RESEARCH Open Access

Two-step Hierarchical Classification of Skin Lesion Using Convolutional Neural Network and Random Forest Classifier

Taofik Ahmed Suleiman[123], Daniel Tweneboah Anyimadu[123], Andrew Dwi Permana[123], Hsham Abdalgny Abdalwhab Ngim[123], Alessandra Scotto di Freca[1]

Corresponding author:
Email:
[1]Department of Electrical and Information Engineering, MAIA, University of Cassino and Southern Lazio, Italy
[2]Department of Medical Imaging and Computing, University of Girona, Spain
[3]Graduate School of Science and Technology, Medical Imaging and Applications (MAIA), University of Burgundy, France

Abstract
Skin lesion classification plays a crucial role in early detection and diagnosis of various skin conditions. In this study, we propose a two-step hierarchical binary classification approach using convolutional neural networks (CNN) and random forest (RF) algorithms. Our methodology aims to overcome challenges associated with the class imbalance and variability in skin lesions. We conducted experiments on the International Skin Imaging Collaboration (ISIC 2017) dataset and evaluated the performance of both the machine learning and deep learning modules. The results demonstrate the effectiveness of the hierarchical approach, with the random forest classifier achieving an overall balanced multiclass accuracy of 87.51% and the CNN achieving 84.33%. Following these results, we also implemented a predictive model that integrates the two-step hierarchical binary classification process which combines the outcomes of the individual steps to obtain the final classification. Overall, the proposed two-step hierarchical classification approach has the potential of providing valuable insights for improving diagnostic outcomes, aid in treatment decisions, and contribute to the development of advanced diagnostic systems for skin cancer.

Keywords: machine learning, random forest classifier, CNN, two-step hierarchical classification, skin lesions

1.0 Introduction

Skin cancer is an excessive growth characterized by the uncontrolled proliferation of abnormal skin cells, often triggered by DNA damage caused by ultraviolet (UV) radiation exposure [1]-[3]. Skin cancer can be broadly categorized into three main types: basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma [4]. Each type has distinct pathophysiological characteristics and varying degrees of severity. However, with significant advances in radiation therapy, chemotherapy, immunotherapy, and targeted medicines, various treatment modalities have been employed [5]-[9]. Despite these progress in treatment options, skin cancer remains a global health concern due to its increasing incidence and associated morbidity and mortality rates [10]-[12].

The severity of skin cancer varies depending on the type and stage of the lesion. For example, melanoma can rapidly progress and metastasize if left untreated or poorly managed, which could result in life-threatening complications such as disfigurement, functional impairment, and a decreased quality of life [17]-[19]. Thus, accurately diagnosing skin lesions is essential for timely, effective, and appropriate treatment of the disease [20]-[22]. However, diagnosing skin lesions poses significant challenges because reading the dermoscopy images relies on expert visual analysis, which has drawbacks such as inter-observer variability and subjective interpretation [23]-[24]. Moreover, the high-resolution and heterogeneity of skin lesions, as well as the presence of clutter (e.g., hairs), further complicate the diagnosis process using dermoscopy. Thus, the need for an advanced computer-aided diagnosis technique for timely intervention and improved patient outcomes.

Big data and AI technology such as machine learning (ML) and deep learning (DL) techniques have been employed in numerous applications including disease diagnosis, prediction, and treatment optimization. In the field of dermatology, these techniques have been used for skin lesion classification, melanoma detection, and skin disease diagnosis. Several ML and DL algorithms such as support vector machines (SVM), k-nearest neighbors (KNN), decision trees, Naïve Bayes, and artificial neural networks have been employed for several classification

tasks including skin lesions and proved effective [30], [31]. Random forests and convolutional neural networks (CNNs) are other algorithms that have gained prominence in recent years due to their ability to handle high-dimensional data [32].

In this work, we used CNN and RFC algorithms to tackle the challenge of classifying three distinct types of skin lesions: benign, melanoma, and seborrheic keratosis. Our approach is based on a 2-step hierarchical binary classification because it offers a unique solution to address the class imbalance and the inherent variability in skin lesions.

2.0 Material and Methodology

This section presents the materials used in the project, the source of the skin lesion image dataset, as well as the methodologies used to achieve the two-step hierarchical classification. The algorithms adopted are CNN for deep learning and Random Forest Classifiers for machine learning. For a better explanation, the method is divided into two modules: the machine learning module and the deep learning module. Additionally, we developed a predictive model to test the efficiency of the project in real-life scenarios which is also discussed in this project. Figure 1 shows the flow diagram of the algorithms used in the implementation of the project.

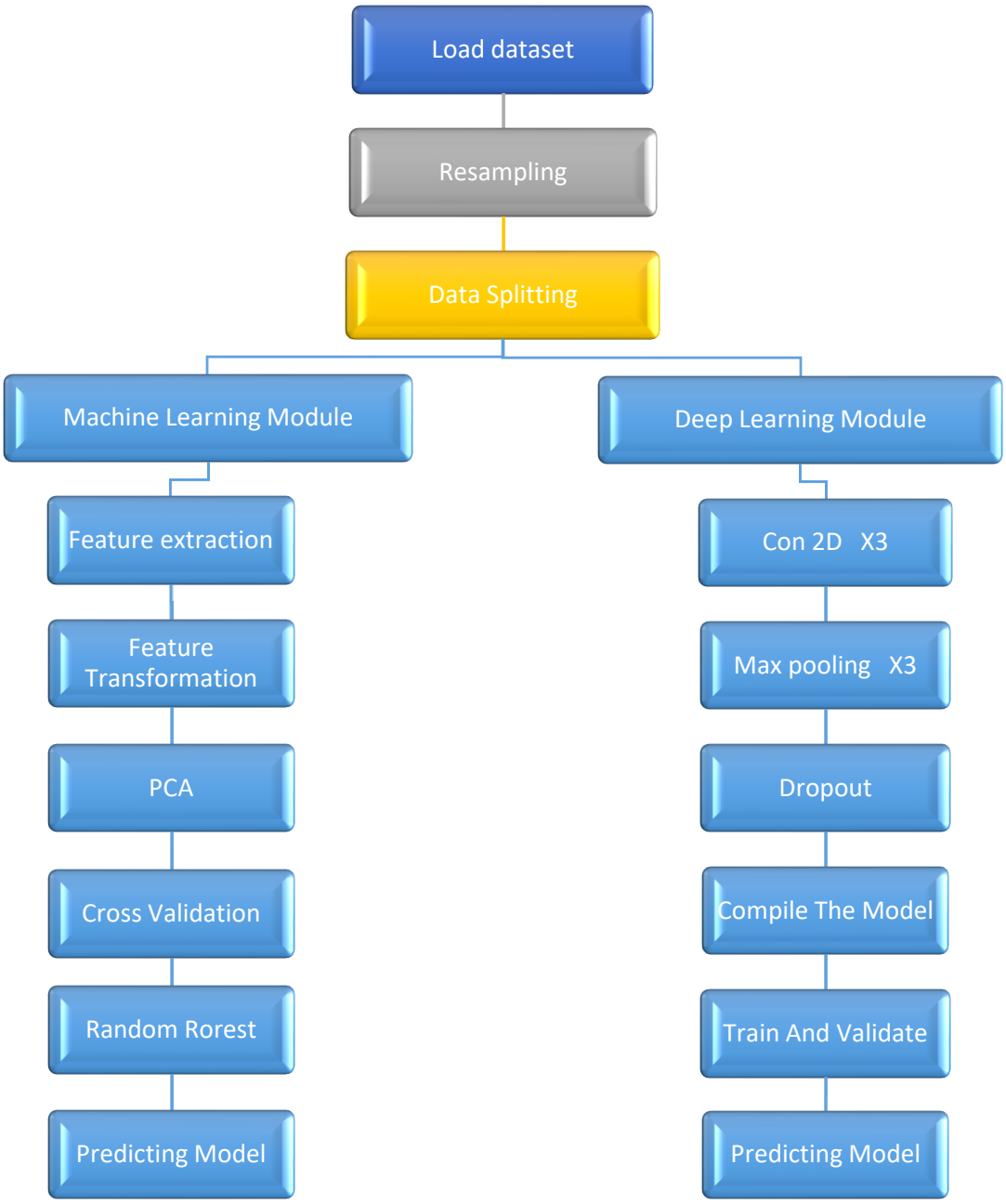


Figure 1: Flow diagram of the algorithms used in the implementation of the project.

2.1 Image Dataset

The image database used in this project consists of 2000 lesion images in JPEG format taken from the International Skin Imaging Collaboration (ISIC) 2017 dataset challenge. The database contains images from three tumor types: melanoma (374 images), seborrheic keratosis (254 images), and benign (1372 images).

Examples of the different tumor types are shown in Figure 2. To ensure accurate labeling and evaluation, the dataset also includes corresponding ground truth labels for all images. Each image is labeled with its respective tumor type corresponding to the image ID. This allows supervised learning and performance assessment of the classification models to be achieved.

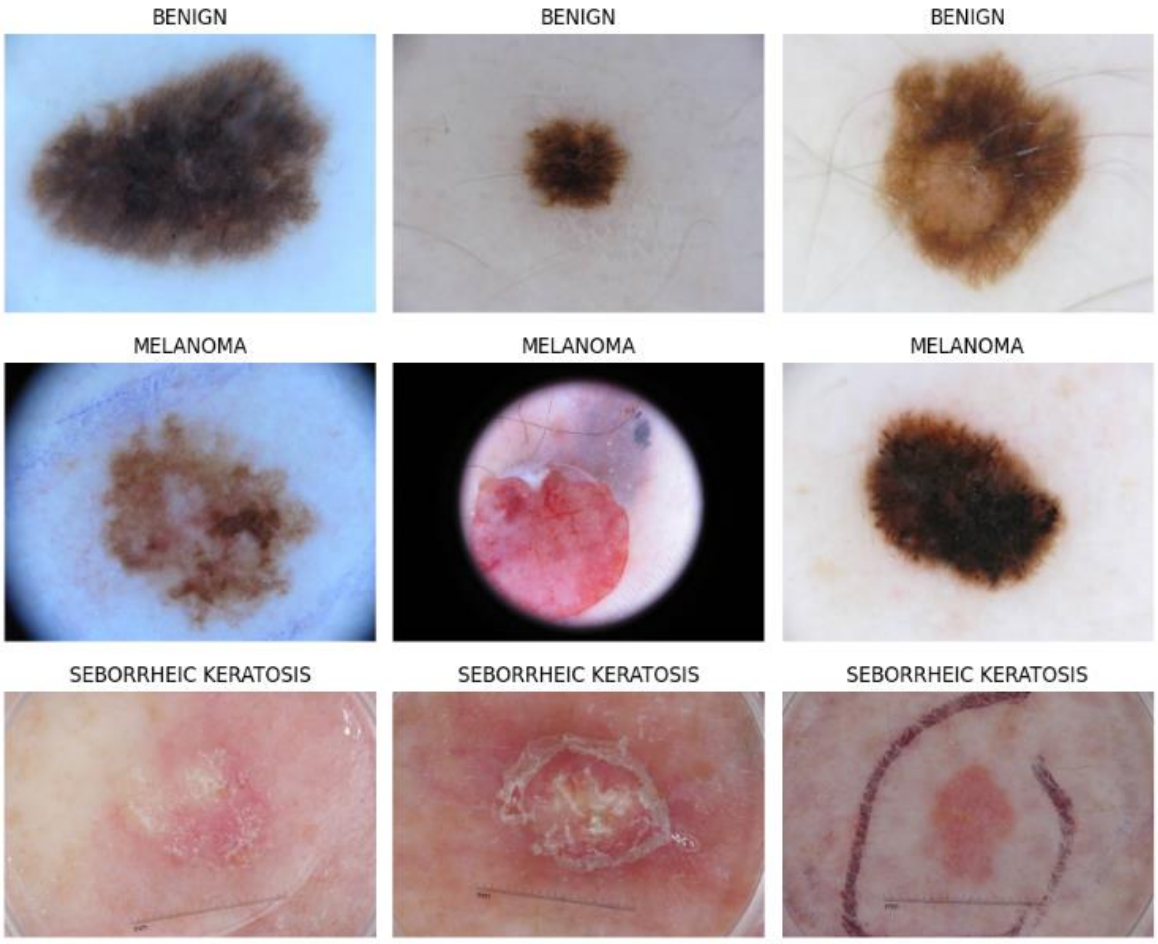


Figure 2: Examples of the different tumor types

3.2 Two-step Hierarchical Binary Classification

In this project, we used the two-step hierarchical binary classification approach to enhance classification performance and address the challenges posed by class imbalance. This technique offers robustness compared to traditional multiclass classification methods. The two-step classification framework involves a sequential process where the majority class (benign) is classified in the first step, followed by the classification of the remaining classes (melanoma vs. seborrheic keratosis) in the second step.

The first step of classification focuses on distinguishing between the benign class and the combination of the melanoma and seborrheic keratosis classes. This step allows us to identify instances that are most likely benign. In the second step, if the first step predicts the sample as "others" (non-benign), it undergoes further classification to determine whether it belongs to the melanoma or seborrheic keratosis class. This two-step approach enables more accurate classification results. It is important to note that although these steps are described sequentially, they are implemented simultaneously, and the final classification results are presented without an intermediate "others" classification. Figure 3 shows a schematic illustration of the two-step classification.

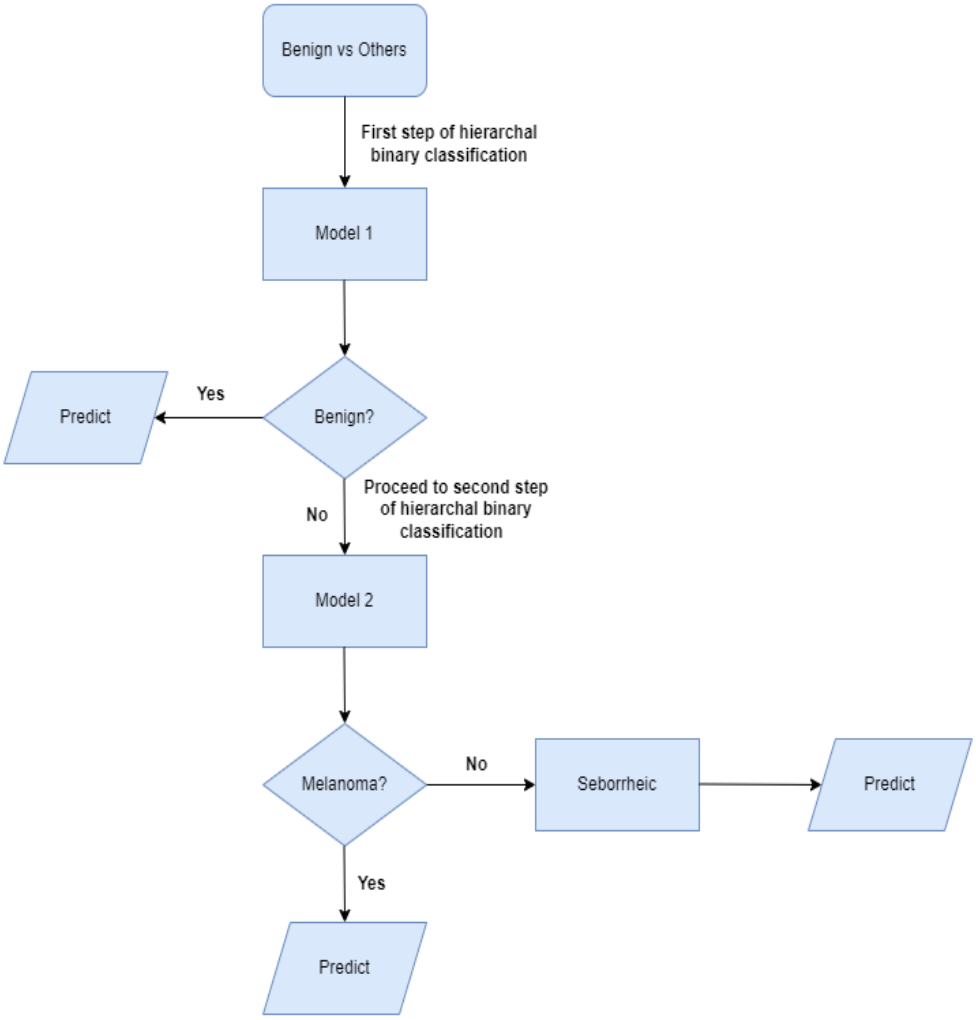


Figure 3: Schematic illustration of the two-step classification

3.3 Resampling and Data Splitting

Considering the significant class imbalance in the dataset, resampling techniques were applied to resolve this issue. The melanoma and seborrheic keratosis classes were combined and labeled as "others." This combined class was then upsampled to 1000 samples, while the benign class was downsampled to 1000 samples. This resampling resulted in a total of 2000 samples, with the "benign" and "others" classes needed for the first step of classification, and a balanced class distribution for the second step.

For data splitting, we used the class-wise splitting method by dividing the dataset such that the first 70% of images from each class were assigned to the training set, while the remaining 30% were allocated to the testing set, maintaining class-wise proportions. This fixed split ensures consistency across different groups when training the model. Furthermore, to prevent any potential bias from the order of the data, the dataset was shuffled after the data split.

3.4 Machine Learning Module

In this module, we tested several machine learning classifiers such as SVM, KNN, logistic regression, and RFC. To extract meaningful features from the data, we leveraged the same convolutional layers employed in the deep learning modules. The output of the last layer was flattened and utilized as a feature extractor, which was then fed into these classifiers. Additionally, we employed various feature engineering techniques to enhance the performance of the model. To evaluate and optimize the model, we employed a 5-fold cross-validation strategy, ensuring robustness and generalization of the results. The random forest outperformed other classifiers and hence, our predictive model was built on this classifier.

3.4.1 Feature Extraction and Transformation

In the feature extraction and transformation stage, we utilized convolutional filters to extract several features from the input data. The convolutional layers of the network effectively capture spatial patterns and hierarchical representations within the images. These learned features carry different information about the characteristics of skin lesions. To ensure fair comparison and prevent features with larger magnitudes from dominating the classification process, we applied the standardized scaler method. This standardized the

extracted features by subtracting the mean and dividing by the standard deviation, thus accounting for variations and bringing the features to a similar scale.

3.4.2 Principal Component Analysis

To address the challenge of high-dimensional feature spaces and mitigate potential overfitting, we have used the principal component analysis (PCA) technique in our machine learning module. PCA is widely used in dimensionality reduction to capture the most important patterns and variances in the data by projecting it onto a lower-dimensional subspace. The selection of the appropriate number of components needed for the PCA was also carefully considered, balancing the trade-off between preserving information and reducing dimensionality. In our approach, we set the number of components to 50 for the first step of binary classification (benign vs. others) and 70 for the second step (melanoma vs. seborrheic keratosis). We determined these values to be optimal for our classification task after several experiments and careful evaluation.

3.4.3 Five-Fold Cross-validation

To ensure the robustness and generalization ability of our machine learning model, we employed the technique of cross-validation. Specifically, we implemented 5-fold cross-validation during the training stage. This helps to assess how well the model is likely to perform on new, unseen data by evaluating its performance on multiple subsets of the training data. In our approach, we divided the training data into five approximately equal-sized folds. The model was then trained and evaluated five times, each time using a different fold as the validation set and the remaining folds as the training set. This process allows us to obtain a more robust estimation of the model's performance and detect any potential issues such as overfitting or underfitting.

3.4.4 Random Forest Classifier

Among several machine learning algorithms considered for this project, the random forest classifier emerged as the most effective in terms of accuracy and performance. Therefore, we adopted the random forest classifier as our final classification model. Since random forest classifier is an ensemble made up of multiple decision trees, we set our number of estimators, or decision trees to 50 for each step of the hierarchical classification. This value was determined based on experimentation and fine-tuning to achieve a good balance between model complexity and performance.

3.5 Deep Learning Module

In this module, we have implemented a deep learning model using a short convolutional neural network (CNN) architecture. The model architecture consists of several sequential layers designed to extract features from the input images and perform the final classification. We then implemented a predictive model for the two-step classification.

3.5.1 Network Architecture

To construct the network architecture, we started by creating a sequential model. We then added convolutional layers to the model, which played a crucial role in the feature extraction. The convolutional layers were configured with a specified number of filters (256, 128, 64) and a filter/kernel size of 3x3. The activation function "relu" was used to introduce non-linearity and enhance the model's ability to learn complex patterns. Max pooling layers were also incorporated into the network architecture. These layers helped in reducing the spatial dimensions of the feature maps, enabling the model to focus on the most relevant features while simultaneously reducing the number of parameters. The max-pooling layers enhanced the model's robustness and efficiency by downsampling the feature maps. Also, to mitigate overfitting, dropout layers were added to the network. Dropout regularization helps prevent the model from relying too heavily on specific features by randomly deactivating a certain percentage of neurons during training, promoting better generalization, and reducing the risk of overfitting. The final output of the model was obtained by flattening the 2D feature maps into a 1D vector, which was then fed into a dense layer for the final classification.

The first dense layer consisted of 32 units, which corresponded to half the size of our image. The "relu" activation function was applied and the number of class units of the final dense layer was set to 1, indicating the binary classification task. The "sigmoid" activation function was used in the final layer, as it is suitable for binary classification, providing output probabilities for the two classes. After constructing the network architecture, the model was compiled, and the model summary, including the number of trainable parameters, layer configurations, and output shapes, was printed. This information was helpful in gaining insights into the model's structure and ensuring its correct implementation. Table 1 shows the summary of the model's architecture, and Figure 4 for a visual representation of the network structure.

Table 1: The summary of the model's architecture

Layer No	Layer Name	Layer Type	Output Shape	Parameters
1	conv2d_6	Conv2D	(None, 62, 62, 256)	7168
2	max_pooling2d_6	MaxPooling2D	(None, 31, 31, 256)	0
3	dropout_6	Dropout	(None, 31, 31, 256)	0
4	conv2d_7	Conv2D	(None, 29, 29, 128)	295040
5	max_pooling2d_7	MaxPooling2D	(None, 14, 14, 128)	0
6	dropout_7	Dropout	(None, 14, 14, 128)	0
7	conv2d_8	Conv2D	(None, 12, 12, 64)	73792
8	max_pooling2d_8	MaxPooling2D	(None, 6, 6, 64)	0
9	dropout_8	Dropout	(None, 6, 6, 64)	0
10	flatten_2	Flatten	(None, 2304)	0
11	dense_4	Dense	(None, 32)	73760
12	dense_5	Dense	(None, 1)	33
Summary	=== Total params: 449,793 Trainable params: 449,793 Non-trainable params: 0 ===			

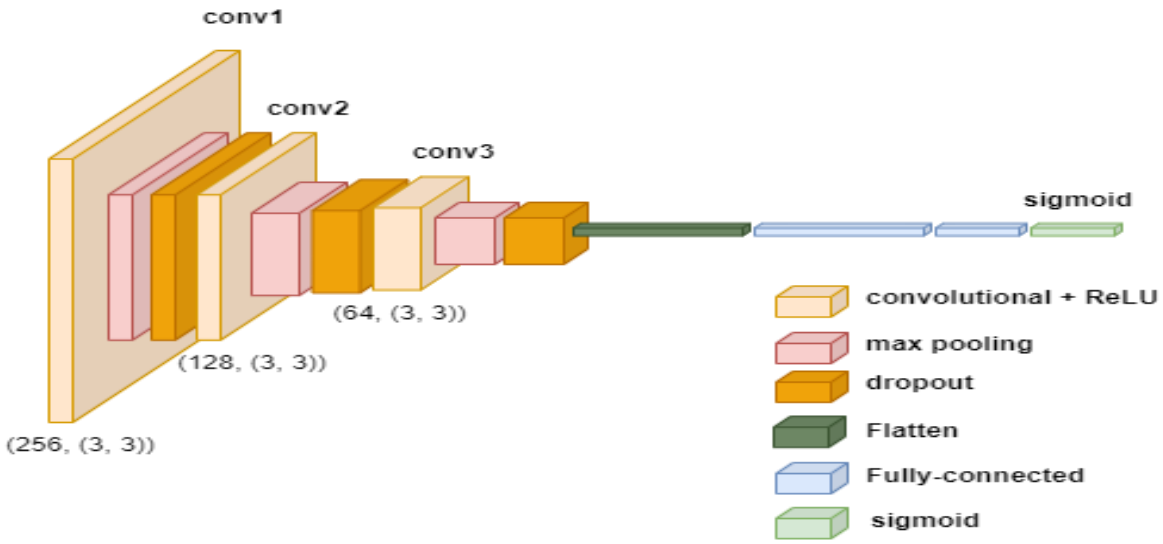


Figure 4: Visual representation of the network architecture

3.5.2 Training Network

To train the network, we used the Adam optimizer which is a popular optimization algorithm in training deep neural networks. The total number of trainable parameters in the CNN model was 449,793, which represents the number of learnable weights and biases. During the training process, we used a batch size of 32, indicating that the model was updated based on 32 images at a time. This batch-based training approach helps in optimizing memory usage and computational efficiency.

We performed training for a total of 50 epochs, which means that the entire dataset was processed 50 times during the training phase.

It is important to note that the same network architecture was utilized in training both steps of the hierarchical classification. This ensured consistency and comparability between the two stages of the classification process and ultimately improved the overall classification performance.

4.0 RESULTS AND DISCUSSION

Achieving high accuracy and reliable classification results using the ISIC 2017 dataset challenge is a challenging task. To address this, we have used a 2-step hierarchical binary classification approach instead of a traditional multiclass classification. This involved the use of both deep learning and machine learning techniques. In this section, we present and discuss the results obtained from each method, highlighting their respective strengths. Finally, we always introduced our predictive model, for easy use in real-life applications.

4.1 Cross-Validation Results for the Machine Learning Module

To assess the performance of the machine learning module, we used the 5-fold cross-validation for each step of the 2-step hierarchical binary classification. This allowed us to evaluate the effectiveness of different classifiers, including Random Forest, SVM, KNN, and logistic regression, in accurately classifying the skin lesions. Table 2 presents the results of cross-validation for the first stage of classification in the machine learning module. The Random Forest classifier consistently outperformed the other classifiers in each fold, achieving an average accuracy of 81.93%. SVM, KNN, and logistic regression yielded lower average accuracies of 67.02%, 69.21%, and 66.50% respectively. These results show that the Random Forest classifier performed best in distinguishing between benign and non-benign lesions.

For the second stage of classification, Table 3 showcases the results of the cross-validation. Once again, the Random Forest classifier exhibited the highest average accuracy of 90.13%, surpassing SVM, KNN, and logistic regression with average accuracies of 76.68%, 65.81%, and 77.83% respectively. This also means that the Random Forest classifier performs better in accurately classifying non-benign lesions into melanoma or seborrheic keratosis.

Table 2: Result of the cross-validation for the first stage in the machine learning module

Cross Validation	Random Forest	SVM	KNN	Logistic regression
1-fold	**0.8714**	0.6857	0.6857	0.6679
2-fold	**0.7929**	0.6250	0.7143	0.6500
3-fold	**0.7929**	0.6464	0.7000	0.6321
4-fold	**0.8250**	0.6786	0.6893	0.6821
5-fold	**0.8143**	0.7179	0.6714	0.6929

Table 3: Result of the cross-validation for the second stage in the machine learning module

Cross Validation	Random Forest	SVM	KNN	Logistic regression
1-fold	**0.8857**	0.7643	0.6714	0.7786
2-fold	**0.9143**	0.7929	0.7429	0.7786
3-fold	**0.8929**	0.7571	0.6071	0.7643
4-fold	**0.8857**	0.7929	0.6143	0.7929
5-fold	**0.9281**	0.7266	0.6547	0.7770

4.2 Performance Evaluation of the Machine Learning Module

To assess the performance of our model in the 2-step hierarchical binary classification, we conducted a comprehensive evaluation of the test data, which comprised 30% of the dataset, which was split class-wise to ensure an unbiased evaluation. Various evaluation metrics were used to measure the effectiveness of the classifiers: random forest, SVM, KNN, and logistic regression. Table 4 presents the performance evaluation results for the first stage of classification in the machine learning module. The random forest classifier exhibited the highest accuracy of 82.67%, outperforming SVM, KNN, and logistic regression, which achieved accuracies of 65.33%, 70.50%, and 63.83% respectively.

For the second stage of classification, table 5 shows the performance evaluation results. Once again, the random forest classifier demonstrated its effectiveness with an accuracy of 90.37%, surpassing SVM, KNN, and logistic regression with accuracies of 80.73%, 76.74%, and 82.06% respectively. Similarly, precision, recall, F1-score, area under the curve (AUC), and balanced accuracy (BA) metrics indicated the superior performance of the random forest classifier in accurately classifying skin lesions.

The results obtained from the performance evaluation of the machine learning module provide strong evidence of

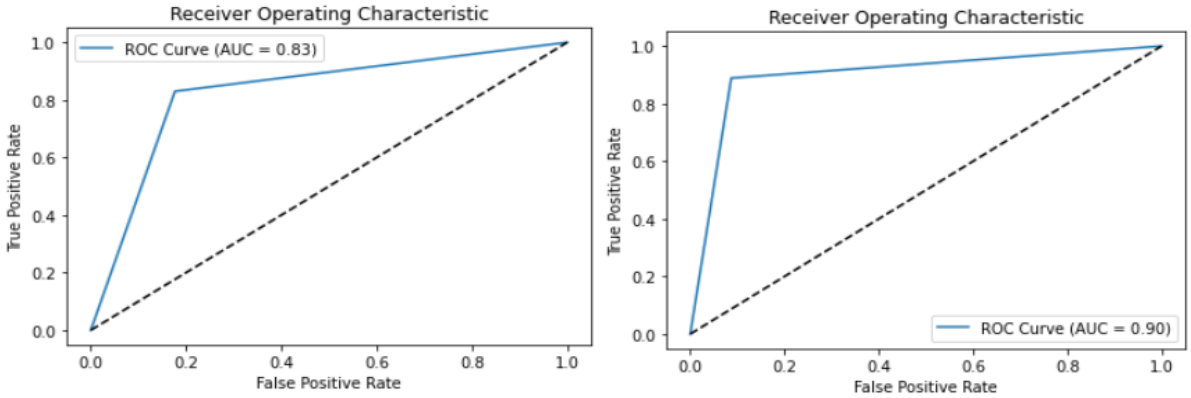
its reliability in accurately classifying skin lesions. The random forest classifier demonstrates the best performance with high accuracy, precision, recall, and F1-score values. The AUC value further shows the ability of the model to discriminate between different lesion classes. Additionally, the AUC-ROC curves depicted in Figure 5 (a and b) highlight the classifier's discriminative power. Figures 6 and figure 7 present the confusion matrix and the fraction of incorrect predictions, respectively, providing insights into the classifier's performance at each step of the classification. These findings emphasize the efficacy of the random forest classifier and its potential for enhancing clinical decision-making in the accurate classification of skin lesions.

Table 4: Performance evaluation of the first stage in the machine learning module

Evaluation Metrics	Random Forest	SVM	KNN	Logistic regression
Accuracy	**0.8267**	0.6533	0.7050	0.6383
Precision	**0.8245**	0.6556	0.7078	0.6402
Recall	**0.8300**	0.6533	0.7050	0.6383
F1-Score	**0.8272**	0.6521	0.7040	0.6371
AUC	**0.8267**	0.6533	0.7050	0.6383
BA	**0.8283**	0.6533	0.7050	0.6383

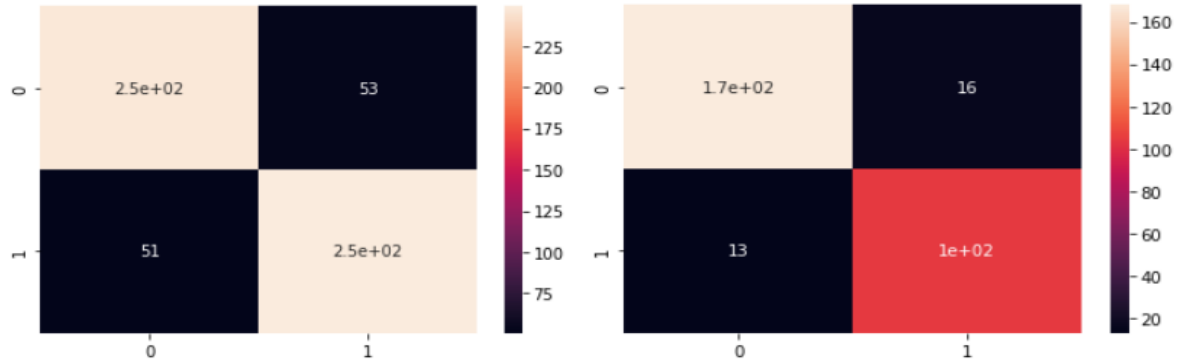
Table 5: Performance evaluation of the second stage in the machine learning module

Evaluation Metrics	Random Forest	SVM	KNN	Logistic regression
Accuracy	**0.9037**	0.8073	0.7674	0.8206
Precision	**0.8667**	0.8068	0.7667	0.8197
Recall	**0.8888**	0.8073	0.7674	0.8206
F1-Score	**0.8776**	0.8070	0.7671	0.8200
AUC	**0.9010**	0.7957	0.7538	0.8081
BA	**0.8963**	0.8073	0.7674	0.8206



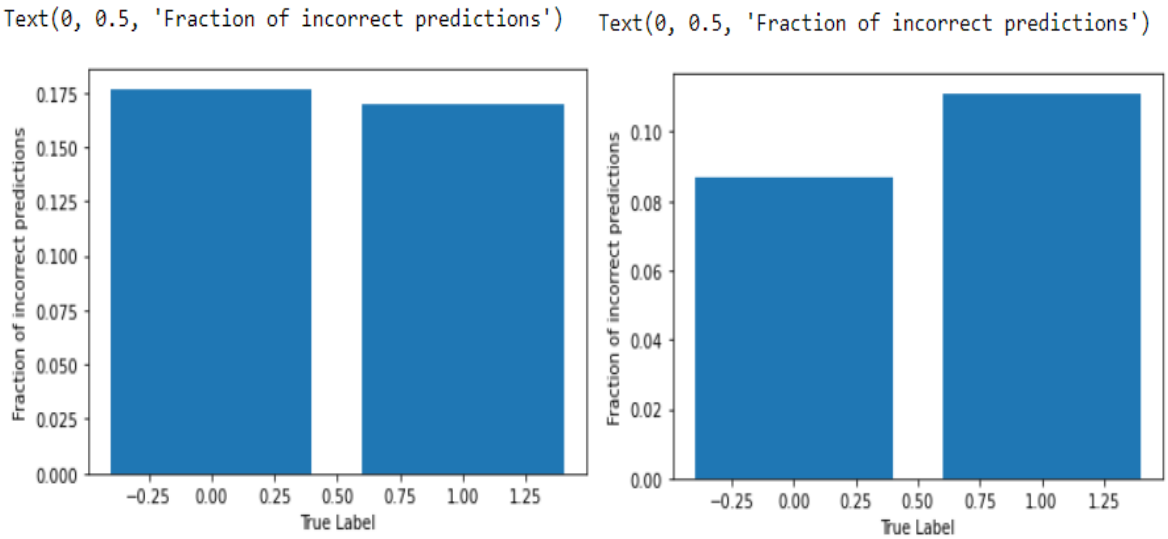
5a. AUC-ROC curve for step 1 5b. AUC-ROC curve for step 2

Figure 5: AUC-ROC curve for the machine learning module



6a. Confusion matrix for step 1 6b. Confusion matrix for step 2

Figure 6: Plot of the confusion matrix for each step



7a. Fraction of incorrect predictions for step 1

7b. Fraction of incorrect predictions for step 2

Figure 7: Plot of fractional incorrect misclassifications

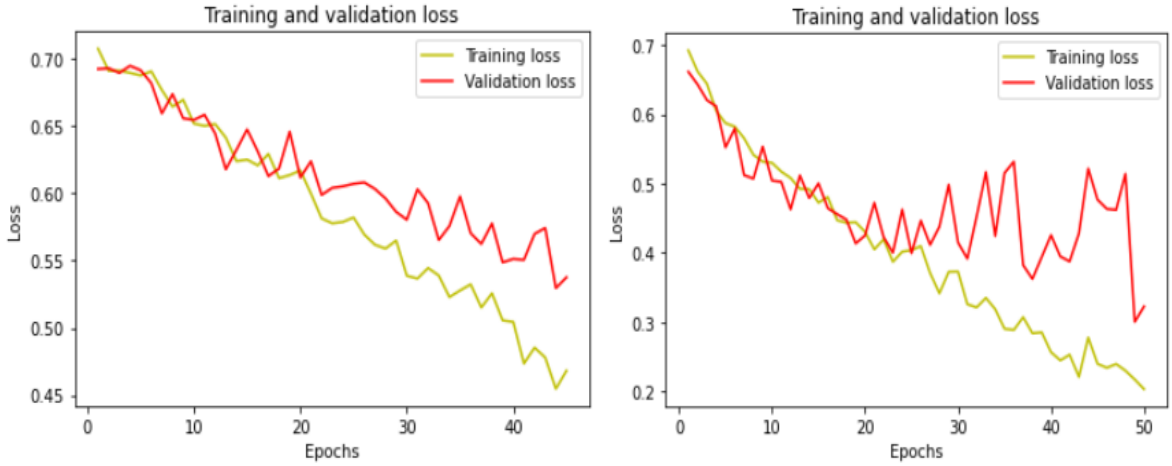
4.3 Performance Evaluation of the Deep Learning Module

The performance evaluation of the deep learning module was achieved by training a convolutional neural network (CNN) and comparing its performance with the random forest classifier. Figures 8 and figure 9 depict the training and validation loss plots, as well as the training and validation accuracy, respectively, for each step of the classification.

The evaluation was conducted using various metrics, and the results are presented in table 6 and table 7, highlighting the comparison between the deep learning CNN and the random forest classifier extracted from the same CNN. In the first step of the hierarchical binary classification, the random forest classifier achieved an accuracy of 82.67%, while the CNN achieved an accuracy of 75.00%. Similarly, the precision, recall, F1-score, AUC, and BMA metrics were slightly lower for the CNN compared to the random forest classifier. However, in the second step, the CNN showcased competitive performance, achieving an accuracy of 89.04%, slightly below the random forest classifier's accuracy of 90.37%. The CNN also demonstrated a higher recall value compared to the random forest classifier. While the random forest classifier outperformed the CNN in terms of accuracy and other metrics, especially in the first step, the CNN showed promising performance in the second step.

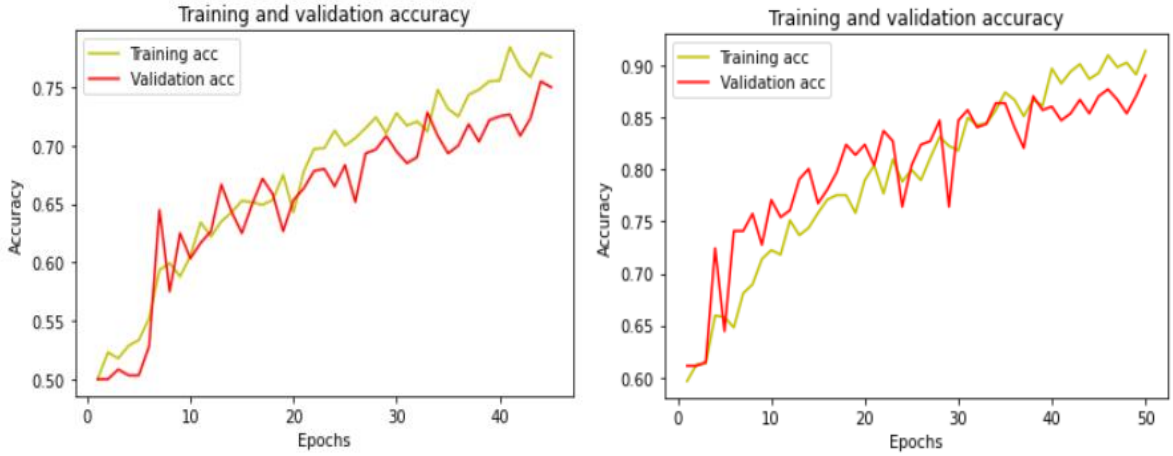
Additionally, CNN's overall performance was not consistently superior to the random forest classifier throughout the entire classification process. The advantage of the random forest classifier lies in its ability to handle high-dimensional feature spaces and performs well even with limited training data. On the other hand, deep-learning CNNs excel in learning intricate features and patterns, particularly in image recognition tasks. They can automatically extract relevant features from the data, eliminating the need for manual feature engineering. However, CNNs require a large amount of training data to generalize effectively and may be susceptible to overfitting when training data is limited. In our case, the limited performance of the deep learning CNN is attributed to the relatively smaller size of the dataset, which did not provide enough training examples for the CNN to effectively learn and generalize the complex features associated with different skin lesion types.

Overall, both the random forest classifier and the deep learning CNN have shown promising results and with more training data, the CNN has the potential to outperform the random forest classifier.



8a. the training and validation loss for step 1 8b. the training and validation loss for step 2

Figure 8: Plot for the training and validation loss



9a. the training and validation accuracy for step 1 9b. the training and validation accuracy for step 2

Figure 9: Plot for the training and validation accuracy

Table 6: Comparison of the result achieved by deep learning CNN and the random forest classifier in the first step of the 2-step hierarchical binary classification

Evaluation Metrics	Random Forest	CNN
Accuracy	**0.8267**	0.7500
Precision	**0.8245**	0.7483
Recall	**0.8300**	0.7533
F1-Score	**0.8272**	0.7508
AUC	**0.8267**	0.7500
BA	**0.8283**	0.7517

Table 7: Comparison of the result achieved by deep learning CNN and the random forest classifier in the second step of the 2-step hierarchical binary classification.

Evaluation Metrics	Random Forest	CNN
Accuracy	**0.9037**	0.8904
Precision	**0.8667**	0.8333
Recall	0.8888	**0.8974**
F1-Score	**0.8776**	0.8642
AUC	**0.9010**	0.8917
BA	**0.8963**	0.8939

4.5 Implementation of Predictive Model for the 2-step Hierarchical Binary Classification

Following the results from the performance evaluation of our classification models, we proceeded to implement a predictive model that integrates the two-step hierarchical binary classification process. This step is important in our project as it enables us to make accurate predictions by combining the outcomes of the individual steps to obtain the final classification.

The implemented predictive model follows a sequential approach. When an image is passed, the model initially distinguishes between benign and non-benign lesions, utilizing the first step of the hierarchical classification. If the lesion is classified as benign, the process concludes at this stage. However, if the lesion falls into the non-benign category (classified as "others"), the model proceeds to the second step for final classification into melanoma or seborrheic keratosis. A separate predictive model was built for both the machine learning and deep learning module of the project. Figure 10 provides a visual representation of the results obtained from the predictive model, showcasing an example of a correctly predicted seborrheic keratosis lesion, alongside the corresponding real image.

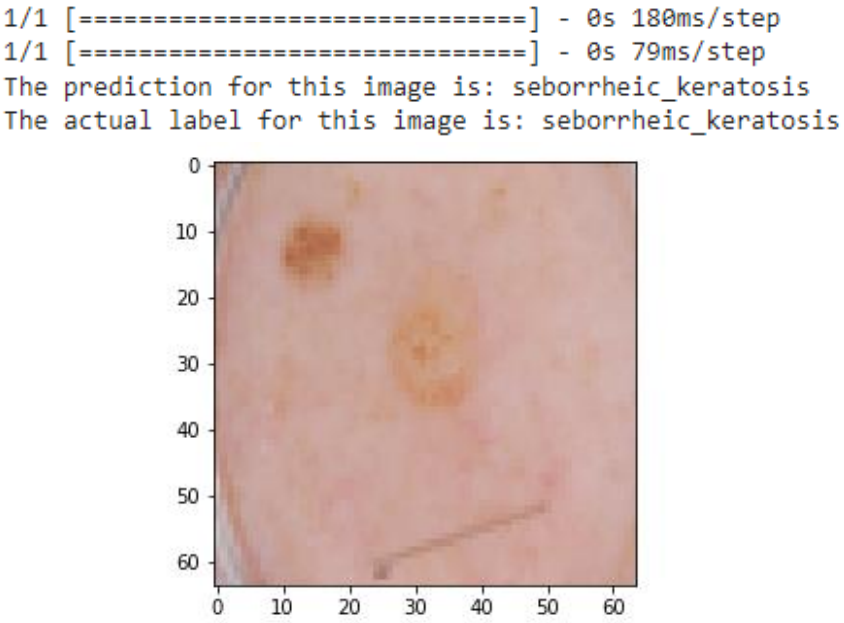


Figure 10: Result of the predictive model

4.4 Balanced Multiclass Accuracy

The Balanced Multiclass Accuracy (BMA) is a crucial metric for evaluating the performance of our classification model. It represents the average accuracy across all classes, considering the imbalanced nature of the dataset. Higher BMA values indicate better overall classification performance of the project. To calculate the BMA, we considered the accuracy of each class and averaged them. In the first step of our 2-step hierarchical binary classification, we calculated the accuracy for the "Benign" class, while in the second step, we calculated the accuracies for the "Melanoma" and "Seborrheic" classes. The BMA was then obtained by averaging the accuracies across all classes.

The random forest classifier achieved a Balanced Multiclass Accuracy of 87.51%, indicating its effectiveness in accurately classifying the different skin lesion classes. On the other hand, CNN achieved a Balanced Multiclass Accuracy of 84.33% which is slightly lower compared to the random forest classifier, but still very good considering the size of the dataset.

5.0 Conclusion

In conclusion, this project aimed to develop a robust and accurate classification system for skin lesion diagnosis. Instead of using the traditional multiclass classification method, we implemented a 2-step hierarchical binary classification using both machine learning and deep learning approaches. The results obtained from the evaluation of the machine learning module showcased the potential of the random forest classifier in accurately classifying skin lesions, achieving high accuracy, precision, recall, and F1-score values amongst other techniques such as SVM, KNN, and logistic regression. Additionally, the deep learning module showed competitive performance in classifying seborrheic keratosis and melanoma, even though the random forest classifier slightly outperformed it in terms of accuracy. Overall, this project has shown

how machine learning and deep learning techniques can help dermatologists in the early detection and diagnosis of skin lesions. The Balanced Multiclass Accuracy (BMA) metric further solidifies the reliability of our models, with the random forest classifier achieving a BMA of **87.51%** and the CNN achieving a BMA of **84.33%**. These findings highlight the significance of using advanced computational techniques to enhance diagnostic accuracy and support efficient management of skin conditions. This project paves the way for future developments in the field of dermatology and adds to the growing body of research in computer-aided diagnosis.

Abbreviation

CNN – Convolutional Neural Networks
ISIC - International Skin Imaging Collaboration
UV – Ultraviolet
BCC - Basal Cell Carcinoma
SCC - Squamous Cell Carcinoma
ML - Machine Learning
DL - Deep Learning
SVM - Support Vector Machines
KNN - K-Nearest Neighbors
RFC - Random Forest Classifier
PCA - Principal Component Analysis
AUC – Area Under the Curve
BA - Balanced Accuracy
BMA - Balanced Multiclass Accuracy

Declarations

Acknowledgments

The authors would like to thank all the reviewers for their valuable comments.

Author contributions

Conceptualization: ASDF. Investigation and Formal Analysis: TAS, DTA. ADP, HAN, Resources: TAS, DTA. ADP, HAN, Methodology, and Validation: TAS, DTA, ADP. Writing – Original Draft: TAS, DTA. ADP, HAN, Writing – Review & Editing: ASDF, TAS. Visualization: TAS, DTA. Supervision: ASDF.

Funding

This review received no funding.

Data availability

The data that support the findings of this study is the publicly available ISIC 2017 training dataset.

Competing interests

The authors declare that they have no competing interests.

References

[1] K. D. Rojas, M. E. Perez, M. A. Marchetti, A. J. Nichols, F. J. Penedo, and N. Jaimes, "Skin cancer: Primary, secondary, and tertiary prevention. Part II.," *J Am Acad Dermatol*, vol. 87, no. 2, pp. 271–288, Aug. 2022, doi: 10.1016/J.JAAD.2022.01.053.

[2] D. Didona, G. Paolino, U. Bottoni, and C. Cantisani, "Non Melanoma Skin Cancer Pathogenesis Overview," *Biomedicines*, vol. 6, no. 1, Jan. 2018, doi: 10.3390/BIOMEDICINES6010006.

[3] B. Ahmed, M. I. Qadir, and S. Ghafoor, "Malignant Melanoma: Skin Cancer-Diagnosis, Prevention, and Treatment," *Crit Rev Eukaryot Gene Expr,* vol. 30, no. 4, pp. 291–297, 2020, doi: 10.1615/CRITREVEUKARYOTGENEEXPR.2020028454.

[4] V. Suhonen, H. Siiskonen, M. Suni, J. Rummukainen, A. Mannermaa, and I. T. Harvima, "Malignant and in situ subtypes of melanoma are associated with basal and squamous cell carcinoma and its precancerous lesions," Eur J Dermatol, vol. 32, no. 2, pp. 187–194, Mar. 2022, doi: 10.1684/EJD.2022.4221.

[5] G. S. Laberge et al., "Recent Advances in Studies of Skin Color and Skin Cancer," *Yale J Biol Med,* vol. 93, no. 1, pp. 69–80, Mar. 2020.

[6] M. Marzi et al., "Applications of Metallic Nanoparticles in the Skin Cancer Treatment," *Biomed Res Int*, vol. 2022, 2022, doi: 10.1155/2022/2346941.

[7] A. Pashazadeh, A. Boese, and M. Friebe, "Radiation therapy techniques in the treatment of skin cancer: an overview of the current status and outlook," *https://doi.org/10.1080/09546634.2019.1573310,* vol. 30, no. 8, pp. 831–839, Nov. 2019, doi: 10.1080/09546634.2019.1573310.

[8] S. J. Quazi, N. Aslam, H. Saleem, J. Rahman, and S. Khan, "Surgical Margin of Excision in Basal Cell Carcinoma: A Systematic Review of Literature," *Cureus,* vol. 12, no. 7, Jul. 2020, doi: 10.7759/CUREUS.9211.

[9] L. Zeng et al., "Open Access Molecular Cancer Advancements in nanoparticle-based treatment approaches for skin cancer therapy," *Mol Cancer,* vol. 22, p. 10, 2023, doi: 10.1186/s12943-022-01708-4.

[10] R. Talty and M. Bosenberg, "The role of ferroptosis in melanoma," *Pigment Cell Melanoma Res*, vol. 35, no. 1, pp. 18–25, Jan. 2022, doi: 10.1111/PCMR.13009.

[11] N. L. Bolick and A. C. Geller, "Epidemiology of Melanoma," *Hematol Oncol Clin North Am*, vol. 35, no. 1, pp. 57–72, Feb. 2021, doi: 10.1016/J.HOC.2020.08.011.

[12] P. Aggarwal, P. Knabel, and A. B. Fleischer, "United States burden of melanoma and non-melanoma skin cancer from 1990 to 2019," *J Am*

Acad Dermatol, vol. 85, no. 2, pp. 388–395, Aug. 2021, doi: 10.1016/J.JAAD.2021.03.109.

[13] N. H. Khan et al., "Skin cancer biology and barriers to treatment: Recent applications of polymeric micro/nanostructures," *J Adv Res,* vol. 36, p. 223, Feb. 2022, doi: 10.1016/J.JARE.2021.06.014.

[14] R. Gall, M. Bongiorno, and K. Handfield, "Skin cancer in the US Military," *Cutis,* vol. 107, no. 1, pp. 29–33, Jan. 2021, doi: 10.12788/CUTIS.0153.

[15] M. Imran Qadir, P. J. Pharm Sci, and M. Imran Qadir, "Skin cancer: Etiology and management," vol. 29, no. 3, pp. 999–1003, 2016.

[16] K. A. Merin, M. Shaji, and R. Kameswaran, "A Review on Sun Exposure and Skin Diseases," *Indian J Dermatol*, vol. 67, no. 5, p. 625, Sep. 2022, doi: 10.4103/IJD.IJD_1092_20.

[17] P. Fontanillas et al., "Disease risk scores for skin cancers," *Nature Communications* 2021 12:1, vol. 12, no. 1, pp. 1–13, Jan. 2021, doi: 10.1038/s41467-020-20246-5.

[18] A. K. Thompson, B. F. Kelley, L. J. Prokop, M. H. Murad, and C. L. Baum, "Risk Factors for Cutaneous Squamous Cell Carcinoma Recurrence, Metastasis, and Disease-Specific Death: A Systematic Review and Meta-analysis," *JAMA Dermatol,* vol. 152, no. 4, pp. 419–428, Apr. 2016, doi: 10.1001/JAMADERMATOL.2015.4994.

[19] T. Brambullo et al., "Current Surgical Therapy of Locally Advanced cSCC: From Patient Selection to Microsurgical Tissue Transplant. Review," *Front Oncol*, vol. 11, p. 783257, Dec. 2021, doi: 10.3389/FONC.2021.783257/BIBTEX.

[20] M. Jindal, M. Kaur, M. Nagpal, M. Singh, G. Aggarwal, and G. A. Dhingra, "Skin Cancer Management: Current Scenario And Future Perspectives," *Curr Drug Saf*, vol. 18, no. 2, pp. 143–158, Apr. 2023, doi: 10.2174/1574886317666220413113959.

[21] V. Madan, J. T. Lear, and R. M. Szeimies, "Non-melanoma skin cancer," *The Lancet*, vol. 375, no. 9715, pp. 673–685, Feb. 2010, doi: 10.1016/S0140-6736(09)61196-X.

[22] W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning," *Healthcare,* vol. 10, no. 7, Jul. 2022, doi: 10.3390/HEALTHCARE10071183.

[23] T. Skuhala, V. Trkulja, M. Rimac, A. Dragobratović, and B. Desnica, "Analysis of Types of Skin Lesions and Diseases in Everyday Infectious Disease Practice—How Experienced Are We?," *Life,* vol. 12, no. 7, Jul. 2022, doi: 10.3390/LIFE12070978.

[24] A. Lallas, Z. Apalla, E. Lazaridou, and D. Ioannides, "Dermoscopy," 2016, doi: 10.1016/B978-0-12-802838-4.00003-0.

[25] K. Batko and A. Ślęzak, "The use of Big Data Analytics in healthcare*," J Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/S40537-021-00553-4.

[26] H. Habehh and S. Gohel, "Machine Learning in Healthcare," *Curr Genomics*, vol. 22, no. 4, p. 291, Dec. 2021, doi: 10.2174/1389202922666210705124359.

[27] B. Mahesh, "Machine Learning Algorithms-A Review," 2019, doi: 10.21275/ART20203995.

[28] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* 2015 521:7553, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[29] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, "Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review," *Diagnostics* 2021, Vol. 11, Page 1390, vol. 11, no. 8, p. 1390, Jul. 2021, doi: 10.3390/DIAGNOSTICS11081390.

[30] M. Bistroń and Z. Piotrowski, "Comparison of Machine Learning Algorithms Used for Skin Cancer Diagnosis," *Applied Sciences* 2022, Vol. 12, Page 9960, vol. 12, no. 19, p. 9960, Oct. 2022, doi: 10.3390/APP12199960.

[31] M. Q. Hatem, "Skin lesion classification system using a K-nearest neighbor algorithm," *Vis Comput Ind Biomed Art*, vol. 5, no. 1, pp. 1–10, Dec. 2022, doi: 10.1186/S42492-022-00103-6/FIGURES/5.

[32] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," Scientific Reports 2022 12:1, vol. 12, no. 1, pp. 1–11, Oct. 2022, doi: 10.1038/s41598-022-22644-9.