Skin Cancer Detection Using Structured Clinical Metadata and VGG16-Based Image Classification

Arun K. Bashambu[1*], Manisha Agarwal[2], Shallu Bashambu[3] and Arushi Bashambu[4]

[1] Department of Computer Science, Banasthali Vidyapith, Rajasthan, 304022, India
[2] Department of Computer Science, Banasthali Vidyapith, Rajasthan, 304022, India
[3] Department of Information Technology, Maharaja Agarsen Institute of Technology, Rohini, Delhi, 110086, India
[4] Capgemini Technology Services, Bangalore, India
arunbashambu@rediff.com*, amanisha@banasthali.in, shallubashambu@mait.ac.in, arushibashambu@capgemini.com

Abstract. Skin cancer is a major dermatological health concern, and timely identification of malignant lesions can support clinical decision-making. This study presents a machine-learning framework using the publicly available HAM10000 dataset, which contains 10,015 dermoscopic images together with structured metadata. The study considers two complementary data pathways: conventional machine-learning classification using structured metadata and VGG16-based transfer learning for dermoscopic images. The binary target distinguishes cancerous from non-cancerous lesions, while the original dataset contains seven diagnostic categories. Data preprocessing included missing-value treatment, categorical encoding, Min-Max normalization for structured variables, image resizing to 224×224 pixels, pixel normalization, and augmentation through flipping, rotation, and zooming. Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) were evaluated on the structured metadata using an 80:20 train-test split, with 5-fold cross-validation used for hyper-parameter tuning. The reported structured-data results show accuracies of 83.45% for LR, 85.32% for DT, and 89.56% for RF. RF also achieved 88.70% precision, 87.90% recall, and 88.30% F1-score. VGG16 is described as the image-classification pathway; however, quantitative CNN results are not reported in the present experimental results table. Therefore, the manuscript does not claim a fused multimodal performance or attribute the 89.56% result to a hybrid model. The principal reported finding is that RF performed best among the evaluated structured-metadata classifiers. The image pathway is retained as a complementary VGG16-based methodology, but its quantitative performance and any multimodal fusion result are not claimed without corresponding experimental evidence. The framework is presented as a potential decision-support approach rather than a replacement for dermatological diagnosis.

Keywords: Skin Cancer, Machine Learning, Random Forest, Logistic Regression, VGG16, HAM10000, Dermoscopic Images, Clinical Metadata.

© The Author(s) 2026
A. Shastri et al. (eds.), *Proceedings of the Fourth Workshop on Applications of Computational Intelligence and Interactive Systems (WACIIS 2026)*, Advances in Intelligent Systems Research 217,
https://doi.org/10.2991/978-94-6239-832-0_22

1 Introduction

Cancer develops when normal cells in the body experience genetic alterations that cause them to multiply without control, resulting in the formation of a mass called a tumor. Tumors may be noncancerous (benign), meaning they remain localized, or cancerous (malignant), meaning they can invade nearby tissues and spread to distant parts of the body.

Skin cancer is a long-term disease that affects people around the world. It is generally divided into two primary groups: melanoma, which is more aggressive and capable of spreading, and non-melanoma skin cancers, which are more common and typically grow more slowly. If not treated in time, skin cancer can cause permanent skin damage, noticeable scarring, pain, and bleeding. Direct exposure to UV radiation is a significant cause of skin cancer. Warning signs include changes in skin color, size, and shape, along with alterations in the color of lesions or moles, the appearance of new growths, swelling, and bleeding. If you notice any of these signs, it is advisable to consult a dermatologist.

According to various reports, the incidence of skin cancer cases and related deaths is rising. The World Health Organization (WHO) estimates around 3 million cases of non-melanoma skin cancer and 132,000 cases of melanoma occur worldwide each year. In the United States alone, nearly 10,000 people are diagnosed with skin cancer annually, with one in every five individuals likely to develop it at some point in their lives. The annual number of skin cancer cases frequently surpasses estimates.

Recent progress in Convolutional Neural Networks (CNNs) has significantly advanced medical image analysis, particularly in image classification tasks. A major breakthrough occurred with the introduction of AlexNet by Krizhevsky et al. in 2012, which demonstrated outstanding performance in large-scale visual recognition challenges [1]. The breakthrough of this design led to its usage for medical imaging problems, which include automated skin lesion classification.

By using deep learning frameworks, Mahbod et al. presented a hybrid deep neural network technique that combines CNN architectures with ensemble learning techniques and handcrafted feature extraction [2]. By integrating multiple models and applying data augmentation methods, their strategy improved the accuracy of distinguishing benign from malignant skin lesions.

Other researchers have explored complementary methods to enhance classification robustness. Takruri et al. introduced a multi-classifier fusion technique aimed at improving detection reliability [3]. Similarly, Sundar and Vadivel emphasized the importance of extracting discriminative features such as color, shape, and texture, and applying Support Vector Machines (SVM) for melanoma classification [4].

From a clinical standpoint, the American Cancer Society highlights the ABCDE rule (Asymmetry, Border irregularity, Color variation, Diameter, and Evolution) as essential criteria for early melanoma detection [5]. These clinical guidelines provide valuable benchmarks that can support the training and evaluation of artificial intelligence systems.

In this work, we investigate two complementary pathways using the same HAM10000 dataset: conventional machine-learning models (Logistic Regression, Decision Tree, and Random Forest) applied to structured clinical metadata, and a VGG16 transfer-learning pathway applied to dermoscopic images. These pathways are evaluated separately rather than fused into a single multimodal predictor. This distinction is important because the quantitative results reported in this manuscript are for the structured-metadata classifiers.

2 Literature Review

The reviewed studies demonstrate substantial progress in the application of machine learning techniques for skin cancer detection. Advances in feature extraction, image processing, and deep learning architectures have contributed to notable improve-ments in diagnostic performance. Despite these developments, further research is needed to enhance model generalization, integrate multi-modal data sources, and develop efficient real-time diagnostic systems to support early detection and improve patient outcomes.

A major milestone in deep learning was the introduction of the AlexNet architecture by Krizhevsky, Sutskever, and Hinton in 2012 [1]. Their work showed that deep CNNs could surpass traditional machine learning methods in large-scale image classification tasks, particularly on the Image Net dataset. Key innovations of this architecture included the use of Rectified Linear Unit (ReLU) activation functions for speeding up training, dropout medhods for decreasing over-fitting, and GPU-based computation for enabling efficient training of deep models. This win of AlexNet showed the efficiency of Deep Learning (DL) in complicated visual recognition works and accelerated its usage in medical imaging applications. It was adopted for in automatic detection of skin cancer.

Mahbod et al. build on this and introduced a hybrid DL system for skin lesion classification [2]. Their technique has multiple CNN models and engineered feature representations using an ensemble strategy. The research included the advantages of data augmentation and transfer learning for making model more robust. It also enhanced its generalization capability. Test results implied that hybrid ensemble approach performed better as compared to single CNN architectures for distinguishing between benign and malignant lesions. This showed that it has high potential for application in clinical dermatology.

The American Cancer Society gives wholistic data and information about the clinical presentation of skin cancer, including melanoma, basal cell carcinoma, and squamous cell carcinoma [3]. The resource emphasizes the ABCDE guideline (Asym-metry, Border irregularity, Color variation, Diameter, and Evolution) for early mela-noma detection and outlines common warning signs such as evolving moles, non-healing lesions, itching, and bleeding. These clinically established indicators serve as valuable reference standards for training artificial intelligence models in dermato-logical image analysis.

In addition, the American Cancer Society details standard diagnostic procedures for melanoma, including dermoscopy, biopsy, and histopathological examination [4]. The document also highlights the growing role of digital imaging and AI-assisted tools in supporting clinical decision-making. Understanding these diagnostic workflows is essential for developing deep learning systems that approximate expert-level lesion assessment.

Takruri et al. introduced a non-invasive and multi-classifier fusion framework which automatically found melanoma [5]. The introduced framework had outputs from three classifiers for improving the performance of detection. It also helped in handling the high mortality rate associated with melanoma. The fusion-based strategy demonstrated enhanced reliability compared to individual classifiers.

Similarly, Sundar and Vadivel developed a machine learning framework for melano-ma classification using image processing techniques [6]. Their approach extracted discriminative features such as color, texture, and shape, followed by classification to distinguish benign from malignant lesions. The study emphasized effective segmentation and feature selection to improve accuracy while reducing computational complexity.

Krohling introduced a smartphone-based deep learning application for skin cancer classification, integrating CNN models with patient demographic data [7]. To address dataset imbalance, the study applied a Differential Evolution-based optimization strategy. The system achieved strong balanced accuracy and recall, demonstrating the feasibility of mobile AI tools for early melanoma screening.

LUCA research (2022) found AI-driven diagnostic technique which used DL models for analyzing the efficiency of clinical data [8]. The study reflected the high scope of AI systems in improving healthcare accessibility. It also showed the high detection precision by automating interpretation of data.

Ahsan et al. introduced an enhanced VGG16-based DL framework for finding Monkeypox using clinical images [9]. By making use of transfer learning methods, the research got high classification accuracy. It showed the fitment of CNN design for detecting infection diseases.

Kanwal et al. presented an automation-based system for finding melanoma and for localization in histopathological whole-slide images [10]. The framework used enhanced image processing and segmentation methods for improving detection of malignant regions, supporting pathologists in diagnosis.

Hao et al. proposed an automation system to identify skin hair follicles. Mueller matrix polarimetry [11] was used. By using polarization-based imaging, the technique advanced visualization and quantitative analysis of skin structures. This research continued to enhancement in dermatological imaging technology.

Mohamad Sofie et al 2025[12] explores skin cancer detection using basic machine learning techniques instead of complex deep learning models, aiming to provide a simple and efficient classification system. The researchers implemented algorithms such as SVM and K-Nearest Neighbors (KNN) to classify skin lesion images into benign and malignant categories using extracted features like texture and color. The proposed system achieved an accuracy of approximately 85–90%.

Gabani V et al. [13] focused on multimodal skin lesion classification by combining dermoscopic images with deep-learning ensemble models. Their work highlights the value of combining complementary information sources. In contrast, the present study evaluates metadata-based conventional classifiers and a VGG16 image pathway separately. The distinction helps define the research gap addressed here: a clear comparison of structured clinical metadata classifiers alongside an image-based transfer-learning pathway, with explicit separation of the two experimental modali-ties.

3 Proposed Methodology

The methodology is organized into separate structured-data and image-data pathways. The two branches are evaluated independently; no feature-level or decision-level fusion is performed in the experiments reported here.

3.1 Dataset Acquisition

The proposed study uses the HAM10000 (Human Against Machine with 10000 training images) dataset. It contains 10,015 dermoscopic images and associated metadata, including age, sex, anatomical location, and diagnosis. The dataset contains seven diagnostic categories listed in Table 1. For the primary binary classification task, the diagnosis labels are grouped into cancerous and non-cancerous classes.

Table 1. Diagnostic categories frequencies

Label	Full Form	Frequency
akiec	Actinic Keratoses and Intraepithelial Carcinoma	327
bcc	Basal Cell Carcinoma	514
bkl	Benign Keratosis-like Lesions	1,099
df	Dermatofibroma	115
mel	Melanoma	1,113
nv	Melanocytic Nevi	6,705
vasc	Vascular Lesions	142

3.2 Tools and Environment

The implementation of this project was conducted using the following tools and platforms:
Programming Language: Python
Development Environment: Visual Studio Code and Jupyter Notebook
Libraries and Frameworks

- Data Handling: pandas, NumPy
- Visualization: Matplotlib, seaborn
- Machine Learning: scikit-learn

- Image Processing: OpenCV

These tools collectively enabled effective data exploration, model development, and performance visualization.

3.3 Data Preprocessing

Structured Metadata. Missing continuous values were imputed using the mean, while missing categorical values were imputed using the mode. Categorical variables such as sex and anatom_site_general_challenge were converted using one-hot encoding. Min-Max normalization was applied to numerical features.
Categorical features such as sex and anatom_site_general_challenge were converted using one-hot encoding. Feature scaling was applied using Min-Max normalization to improve model efficiency.

- Image Data: All dermoscopic images were resized to 224×224 pixels to provide a consistent input size for VGG16. The image intensities were normalized to the range [0, 1].
- Data augmentation was performed using flipping, rotation, and zooming to increase variation in the training images and reduce overfitting.

Class imbalance was examined from the dataset distribution. The binary distribution shown in Fig. 3 contains approximately 80.5% non-cancerous and 19.5% cancerous samples, indicating a substantial imbalance. The current experiments do not report a separate resampling or class-weighting procedure; this is therefore treated as a limitation when interpreting aggregate accuracy.

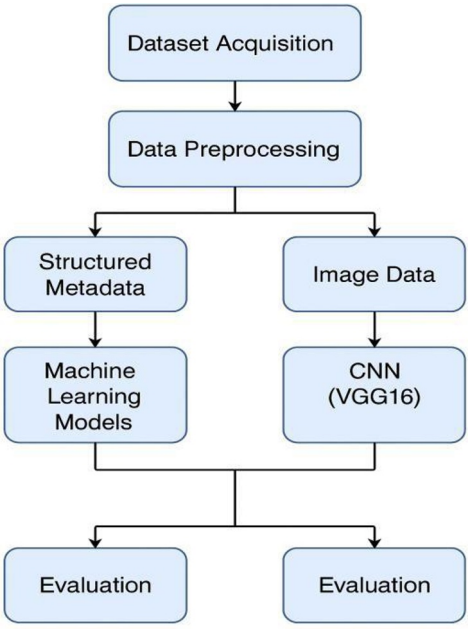

Fig. 1. Proposed Methodology block diagram

3.4 ML on Structured Data

The structured metadata was used to train three conventional supervised learning models: Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). The target was binary (0 = non-cancerous, 1 = cancerous). The data were divided into 80% training and 20% testing subsets. Accuracy, precision, recall, and F1-score were used for evaluation.

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)

The data was split into an 80:20 train-test ratio, and model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score.

3.5 CNN-Based Image Classification

A VGG16 model pre-trained on ImageNet was used as the image-classification pathway. The original top classification layers were removed and a custom dense layer with sigmoid activation was added for binary classification. The input size was 224×224 pixels. The Adam optimizer was used with a learning rate of 0.0001, batch size 32, and training for more than 20 epochs. Binary Cross-Entropy was used as the loss function. Training and validation accuracy/loss were monitored to assess convergence and possible overfitting. The manuscript does not specify a separate fine-tuning schedule or the number of unfrozen VGG16 layers; therefore, no such configuration is claimed.
Top classification layers were removed. A custom dense layer having sigmoid activation for binary classification was added. Learning rate of 0.0001 was used for optimization carried out with the Adam optimizer.
For model training with over 20 epochs having batch size of 32. Tensor Flow and Keras were used. For the loss function, Binary Cross-Entropy was deployed. For ensuring convergence and detecting over-fitting, Training and validation accuracy/loss were monitored.

3.6 Evaluation

For the structured-data models, accuracy, precision, recall, and F1-score were reported. The methodology also specifies confusion matrices and ROC-AUC analysis for model evaluation, but corresponding CNN numerical results and curves are not included in the present results section. Five-fold cross-validation was used during hyperparameter tuning. Because no fused multimodal experiment is reported, the structured and image pathways are discussed separately.

- Confusion metrices for ML and CNN models
- For CNN findings, ROC-AUC curves
- Cross-validation (k=5) to determine model robustness
- Comparative analysis of structured data classifiers and CNN for analyzing performance trade-offs

4 Results

The experiments were conducted using Python 3.9 in Jupyter Notebook and Visual Studio Code. HAM10000 contains 10,015 dermoscopic images distributed across seven lesion categories. For the reported structured-data experiment, metadata were preprocessed and evaluated using LR, DT, and RF. The image pathway used VGG16 transfer learning with the preprocessing described above. The two pathways were not fused, so the term "hybrid model" is not used for the reported results.

- Dataset Split Training Set: 80%
- Testing Set: 20%
- Cross-validation: 5-fold cross-validation was used for hyper-parameter tuning. The reported performance values in Table 2 are aggregate test-set metrics for the structured-metadata classifiers.

Table 2. Models Performance Matrix

Model	Accuracy (%)	Precision (%)	Recall (%)	F1Score (%)
Logistic Regression	83.45	81.20	80.75	80.97
Decision Tree (DT)	85,32	84.10	82.65	83.37
Random Forest (RF)	89.56	88.70	87.90	88.30

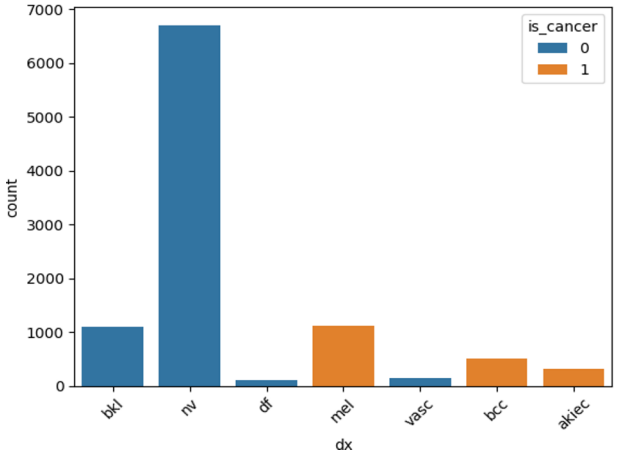

Fig. 2. Cancer Presence by Diagnosis

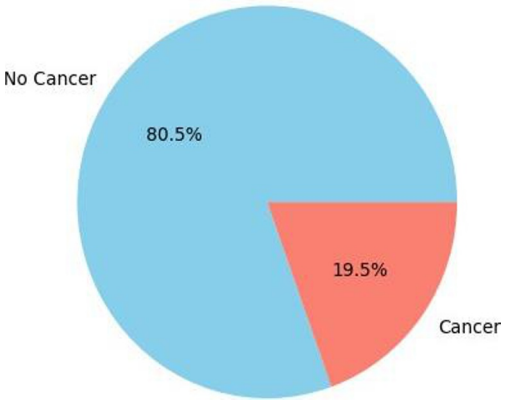

Fig. 3. Cancer Distribution

5 Result Analysis and Scope

5.1 Structured Metadata Results

The structured-data results are summarized in Table 2. RF achieved 89.56% accuracy, compared with 83.45% for LR and 85.32% for DT. RF also achieved the highest reported precision (88.70%), recall (87.90%), and F1-score (88.30%). These results support the conclusion that RF performed better than the other two structured-metadata classifiers evaluated in this study.

5.2 Image Pathway and Scope of Reported Results

VGG16 was included as the image-classification pathway, but the present manuscript does not contain its numerical predictions, confusion matrix, ROC-AUC value, or class-wise metrics. These quantities have therefore not been fabricated or estimated. Similarly, no feature-level or decision-level fusion result is reported. The discussion has been revised accordingly so that the 89.56% accuracy is explicitly attributed only to Random Forest on structured metadata.

5.3 Experimental Limitations

The dataset has a pronounced class imbalance, and the current manuscript does not document a dedicated resampling or class-weighting strategy. In addition, the available results do not provide the CNN metrics requested for a complete image-versus-metadata comparison. These limitations restrict the strength of claims about a hybrid multimodal system and are identified as directions for additional experimentation.

6 Conclusion

In this work, a machine-learning framework was examined using the HAM10000 dataset containing dermoscopic images and structured clinical metadata. The study separates the experimental pathways: LR, DT, and RF were applied to structured metadata, while VGG16 transfer learning was used for image classification. The reported quantitative results correspond to the structured-metadata experiment. RF achieved the highest reported accuracy of 89.56%, with precision of 88.70%, recall of 87.90%, and F1-score of 88.30%. These values indicate that RF performed better than LR and DT for the structured metadata evaluated in this study. The VGG16 pathway is retained as a complementary image-analysis component, but no numerical CNN performance or multimodal fusion result is claimed because such results are not included in the current experimental record. The dataset is also imbalanced, with approximately 80.5% non-cancerous and 19.5% cancerous samples; consequently, accuracy should be interpreted with this limitation in mind. The framework is intended to support, not replace, dermatological assessment. Future work should report complete CNN metrics, class-wise performance, confusion matrices and ROC-AUC, evaluate imbalance-handling strategies, and conduct an explicit multimodal fusion experiment if a genuinely hybrid model is to be claimed.

References

1. Sun, X., Zhang, M.: Exploring Atmosphere Creation of Light Environment Design in Visual Space. In: International Conference on Human-Computer Interaction. Cham: Springer Nature Switzerland (2023).
2. Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105 (2012).
3. Mahbod, A., Schaefer, G., Wang, C., Ecker, R., Ellinger, I.: Skin lesion classification using hybrid deep neural networks. Computer Methods and Programs in Biomedicine, vol. 182 (2020).
4. Signs and Symptoms of Melanoma Skin Cancer. American Cancer Society (2020). https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/signs-and-symptoms.html.
5. Tests for melanoma skin cancer. American Cancer Society (2020). https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/how-diagnosed.html
6. Takruri, M., Al-Hazaimeh, M.H., Al-Smadi, M.: Automated non-invasive multi-classifier system for melanoma skin cancer detection. International Journal of Advanced Computer Science and Applications, vol. 7(3), pp. 1–7 (2016).
7. Sundar, R.S.S., Vadivel, M.: Melanoma skin cancer classification using image processing and machine learning. Procedia Computer Science, vol. 85, pp. 539–546 (2016).

8. Krohling, B.: A smartphone-based application for skin cancer classification using deep learning. Sensors, vol. 21(2) (2021).
9. Myluca.ai: LUCA: AI-driven medical diagnostic system (2022) https://www.myluca.ai
10. Ahsan, M.M., et al.: Machine learning-based detection of monkeypox using modified VGG16 model. Future Internet, vol. 14(7) (2022).
11. Kanwal, T., et al.: Detection and localization of melanoma in histopathological whole-slide images using advanced image processing techniques. IEEE Access, vol. 11, pp. 1–12 (2023).
12. Hao, R., et al.: Automatic detection of skin hair follicles using Mueller matrix polarimetry. Biomedical Optics Express, vol. 14(3), pp. 1–15 (2023).
13. Sofie, M., Rofi'i, M., Wahyudi, B.: Classification of Skin Cancer Diseases Using KNN, CNN and SVM Methods. Journal of Engineering, Electrical and Informatics, vol. 5(2), pp. 14–26 (2025)
14. Gabani, V., Navamani, T.M., Shyamala, K., Vaswani Rajpal, V.K.: Multimodal skin lesion classification for early cancer diagnosis using deep learning. Frontiers in Physiology, vol. 17, p. 1717517 (2026).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

