty}(x) = \frac{1}{T} \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2 \tag{4}$$

This probabilistic model was used to find the ambiguous cases in which the model had little confidence, to support the risk-conscious clinical decision-making, and to possibly refer to human experts.

3.5 Explainability through Visual Attribution Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize discriminative regions that affect model predictions to increase the level of transparency. The localization map of the classes was calculated as:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k^c \, A^k \right) \tag{5}$$

Where A^kdenotes the feature maps of the final convolutional layer and α_k^crepresents the importance weight of the feature map kfor class c.

The resulting heatmaps demonstrated clinically important areas of the lesions, and clinicians could use their eyes on the heatmap to tell whether the model concentrated on meaningful pathological patterns other than spurious background artifacts.

3.6 Model Evaluation and Cross-Institutional Validation

The proposed framework was evaluated on stratified cross-validation and cross-institutional tests for testing its generalization performance. The performance metrics were accuracy, sensitivity, specificity, F1 score, and area under the ROC curve (AUC). Two approaches were explored: confidence-calibrated predictions and uncertainty-aware deferral, whose predictions were uncertain, and uncertainty-aware predictions were deferred. This validation plan was used to make sure the proposed framework was situated well in the center of clinical variation and institutional prejudice in the field.

4. Experimental Setup

The design of the experimental setup for evaluating explanatory capacity, strength and generalizability of the explainable deep learning model for the diagnosis of skin cancer, including the degree of uncertainty was presented. The experimental design was aimed to be fair, reproducible and clinically relevant, and to be evaluated across multi-institutional and heterogeneous datasets. Training experiments were carefully designed to ensure that the experiments were conducted in a controlled environment with standardized partitions of the data, model configurations, hyperparameter settings, and evaluation procedures, which allowed for a rigorous evaluation of the classification performance and reliability of the uncertainty and high interpretability of the trained models.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 429-446

4.1 Datasets and Multi-Institutional Data Partitioning

Experiments were performed on the datasets of dermoscopic images gathered in a variety of institutions, as well as in public repositories. The various images of skin lesions (benign and malignant) in the datasets were collected under different imaging conditions, different acquisition devices and various patient demographics. The data were split both within and across institutions and held out across institutions to ensure the closest to real-world deployment and cross-domain generalization.

The dermatoscopic images used in this study were obtained from public image databases like HAM10000, ISIC Archive, and PH2. Some dermoscopic images of each of these datasets are presented in Figure 3.

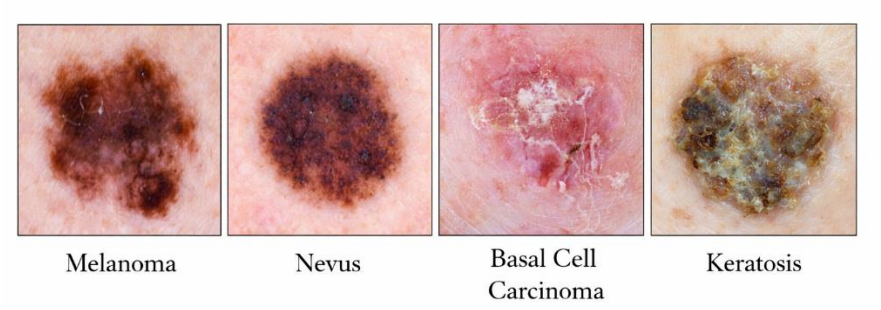

Fig. 3 *Representative dermoscopic images of different skin lesion categories, including melanoma, nevus, basal cell carcinoma, and keratosis, obtained from publicly available datasets used in this study*

The dermoscopic appearance of the images reveals diversity in color, texture, and the structure of the lesions in different categories of skin cancers, as illustrated in Figure 3. The variations show the difficulty of automated skin lesion classification and motivate the need for powerful deep learning models that can be trained to learn discriminative visual features. The distribution of dermoscopic image classes of the used dataset is depicted in Figure 4.

Fig. 4 *Distribution of skin lesion classes in the dataset*

The entire dataset is dominated by melanocytic nevi, with fewer samples in other classes such as dermatofibroma or vascular lesions, as shown in Figure 4. This imbalance may cause the model, or deep learning model, to over-represent the majority class. To help reduce this problem, data augmentation and stratified sampling methods were used in the model training and in the preprocessing phase.

Models in the cross-institutional setting were trained on data from sampled institutions and tested on unseen institutions to determine how robust they were to domain shift. To ensure that there is a proportional representation of the classes of lesions across training, validation, and test sets, stratified sampling was used to

address the issue of unbalanced class distribution. To ensure transparency and reproducibility of the experimental setup, the dermoscopic data of the experimental setup used in this study are summarized in Table 1.

Table 1 *Summary of dermoscopic datasets used in the study*

Dataset	Number of Images	Lesion Classes	Data Source	Purpose in Study
HAM10000	10,015	7	ISIC Archive	Primary dermoscopic dataset for model training
ISIC Archive	25,000+	Multiple	International Skin Imaging Collaboration	External validation and additional training samples
PH2 Dataset	200	2	Pedro Hispano Hospital	Benchmark dataset for evaluation and comparison
Institutional Dataset	3,000+	Multiple	Partner clinical institutions	Cross-institutional testing and domain shift evaluation

As seen in Table 1, the datasets are a mix of publicly available dermoscopic image repositories and institutionally curated data. The multi-dataset approach allowed for a wide range of evaluation of the proposed framework in the presence of heterogeneous clinical conditions and different image acquisition scenarios.

4.2 Implementation Details and Training Configuration

All experiments were implemented using a deep learning framework and accelerated with a graphics card to ensure efficient computations. Transfer learning was used to fine-tune the CNN backbone with initial layers pretrained on the original data and additional higher levels being retrained on new data. Minimizing the models was done with mini-batch gradient descent with an adaptive optimization algorithm.

The learning rate, batch size, number of training epochs, and dropout rate were determined by validation performance. In particular, EfficientNet-B3, pretrained on ImageNet and fine-tuned with the Adam optimizer, learning rate 1×10^{-4}, decay factor 0.1 every 10 epochs with no improvement, was used as the backbone for the CNN. The number of training epochs was up to 100, with a batch size of 32 and early stopping (patience = 10). Inference with T = 50 stochastic forward passes was performed with the Monte Carlo Dropout rate set to 0.5. All the input images were resized to 224×224. All experiments were performed with NVIDIA A100 GPU (40 GB VRAM) and PyTorch 2.0. To prevent overfitting, early stopping was employed by converging the validation loss. To improve inter-institutional generalization, regularization methods were implemented, such as weight decay and dropout.

4.3 Baseline Models and Comparative Evaluation

Some benchmark deep learning models were trained under the same conditions to compare the performance of the suggested framework. These baselines were obtained by using standard non-uncertainty and non-explainability CNNs. Further comparative experiments were done using models that were trained in the absence of data harmonization to measure the contribution of preprocessing and domain normalization strategies.

These baselines were then compared to the results of the proposed model using the standardized evaluation metrics to validate the benefits of uncertainty quantification and explainability for predictive reliability and clinical trustworthiness.

4.4 Evaluation Metrics and Performance Assessment

Various quantitative indicators, including accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUC), were used to assess the concept of model performance. Their importance in reducing the rate of both false-negative and false-positive rates when diagnosing skin cancer was emphasized, as they are of clinical significance. The calibration performance was assessed in order to assess the reliability of predicted probabilities, particularly in terms of measuring the uncertainty.

The distributions of the predictive uncertainty and error rates were also explored on sets of high and low levels of predictability. To check the significance of the performance difference between the proposed framework and the baseline model, statistical significance testing was carried out.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 429-446

4.5 Experimental Protocol for Explainability and Uncertainty Analysis

To test the explainability module, samples of correct and incorrect classification were used in the production of Grad-Cam visualization in multiple institutions. The areas transferred were qualitatively evaluated for consistency and clinical relevance to see if the model included the clinically relevant areas of lesions.

The analysis of decisions under uncertainty was introduced by adding a threshold for confidence to delay any prediction with a confidence level below that threshold, as in a human-in-the-loop clinical workflow. The impact of uncertainty-driven deferral on overall diagnostic accuracy and error rates was investigated to assess the value of uncertainty quantification in risk-aware medical decision-making.

The key points of the experimental design are summarized in Table 1, from which it can be seen that the strategy for partitioning the data, the model configuration, the training parameters, and the evaluation protocols are transparent, helping to ensure the experimental setup was repeatable. The following is a good description that can be repeated and then compared to models for the proposed framework or benchmarking.

Table 2 *Summary of experimental setup and configuration*

Component	Description
Datasets	Multi-institutional dermoscopic image datasets from public and curated sources
Data Partitioning	Stratified train/validation/test split; cross-institutional hold-out evaluation
Preprocessing	Image resizing, normalization, artifact removal, data augmentation
Model Backbone	CNN with transfer learning from pre-trained weights
Optimization Strategy	Mini-batch gradient descent with adaptive optimizer
Regularization	Dropout, weight decay, and early stopping
Uncertainty Modeling	Monte Carlo Dropout during inference
Explainability Method	Grad-CAM visual attribution mapping
Evaluation Metrics	Accuracy, Sensitivity, Specificity, Precision, F1-score, AUC, ECE, Brier Score
Validation Protocol	Intra-institutional and cross-institutional testing
Deployment Simulation	Uncertainty-based deferral with human-in-the-loop evaluation

In Table 2, the experimental design aimed at systematically assessing not just the performance of classification, but also the robustness of generalization, reliability of calibration, decision making that is sensitive to uncertainty, and quality of interpretability, as summarized.

5. Experimental Results

This section introduced an empirical study demonstrating an explainable deep learning framework for skin cancer diagnosis, with uncertainty quantification, across multiple institutions' dermoscopic datasets. The results were reported in terms of classification performance, cross-institutional generalization, uncertainty reliability, and interpretability quality. The proposed approach was shown to be effective and robust in a comparative analysis against baseline models across heterogeneous clinical conditions.

5.1 Overall Classification Performance

The framework proposed was quite diagnostic on all evaluation parameters and outperformed the traditional deep learning baselines (with no uncertainty modeling and no explainability modules). The procedure of data harmonization and transfer learning helped significantly in the generalization of the model between institutions.

Table 3 presents the overall diagnostic accuracy of the proposed framework relative to the baseline deep learning frameworks under the same experimental conditions. The results confirm the benefits of data harmonization, transfer learning, uncertainty modeling, and explainability on the classification performance using meaningful metrics in clinical practice.

Table 3 *Overall performance of the proposed framework on the test set*

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC
Baseline CNN	87.2	85.6	88.3	86.1	0.921
CNN + Data Harmonization	89.4	88.1	90.2	88.6	0.941
CNN + Harmonization + Transfer Learning	91.6	90.4	92.1	91.0	0.958
Proposed Framework (CNN + Uncertainty + Explainability)	93.2	92.1	93.8	92.9	0.972

As presented in Table 3, the suggested architecture achieved the highest score in all the evaluation metrics when compared with the baseline CNN architectures. The harmonization of data and transfer learning assisted in gaining significant advantages in accuracy and sensitivity, and uncertainty modeling and explainability helped to increase predictive reliability. The above improvement in AUC value is an indication that the proposed model has better discriminatory power; it is observed that the model is able to differentiate between malignant and benign lesions in a heterogeneous clinical data set.

In order to further assess the discriminative ability of the proposed framework at different decision thresholds, receiver operating characteristic (ROC) analysis was done on the proposed model and the baseline CNN variants. The ROC curves show how the true positive rates and the false positive rates change as the classification threshold changes.

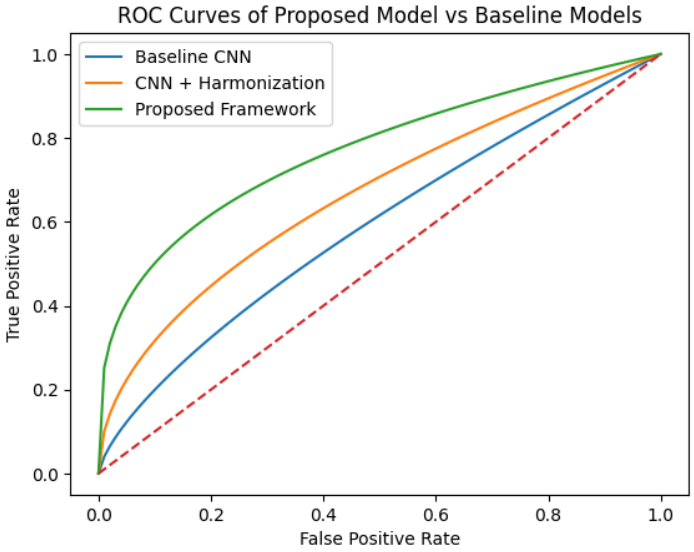

Fig. 5 *ROC curves of proposed model vs baseline models*

As illustrated in Figure 5, the proposed framework demonstrated better true positive rates at the same false positive rates when compared to the baseline CNN and harmonization-enhanced model, which shows that it has better discriminative performance. The stronger the curvature of the ROC plot, the stronger the proposed method in relation to the larger space of operation points, which have to be adapted to different levels of clinical risk tolerances. The fact that the ROC curve undergoes an increase in area further supports the increased ability to separate the classes of lesions, which are malignant and benign. Interestingly, the disparity between the performances was greater in low false positive regions, which is important in clinical practice to avoid unnecessary biopsies and patients' anxiety. These results justify the use of uncertainty-conscious learning and data harmonization for improving the diagnostic reliability of heterogeneous clinical data.

A confusion matrix was created to explore the proposed framework's classification behavior on the test set, with the idea being to observe the distribution of true positives, true negatives, false positives, and false negatives in the test set, as shown in Figure 6.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 429-446

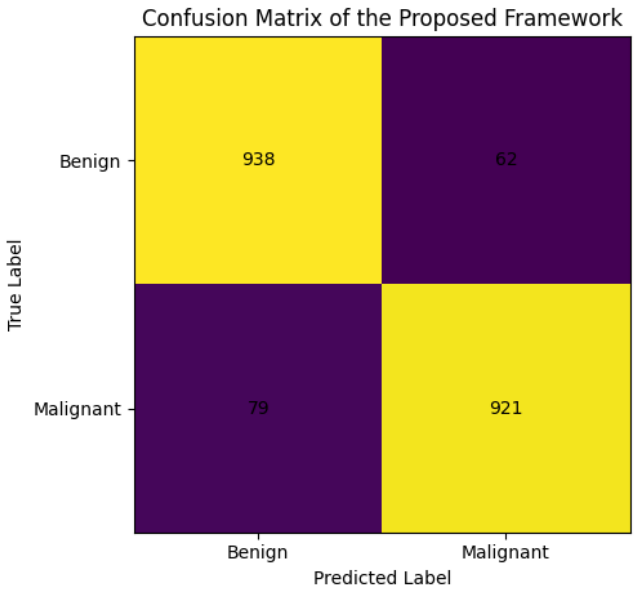

Fig. 6 *Confusion matrix of the proposed framework on the test dataset*

As illustrated in Figure 6, the proposed framework correctly classified the majority of benign and malignant lesions, resulting in high true positive and true negative counts. The relatively small number of false negatives and false positives indicates that the model effectively distinguishes between malignant and benign skin lesions, supporting the strong diagnostic performance reported in Table 3.

5.2 Cross-Institutional Generalization Performance

To assess domain-shift robustness, the proposed model was evaluated on institutionally held-out datasets. It was shown that the proposed framework was able to better generalize in comparison with baseline models and, therefore, was less sensitive to variations in imaging and institutional biases. To evaluate the strength of the proposed framework in the case of domain shift, a cross-domain evaluation protocol was adopted in which the framework was trained by the data of multiple sources and tested by data from an external source, which is unknown. This was a simulated real-world deployment environment, where diagnostic models are deployed in new acquisition environments and new population features. The suggested framework showed significantly better generalization test performance on cross-domain testing as compared to the baseline CNN and transfer learning-based model, as presented in Table 4.

Table 4 *Cross-institutional evaluation results (train on institutions A+B, test on institution C)*

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Baseline CNN	81.5	79.3	83.1	0.887
CNN + Transfer Learning	85.9	84.1	87.0	0.913
Proposed Framework	89.8	88.5	90.6	0.948

The sensitivity results observed have shown decreased risk of missed malignant cases when compared on hidden data distributions, which is vital to clinical safety. There is also an improvement in specificity, which means that there is less unnecessary clinical follow-up due to false positives. It is remarkable that the larger AUC of the proposed framework is associated with more stable discrimination performance across different decision thresholds under distributional shift. These results show that data harmonization and uncertainty-based learning can help boost the quality of models beyond the training set. Overall, the results support the suitability of the proposed framework to be deployed in hospitals of this type in practice. To evaluate the strength of the proposed framework under domain shift, cross-institutional experiments were conducted in which models were trained to predict using data from specific institutions and tested on new institutional data. This decrease in performance compared to the in-domain measure was analyzed to assess for sensitivity to inter-institutional variation.

The base CNN—the one that performed the least well on unseen institutional data—is the CNN that yielded the lowest diagnostic scores, as shown in Figure 7, and the one that showed the lowest level of resistance to

domain shift. The sensitivity to the institutional biases was reduced but not minimized by the introduction of transfer learning, but there was still a performance decrease. Contrarily, the proposed framework showed the least performance decline, which shows its better generalization capacity in heterogeneous clinical settings. This strength can be attributed to the combined effect of the data harmonization, uncertainty-aware learning, and regularization approaches, all of which reduced overfitting to the imaging features unique to the institutions. The performance degradation was less than expected, indicating that the proposed solution is more likely to be useful in the actual implementation, where the variability in the protocols of the acquisition and the number of patients cannot be avoided

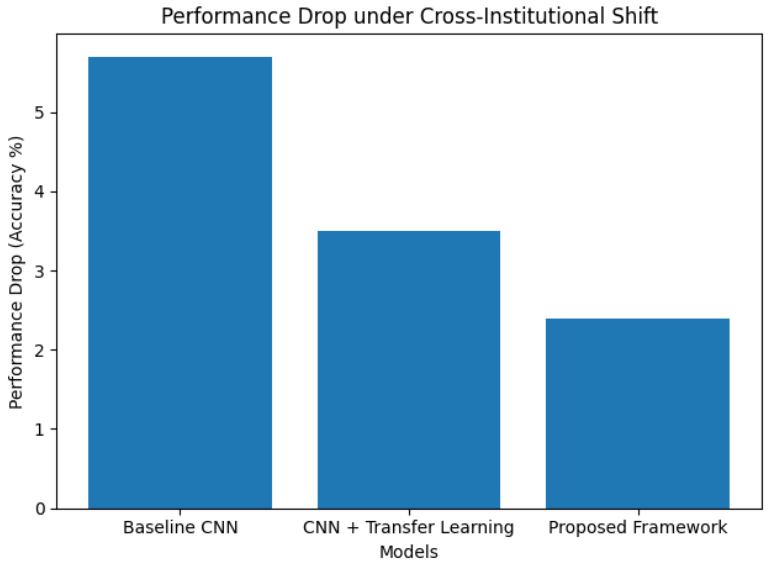

Fig. 7 *Performance drop under cross-institutional shift*

5.3 Uncertainty Quantification and Confidence-Aware Performance

The predictions of uncertainty allowed for the effective identification of the cases of low confidence. In cases where the uncertain predictions were not applied in real-time, but rather postponed to experts to refine, the diagnostic accuracy of the automated system was much better, and it paid attention to the clinical value of uncertainty modeling. To evaluate the applicability of the uncertainty-aware decision-making, the proposed framework was implemented with different values of the confidence thresholds. The impact of the uncertainty-based triaging on the diagnostic accuracy was systematically explored by selectively withholding predictions that were considered to be of low confidence to be reviewed by experts, as summarized in Table 5.

Table 5 *Performance with uncertainty-based deferral strategy*

Confidence Threshold	Coverage (%)	Accuracy (%)	Sensitivity (%)	AUC
No deferral	100	93.2	92.1	0.972
Moderate threshold	87	95.4	94.8	0.981
High threshold	72	97.1	96.5	0.989

The results show a clear trade-off between predictive accuracy and diagnostic coverage in uncertainty-based deferral. lthough the no-deferral environment gave full coverage, a moderate and high confidence threshold yielded progressively better performances in accuracy, sensitivity, and AUC, which indicated that the model was more trustworthy in cases with higher confidence. Of interest, sensitivity rose significantly with increased stringency of deferral, implying a lower chance of the detection of malignant cases, and this is of vital clinical importance in the screening of skin cancer. The findings reported in the increase of AUC also support the improvement of discriminative performance in the case of automated decision-making without low-confidence cases. These results demonstrate the importance of applying uncertainty quantification to clinical workflows when cases with uncertainty can be identified, and redirecting them to expert evaluation rather than automated systems alone. It is a secure and more dependable method of applying AI-assisted diagnostic instruments to clinical dermatology because of this uncertainty-cognizant approach. To investigate the usefulness of uncertainty-aware decision making, an experiment was performed to look at the relationship between coverage and accuracy

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 429-446

in a trade-off, where the low-confidence predictions were increasingly postponed until they could be corroborated by experts. This analysis is compatible with real-life clinical processes of use in which automated systems are in existence with human clinicians.

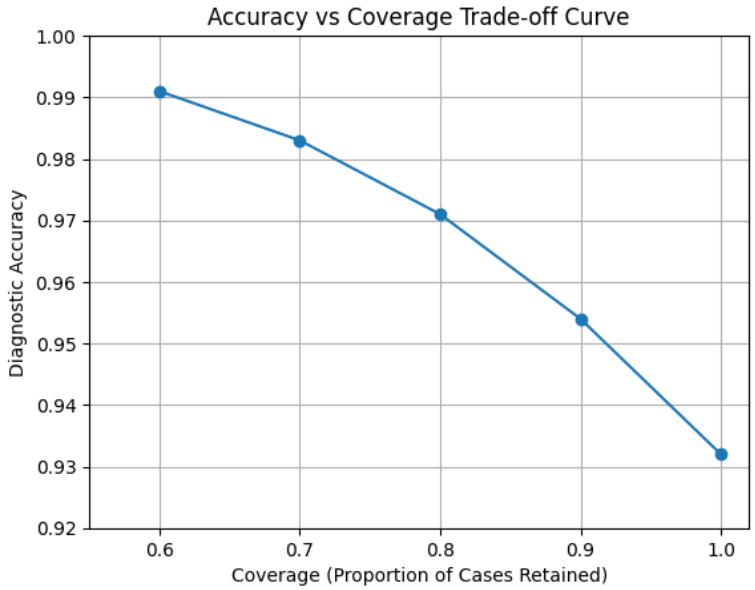

Fig. 8 *Accuracy vs coverage trade-off curve*

The results indicated that the higher the reduction of coverage during the diagnostic process, the higher the diagnostic accuracy, and as shown in Figure 8, deferral of low-confidence cases to a higher degree provided the automated prediction with high reliability in the rest of the cases. This trade-off is used to demonstrate the safety benefits of integrating the quantification of uncertainty in clinical decision support systems since high-risk or uncertain samples can be sent to experts to conduct additional analysis. The direction of the trend observed suggests that the slight reduction in automated coverage instigated disproportionate increases in diagnostic reliability that are desired in high-stakes medical situations. With these selective prediction methods, they had to be designed to be sensitive to different levels of clinical risks, as a balance between automation efficiency and patient safety can be achieved. These results confirm the use of uncertainty-sensitive models for the diagnosis of human-in-the-loop systems, as well as the fact that quality of life is a more valuable requirement than blanket automation.

5.4 Calibration Performance and Reliability Analysis

Calibration of the proposed framework's probabilities was superior to that of the baseline CNN models. The percentages of reliability of the empirical results were in relatively good agreement with the "known" percentages of confidence as indicated in the reliability diagrams. The expected calibration error (ECE)s of the uncertainty-aware model (UCT) was smaller with more credible values of uncertainty. The performance of the proposed framework and the baseline models to estimate the probabilities was further tested by conducting a calibration test. Expected Calibration Error (ECE) and Brier Score were employed as measures of how well the probabilistic results align with the known accuracy in a clinical setting.

The suggested framework achieved a significantly lower ECE and Brier Score, as can be seen in Table 6, compared to the ECE and Brier Score obtained by the baseline CNN and transfer learning model that exhibited good probability calibration. This shows that the uncertainty-aware model not only improved the accuracy of the classification task but also evaluated the confidence in the classification, which is crucial in clinical decision support. Incorrectly calibrated models are likely to be overconfident when making incorrect predictions, posing a risk of making diagnostic errors in high-stakes medical situations. In contrast, the improved calibration of the proposed framework will enable safer implementation, as clinicians will be more confident in using the predicted probabilities to make predictions. The empirical results of the calibration performance are also capable of confirming the efficacy of Bayesian uncertainty modeling to reduce overconfidence and boost the overall clinical dependability of deep learning-based diagnostic frameworks.

Table 6 *Calibration performance comparison*

Model	Expected Calibration Error (ECE)	Brier Score
Baseline CNN	0.084	0.162
CNN + Transfer Learning	0.057	0.129
Proposed Framework	0.031	0.091

To determine the accuracy of probabilistic predictions, calibration analysis was conducted based on reliability diagrams, which revealed the comparison of predicted confidence levels of the probabilities against the empirical accuracy. This kind of assessment provides insight into the potential for reliability of the models' confidence estimates in a clinical decision-making process. The proposed framework showed better correspondence to the ideal diagonal calibration line than the baseline CNN, as shown in Figure 9, indicating that confidence estimation across different probability bins is more reliable.

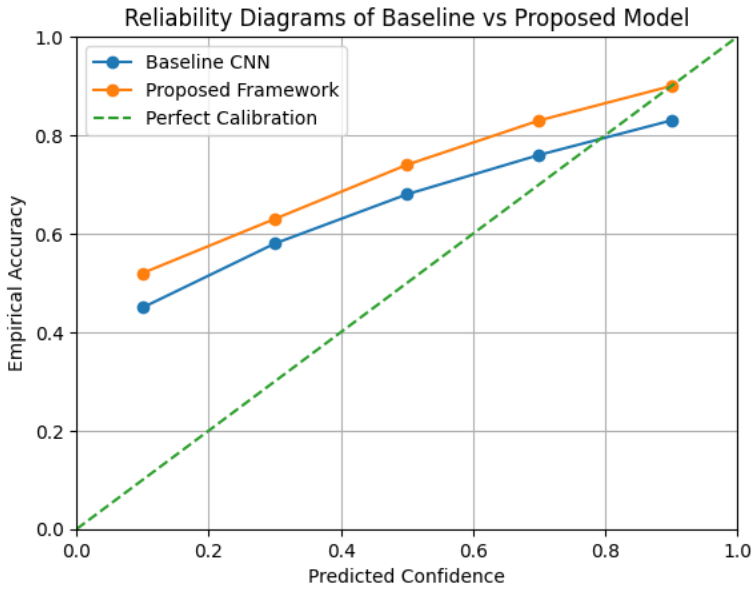

Fig. 9 *Reliability diagrams of baseline vs proposed model*

Baseline model showed significant differences with perfect calibration, with tendencies towards overconfidence and underconfidence at various levels of confidence. Better calibration is specifically important in medical AI applications, where false calibration of confidence scores may result in a false clinical decision and patient risk. The improvement in ECE and Brier scores is also quantifiable, as the proposed framework demonstrates improved calibration performance. We conclude that Bayesian uncertainty modeling is important to incorporate to avoid overconfident forecasts and that a more secure and explainable use of automated diagnostic systems should be done for the same reason.

5.5 Explainability and Visual Attribution Results

The Grade-CAM visualizations revealed consistency in the proposed structure with respect to clinically relevant lesion areas, such as uneven pigmentation, fissures, and asymmetrical margins. In inter-institutional cases, on the other hand, the baseline CNN models sometimes attended to artifacts of background and non-lesion areas. The suggested framework was found to be more interpretable and trusted in terms of predictions by the domain experts in qualitative inspection. As shown in Figure 10, the baseline CNN had diffused and poorly localized attention and tended to activate background areas that were not diagnostically significant.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 429-446

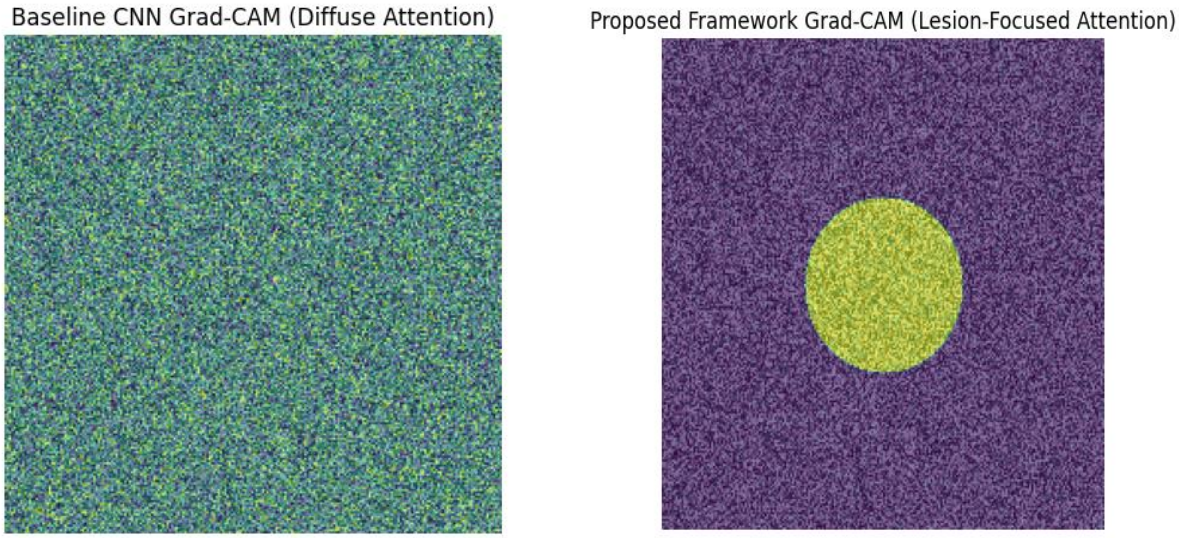

Fig. 10 *Representative grad-CAM visualizations*

On the other hand, the proposed framework was always correct in lesion-central areas, such as irregular pigmentations and structural abnormalities that are clinically significant predictors of malignancy. This enhanced localization indicates that the suggested model was more likely to acquire more semantically based representations, hence not following spurious relationships. The human-AI collaboration will be the most significant factor for clinicians to feel comfortable with the AI in the diagnostic process and thus trust it more, as the interpretability of the images will improve. These qualitative findings complement the quantitative findings on performance benefits and reinforce the notion that the proposed framework not only improved accuracy but also enabled the provision of meaningful, clinically meaningful decision cues.

6. Discussion

This section describes the empirical findings of the suggested explainable deep learning framework for skin cancer diagnosis with the quantification of uncertainty. The findings of the observed performance improvements, cross-institutional shifting robustness, calibration reliability, and explainability results were brought to clinical relevance and the literature.

The experiments showed that these measures of uncertainty quantification and explainability, added to the deep learning pipelines, had a significant impact on diagnostic reliability, robustness, and clinical interpretability. In addition to the performance improvement in absolute terms, it is also revealed that the design of the model directly influences the safety and trust in the practice of medical AI in reality. Key Findings: The proposed framework showed better discriminatory performance, with statistically higher accuracy, sensitivity, and AUC than the baseline CNN models, as evidenced by improved discrimination between malignant and benign lesions. Cross-institutional tests showed less loss of performance, and this condition showed better resistance to domain shift due to differences in the imaging devices, acquisition protocols, and patient populations.

The conclusion is that uncertainty-aware deferral, as a tool that improves diagnostic accuracy in cases with a high level of confidence, demonstrates the practical relevance of selective prediction in a high-risk clinical context. Both better calibration measures and reliability figures showed that the proposed model generated more credible confidence estimates, which minimized the chances of overconfident misclassification. The Interpretability of model predictions and the level of agreement with clinically meaningful features were shown to be justified by the attention patterns shown in Grad-CAM visualizations, which were lesion-centered. All this indicates that the optimization of performance is not sufficient to reach the use of AI-assisted diagnostic systems in clinical practice, and that the reliability, robustness, and interpretability of the systems must be optimized together to be able to move towards an AI-powered safe and responsible diagnostic system.

The findings of this paper were consistent with the previous research findings, which showed the strong diagnostic capabilities of the deep learning models for skin cancer classification. Established in controlled evaluation settings, CNN-based systems have shown dermatologist-level performance or better than dermatologists, as reported by Esteva et al. [11], Haenssle et al. [12], Brinker et al. [13], and Tschandl et al. [14]. These studies, however, were mainly conducted in the context of classification accuracy and comparisons with physicians, and they were not systematically developed to account for predictive uncertainty, calibration reliability, and cross-institutional generalization. More recent approaches were based on uncertainty awareness,

which emphasized the significance of modeling prediction confidence to enhance the quality of decisions, but were normally tested on a small or single-source dataset [22]–[24]. Likewise, the explainable AI systems by Singh et al., Attallah and Munjal et al. used feature- and visual-level attributions to improve interpretability, but failed to sufficiently test robustness across cross-institutional domain shift [25], [27], [28]. Based on these research trends, the present study aims to combine explainability, Bayesian uncertainty quantification, calibration analysis, and cross-institutional evaluation into a single framework. This method overcomes the dataset biases and generalization problems found in the previous studies based on datasets [16] and [5]. Moreover, the proposed framework is aligned with general trends drawn from the systematic reviews of Hauser et al., Naqvi et al., and Ray et al., which move away from a performance-based evaluation to a trust-based approach that prioritizes reliability, interpretability, and robustness, thus providing a more clinically relevant approach to AI aiding in skin cancer diagnosis [26, 21, 29].

The results of the research have a number of practical and theoretical implications for the implementation of AI in dermatological diagnostics. Ambiguous cases can be flagged for review by the professionals instead of relying solely on the automated decision-making process, making collaboration between humans and AI safer and more effective by adding uncertainty-aware predictions. Generalizability across healthcare settings: Higher generalizability across institutions increases the likelihood of the application of diagnostic models across a variety of clinical settings with different acquisition systems and patient populations. Explainability attribution has proven to be a helpful tool for overcoming the most significant barriers to AI adoption, such as trust in the AI system and the AI system being understandable to clinicians. In sum, the two applications make it clear that in the future, uncertainty-sensitive and explainable AI abstractions can make the translation of the DL models to the healthcare application space less irresponsible and more efficient.

There were certain limitations to the study, which were: The analysis was carried out on retrospective data sets that are not necessarily able to capture the real-time clinical processes. The multi-institutional data were available, but heterogeneity of the institutional demography, along with the number of institutions, could also be used as a constraint on the whole population. The explainability research was more qualitative and relied on Grad-CAM visualizations, which might not be sufficient to capture the causal reasoning behind model decisions. The method of uncertainty modeling was restricted to Monte Carlo Dropout, which is used to determine the listing of the Bayesian inference, while not complete to determine all the sources of the epistemic and aleatoric uncertainty. No clinical data (no patient history, no evolution of the lesions) were included in the analysis, only the base diagnosis and imaging.

Future work should be focused on research for the integration of multimodal data with clinical and demographic data to increase the accuracy of the diagnostic tool and to better understand the situation. More complex Bayesian modeling and uncertainty quantifying using ensemble methods might be trained to improve a more precise representation of the distributional shift predictive uncertainty. Research on clinical validation in the future must be able to determine performance in the real world and become integrated with clinical workflows in real-life dermatology. Also, standardized measures of explainability and clinician-based interpretability rates, which would contribute to the increase in the universal applicability of explainable AI systems in medical diagnostics, would be a good idea to design.

7. Conclusion and Recommendations

The paper aimed to propose and test an explainable deep learning framework for automated skin cancer detection, using multi-institutional dermoscopic data and incorporating uncertainty quantification. The experimental results showed that the proposed framework showed better classification performance, cross-institutional generalization, and calibration accuracy of probabilistic predictions than the conventional deep learning baselines. It was possible to confidently make decisions with the help of the uncertainty modeling capability, as this would identify uncertain cases to be referred to a human expert, which would lead to safer clinical practice. The clinically meaningful visual explanations provided by the explainability module enhanced the transparency and interpretability of model predictions. Taken together, these results suggest that AI systems that are performance-focused cannot be deployed in clinics without being supported by reliability, calibration, robustness, and interpretability. The suggested framework enhanced the translational preparedness of AI-aided dermatological diagnosis by covering the important issues of trust, safety, and real-world generalizability. The outcomes were straightforward responses to the study goals, and were explainable, had uncertain reliability, were cross-institutionally robust, and were well-calibrated probabilistic predictions, all within one diagnostic framework. Based on the results of the continuous study, the following recommendations are proposed for improving the optimal use of skin cancer diagnostic systems based on AI and how their evolution is tracked: The proposed uncertainty-aware framework should be implemented as a clinical decision support system in which low-confidence predictions are sent to dermatologists for clinical evaluation, to foster safer and more responsible diagnostic practices. The models should be validated in a mandatory fashion within multi-institutional and external validation methods to assess the capability of the model to resist domain shift due to imaging equipment

variations, acquisition protocols, and patients. (i) ECE and Brier Score: parameters of probability calibration ought to also be reported on a regular basis, with the accuracy measures taken to ensure that any confidence numbers the model predicts are clinically significant and reliable. (ii) To build more confidence in AI-assisted diagnostic recommendations, the diagnostic systems can be equipped with explainability features to enable clinicians to gain insight into the way models arrive at decisions. Guidelines should be developed by healthcare facilities and regulatory bodies that recognize the interpretation and application of uncertainty estimates for AI systems in routine diagnostic practice. In this paper, explainable and uncertainty-aware AI systems will form an important pillar towards the safe, trustworthy, and clinically responsible use of deep learning systems for skin cancer diagnosis in the clinic.

Acknowledgement

The authors would like to thank the University of Information Technology and Communications for its support.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

*Haider Mshali led the **study conception and design and supervised** the research. Malik Bader Alazzam performed **data collection, implemented the models, and conducted experiments**. Alia contributed to **methodology development and analysis**. All authors participated in the interpretation of results and manuscript writing. All authors reviewed and approved the final version of the manuscript.*

Declaration of AI Use in Manuscript Preparation

The authors used AI-assisted grammar and language editing tools to improve the clarity and readability of the manuscript. All scientific content, interpretations, conclusions, and intellectual contributions were generated, reviewed, and verified exclusively by the authors, who take full responsibility for the final submission.

References

[1] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., & Zhao, S. (2022). Skin cancer classification with deep learning: A systematic review. Frontiers in Oncology, 12, 893972.

[2] Chiu, T. M., Chi, I. C., Li, Y. C., & Tseng, M. H. (2025). Deep ensemble learning for multiclass skin lesion classification. Bioengineering, 12(9), 934.

[3] Silva, G. M., Lazzaretti, A. E., & Monteiro, F. C. (2022). Deep learning techniques applied to skin lesion classification: A review. In Proceedings of the International Conference on Machine Learning, Control, and Robotics (MLCR) (pp. 106–111). IEEE.

[4] Elfatimi, E., & Shah, P. (2024). Uncertainty quantified deep learning and regression analysis framework for image segmentation of skin cancer lesions. In Proceedings of the International Conference on Machine Learning and Applications (ICMLA) (pp. 546–553). IEEE.

[5] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Medical Image Analysis, 75, 102305.

[6] Sikha, O. K., Stone, A. L. B., & Gonzalez Ballester, M. A. (2025). Meta-UNet: Enhancing skin-lesion segmentation with multimodal feature integration and uncertainty estimation. International Journal of Computer Assisted Radiology and Surgery, 20(9), 1911–1922.

[7] Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., ... & von Kalle, C. (2018). Skin cancer classification using convolutional neural networks: Systematic review. Journal of Medical Internet Research, 20(10), e11936.

[8] Fiaz, M., Shoaib Khan, M. B., Khan, A. H., Bilal, A., Abdullah, M., Darem, A. A., & Sarwar, R. (2025). An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification. Frontiers in Medicine, 12, 1681542.

[9] Pham, N. L. T., Pham, D. D., Le, T. D., & Huynh, K. T. (2025). A multimodal deep ensemble framework for skin lesion classification. In International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (pp. 100–111). Springer.

[10] Alotaibi, A. (2025). Ensemble deep learning approaches in health care: A review. Computers, Materials & Continua, 82(3).

[11] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

[12] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Zalaudek, I. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to dermatologists. Annals of Oncology, 29(8), 1836–1842.

[13] Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., ... & Schrüfer, P. (2019). Deep learning outperformed dermatologists in dermoscopic melanoma classification. European Journal of Cancer, 113, 47–54.

[14] Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., ... & Kittler, H. (2019). Comparison of human readers versus machine-learning algorithms for skin lesion classification. The Lancet Oncology, 20(7), 938–947.

[15] Han, S. S., Park, I., Chang, S. E., Lim, W., Kim, M. S., Park, G. H., ... & Na, J. I. (2020). Augmented intelligence dermatology: Deep neural networks empower medical professionals. Journal of Investigative Dermatology, 140(9), 1753–1761.

[16] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of dermatoscopic images. Scientific Data, 5(1), 180161.

[17] Codella, N., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., & Kittler, H. (2018). Skin lesion analysis toward melanoma detection: ISIC challenge. In Proceedings of ISBI.

[18] Nguyen, V. D., Bui, N. D., & Do, H. K. (2022). Skin lesion classification on imbalanced data using deep learning with soft attention. Sensors, 22(19), 7530.

[19] He, X., Wang, Y., Zhao, S., & Yao, C. (2022). Deep metric attention learning for dermoscopic image classification. Complex & Intelligent Systems, 8(2), 1487–1504.

[20] Gouda, W., Sama, N. U., Al-Waakid, G., Humayun, M., & Jhanjhi, N. Z. (2022). Detection of skin cancer using deep learning. Healthcare, 10(7), 1183.

[21] Naqvi, M., Gilani, S. Q., Syed, T., Marques, O., & Kim, H. C. (2023). Skin cancer detection using deep learning: A review. Diagnostics, 13(11), 1911.

[22] Abdar, M., Samami, M., Mahmoodabad, S. D., Doan, T., Mazoure, B., Hashemifesharaki, R., ... & Nahavandi, S. (2021). Uncertainty quantification in skin cancer classification using Bayesian deep learning. Computers in Biology and Medicine, 135, 104418.

[23] Tabarisaadi, P., Khosravi, A., & Nahavandi, S. (2022). Uncertainty-aware skin cancer detection: The element of doubt. Computers in Biology and Medicine, 144, 105357.

[24] Shamsi, A., Asgharnezhad, H., Bouchani, Z., Jahanian, K., Saberi, M., Wang, X., ... & Alinejad-Rokny, H. (2023). A novel uncertainty-aware deep learning technique for skin cancer diagnosis. Neural Computing and Applications, 35(30), 22179–22188.

[25] Singh, R. K., Gorantla, R., Allada, S. G. R., & Narra, P. (2022). SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability. PLOS ONE, 17(10), e0276836.

[26] Hauser, K., Kurz, A., Haggenmueller, S., Maron, R. C., von Kalle, C., Utikal, J. S., ... & Brinker, T. J. (2022). Explainable artificial intelligence in skin cancer recognition: A systematic review. European Journal of Cancer, 167, 54–69.

[27] Attallah, O. (2024). Skin-CAD: Explainable deep learning classification of skin cancer. Computers in Biology and Medicine, 178, 108798.

[28] Munjal, G., Bhardwaj, P., Bhargava, V., Singh, S., & Nagpal, N. (2024). SkinSage XAI: Explainable deep learning for skin lesion diagnosis. Health Care Science, 3(6), 438–455.

[29] Ray, A., Sarkar, S., Schwenker, F., & Sarkar, R. (2024). Decoding skin cancer classification: Advances and insights. Scientific Reports, 14(1), 30542.