Ensemble Models for Robust Skin Cancer Detection with 3D Total-Body Photographs

Yuehan He, Chunyu Li, Xiaoqian Zhou, Zongwei Huang, Mira Parikh
{mperform, chunyuli, zhouxq, zongweih, miparikh}@umich.edu

April 30 2025

Abstract

Skin cancer, particularly melanoma, remains among the deadliest cancers worldwide. Early detection is critical for improving patient outcomes. However, traditional diagnostic methods relying solely on visual examination by dermatologists are subjective and often inaccessible in underserved regions. To address these challenges, we propose an ensemble machine learning approach utilizing the ISIC 2024 dataset [16], which includes dermoscopic lesion images and patient metadata. Our methodology integrates convolutional neural networks (CNNs) (EfficientNet-B0[19], ResNet50[11], and DenseNet121[12]) for visual feature extraction, alongside gradient boosting models (XGBoost [5], LightGBM [13], CatBoost [8]) for structured metadata analysis. These two modalities are combined through different ensemble strategies. Performance is evaluated using the partial Area Under the Curve (pAUC) above 80% True Positive Rate (TPR), emphasizing high-sensitivity predictions.

1 Introduction

Skin cancer is among one of the most prevalent and deadly forms of cancer worldwide, with melanoma being the most aggressive type [7]. Early detection substantially improves survival rates, as prognosis deteriorates significantly in advanced stages. Traditional diagnostic methods mostly depend on visual examination by dermatologists and histopathological confirmation via biopsies. However, these approaches are time-consuming, subjective, and often inaccessible to individuals in underserved or remote areas [3].

The rapid advancement of machine learning, particularly deep learning, has transformed medical image analysis [4]. Large-scale dermatological images now enable automated skin lesion classification with accuracy approaching expert-level performance [10], potentially reducing the burden on healthcare systems and improving diagnostic efficiency. However, clinical diagnosis typically involves more than image-based assessment. Metadata—such as age, sex, and lesion location—provides vital contextual information that can further enhance diagnostic precision that can enhance diagnostic precision.

We aim to build a multimodal ensemble framework that combines image-based deep learning models with metadata-driven tree-based models. By leveraging both data modalities, our goal is to provide an effective and accessible solution for skin cancer detection that improves diagnostic reliability and supports clinical decision-making.

2 Related Work

Skin cancer detection has been extensively studied in the field of medical image analysis, with deep learning playing a pivotal role in recent advancements. Traditional methods relied on handcrafted features extracted from dermoscopic images, while modern approaches leverage deep neural networks for automated feature extraction and classification.

2.1 Traditional Methods

Early methods primarily relied on image processing and machine learning techniques to extract features such as asymmetry, border irregularity, color variation, and texture patterns [3]. These features were typically fed into classifiers like Support Vector Machines (SVMs), k-Nearest Neighbor (kNN), and Random Forests (RFs) for lesion classification.

For example, Alquran, et al. [2] developed a system using segmentation and feature extraction to process the dermoscopy images. SVM with a nonlinear classifier was utilized to distinguish benign and malignant lesions. Principle Component Analysis (PCA) was also employed to reduce the number of irrelevant features during feature extraction. Murugan, et al. [15] applied segmentation and feature extraction techniques including Watershed algorithm, ADCB rule (Asymmetry, Border, Color, Diameter), and Gray Level Co-occurrence Matrix (GLCM) extraction method, combined with kNN and RF classifiers, to classify melanoma cases.

While these methods achieved reasonable accuracy, their reliance on manually engineered features limited their generalizability. Additionally, methods such as SVM and kNN exhibit high computational complexity, making them challenging and unsuitable for large-scale datasets [4].

2.2 Deep Learning Methods

The rise of deep learning has significantly advanced skin cancer detection by enabling automatic feature extraction. Convolutional Neural Networks (CNNs) have become the dominant approach, achieving state-of-the-art performance on datasets like ISIC [6], HAM10000 [20], and PH2 [7]. CNN architectures such as ResNet [11], EfficientNet [19], and MobileNet have demonstrated high accuracy in skin lesions classifications [7].

Recent studies have further enhanced classification accuracy by finetuning pretrained CNN models—including InceptionV3 [18], VGG16 [17], and EfficientNet [19]—on dermatological datasets [21]. This approach leverages transfer learning to the need for large, labelled datasets while improving robustness.

2.3 Ensemble Methods

Ensemble learning has emerged as a promising strategy to boost classification performance and model robustness [4]. By combining multiple CNN architectures, researchers have achieved improved generalization, particularly for real-world applications where data variability is high [9, 21]. These approaches have also proven effective in leveraging unlabeled data and reducing prediction uncertainty.

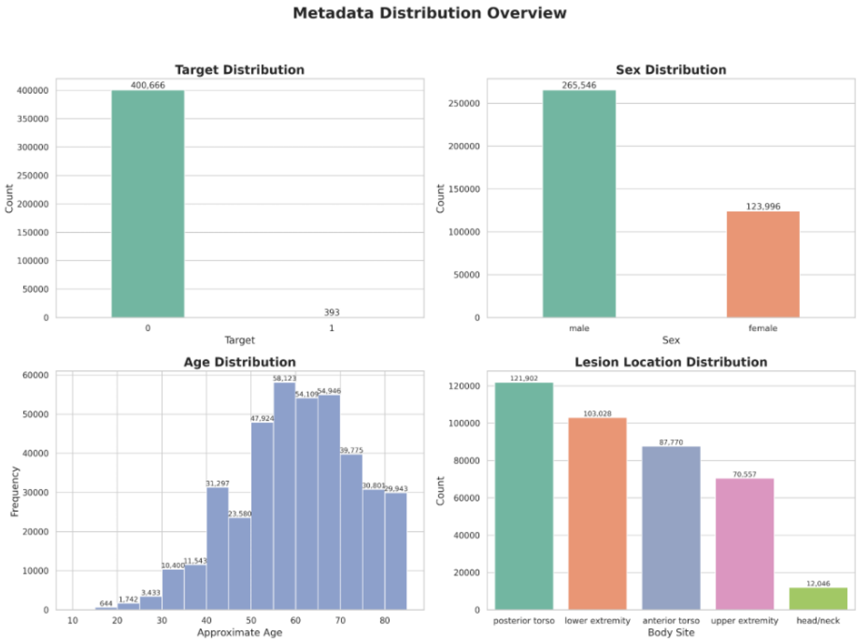

Figure 1: Distributions of a Selection of Features in Metadata

3 Dataset and Features

3.1 Dataset

The ISIC 2024 competition dataset [16] consists of dermoscopic lesion images and structured metadata. The lesion sizes range from 1.0mm to 28mm. Metadata includes patient demographics (e.g., age, sex, lesion-specific details (e.g., anatomical location), and additional carefully engineered image-derived features. The median patient age is 60 years old. The dataset was split into training (60%), validation (15%), and test (25%) sets.

3.2 Data Preprocessing

For metadata, features that would cause information leakage, such as thickness in depth of melanoma invasion and lesion confidence score, were removed. Irrelevant and uninformative features, such as image attribution and copyright license, were also removed. Missing data in age were imputed using kNN imputation with k=10.

For images, all samples were resized to 224x224 pixels, and the pixel values were normalized to [0, 1] for input to CNNs.

3.3 Data Augmentation

Out of the 401059 training samples, only 393 samples (approximately 0.1%) are malignant. This is insufficient and challenging for most neural networks to effectively learn the patterns in malignant samples.

To combat this issue, extensive image augmentation was employed, including random horizontal/vertical flips, rotations, and color augmentations such as contrast and saturation adjustments. These approaches boosted the malignant sample size to 2358. Additionally, malignant samples from external history datasets were incorporated to further diversify the malignant samples, increasing the number of malignant samples to 11597.

Other augmentation techniques, such as Variational Autoencoders and Gaussian Diffusion model with UNet architectures, were also explored. However, these results yielded suboptimal results due to the small sample size, and thus they were not included in the final training set.

4 Methods

4.1 Model Selection

Two classes of models have been trained on the structured metadata and image data separately. For metadata, we have chosen three tree-based models: (1) XGBoost, (2) CatBoost, and (3) LightGBM. They are all gradient boosting frameworks that build ensembles of decision trees, which are well-suited for handling tabular data. XGBoost is known for its robust regularization and efficient handling of missing or sparse data. CatBoost introduces techniques such as ordered boosting and native support for categorical features to reduce overfitting and improve generalizability. LightGBM is optimized for speed and memory efficiency, using histogram-based algorithms and leaf-wise tree growth to accelerate training while efficiently capturing complex feature interactions and maintaining accuracy.

Augmented images are trained by finetuning three CNN-based models: (1) ResNet50, (2) EfficientNet-B0, and (3) DenseNet121. These three models are widely adopted CNN architectures, each offering distinct advantages for image classification tasks. ResNet leverages residual connections to ease the training of deep networks and effectively capture hierarchical features, which would be essential for distinguishing subtle skin lesion patterns. EfficientNet introduces a compound scaling method to balance network depth, width, and resolution, achieving strong performance with relatively low computational cost. DenseNet promotes feature reuse through dense connectivity between layers, which helps strengthen gradient flow and capture fine-grained texture and shape details that may be critical for differentiating subtle lesion patterns.

4.2 Training and Hyperparameter Tuning

These six models (three tree-based, three CNN-based) were trained with the training set separately. For the tree-based methods, different sets of hyperparameters were selected for different models. These hyperparameters were tuned using the Optuna package [1]. Optuna is a flexible hyperparameter optimization framework designed to automate the search for optimal model configurations. It employs advanced techniques such as Bayesian optimization and pruning strategies to efficiently explore the hyperparameter space, reducing the computational cost.

The CNN-based models were finetuned with the binary cross entropy with logits loss. A weighting parameter was applied to the loss function to further address the class imbalance issue with the dataset by encouraging the network to pay more attention to the minority malignant class. The Adam optimizer [14] was adopted for training.

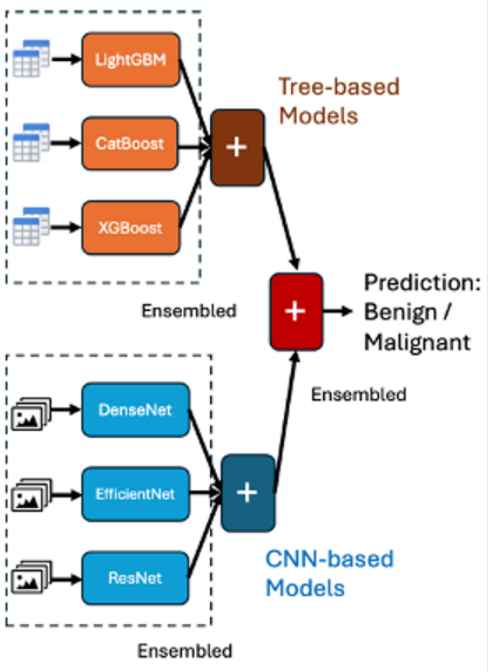

Figure 2: Schematic Diagram of One Possible Ensemble Model Structure

4.3 Ensemble Strategies

Once all the models were successfully trained, predicted probabilities were obtained on the validation set and test set for each model. Then three different ensemble strategies were adopted: (1) average ensemble, (2) weighted ensemble, and (3) stack ensemble.

For average ensemble, the final test predicted probabilities were simply the average of the test probabilities for each individual model. For weighted ensemble, a weight was assigned to each individual model, and this weight was treated as a hyperparameter and tuned on validation predicted probabilities using Optuna. For stack ensemble, a meta classifier was fit by treating the validation predicted probabilities as covariates and the validation target labels as response. Only two classifiers were considered: logistic regression and random forest. In our experiments, we had tried different combinations of model ensemble. Figure 2 illustrates a schematic diagram for one possible ensemble model architecture.

5 Results

To evaluate the model performance, the metric of partial AUC (pAUC) above 80% true positive rate (TPR) was adopted. The metric was chosen because it was the same metric used in ISIC 2024 competition, and it has a great practical significance. In clinical applications such as skin cancer detection, achieving high sensitivity (TPR) is of paramount importance. Missing malignant cases (i.e., false negatives) can lead to delayed diagnosis and severe outcomes, so the model's ability to maintain strong performance at high TPR is particularly critical. By focusing on the partial area under ROC curve above 80% TPR, the evaluation emphasizes how well the model performs in the

	Average Ensemble	Weighted Ensemble	Stacked Ensemble (Logistic Regression)	Stacked Ensemble (Random Forest)
XGBoost + CatBoost + LightGBM	0.1710	0.1724	0.1691	0.1012
ResNet + EfficientNet + DenseNet	0.1637	0.1616	0.1635	0.1330
XGBoost + DenseNet	0.1561	0.1617	0.1559	0.1157
All	0.1645	**0.1748**	0.1639	0.1520

Table 1: pAUC results for different ensemble strategies and model combinations. The highest pAUC score is bolded.

high-sensitivity regime.

Table 1 presents the performance of various ensemble strategies applied to different combinations of models using the pAUC metric. Note that several ensemble models of XGBoost and DenseNet were displayed on the third row, because XGBoost achieved the highest validation pAUC among all the tree-based models and DenseNet had the highest validation pAUC for all the CNN-based models.

Several important patterns can be observed. Firstly, the combination of tree-based models (XGBoost, CatBoost, LightGBM) generally yielded higher pAUC compared to CNN-based models (ResNet, EfficientNet, DenseNet), indicating that structured metadata plays an important role in the skin cancer detection task.

Secondly, the stack model approaches, particularly when using random forest as the meta learner, resulted in lower pAUC compared to average and weighted ensembles. This is particularly evident in the tree-based model ensembles alone, where random forest stack ensemble only achieved 0.1012 in pAUC, which is substantially lower than all other ensemble models. One possible explanation is that random forest stacking may have introduced additional complexity without effectively capturing the synergistic relationships among base models. In contrast, logistic regression stacking produced more competitive results, although still generally lower than simple averaging and weighted averaging.

Finally, when examining cross-modality combinations, the “All” category consistently outperformed single-modal ensembles (tree-based or CNN-based) in the weighted ensemble. This demonstrates that integrating predictions from both metadata and images leads to improved model performance.

6 Conclusion

In this project, we have investigated the performance of various machine learning models and ensemble strategies for skin cancer detection using both structured metadata and image data. For metadata, we adopted three tree-based models (XGBoost, CatBoost, and LightGBM), while for image data, we finetuned three well-established CNN architectures (ResNet50, EfficientNet-B0, DenseNet121). Each model was trained individually, with hypeparameters tuned using Optuna for tree-based models and class imbalance addressed through data augmentation and weighted binary cross-entropy loss in CNN training.

We explored multiple ensemble approaches, including average ensembles, weighted ensembles, and stacked ensembles with logistic regression and random forest as meta-learners. Evaluation was conducted using the pAUC above 80% true positive rate, reflecting the clinical priority of achieving high sensitivity in skin cancer detection.

The results revealed several key insights. Tree-based models trained on structured metadata generally achieved higher pAUC compared to CNN-based models, underscoring the importance of metadata in this task. Weighted ensemble methods, particularly those integrating both metadata and imaging modalities, achieved the highest overall performance, demonstrating the advantage of multimodal integration. This highlights the complementary nature of clinical metadata and imaging data in skin cancer prediction and reinforces the value of multimodal ensemble strategies.

While our current approach has demonstrated promising results, several avenues remain for further improvement. Although weighted ensembles proved effective, more sophisticated strategies, such as attention-based multimodal networks, could better exploit interactions between metadata and image features. In addition, the relatively poor performance of random forest-based stacking suggests room for improvement in meta-model selection. Future efforts could explore neural network-based meta-learners or gradient boosting frameworks for stacking, which may capture more subtle nonlinear relationships among base model predictions more effectively.

Author Contributions

X.Z. and Z.H. performed data preprocessing and conducted the literature review; C.L. and Y.H. developed the overall model pipeline and conducted the experiments; M.P. assisted with data preprocessing and experimental implementation; all authors contributed to the writing of this report.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *KDD*, pages 2623–2631, 2019.

[2] Hiam Alquran, Isam Abu Qasmieh, Ali Mohammad Alqudah, Sajidah Alhammouri, Esraa Alawneh, Ammar Abughazaleh, and Firas Hasayen. The melanoma skin cancer detection and classification using support vector machine. In *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5, 2017.

[3] Uzma Bano Ansari and Tanuja Sarode. Skin cancer detection using image processing. *Int Res J Eng Technol*, 4(4):2875–2881, 2017.

[4] Harsh Bhatt, Vrunda Shah, Krish Shah, Ruju Shah, and Manan Shah. State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review. *Intelligent Medicine*, 3(3):180–190, 2023.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *KDD*, pages 785–794, 2016.

[6] Noel C Codella, Veronica Rotemberg, Philipp Tschandl, and et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1710.05006*, 2018.

[7] Mehwish Dildar, Shumaila Akram, Muhammad Irfan, Hikmat Ullah Khan, Muhammad Ramzan, Abdur Rehman Mahmood, Soliman Ayed Alsaiari, Abdul Hakeem M Saeed, Mohammed Olaythah Alraddadi, and Mater Hussen Mahnashi. Skin cancer detection: A review using deep learning techniques. *International Journal of Environmental Research and Public Health*, 18(10):5479, 2021.

[8] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[9] Pratik Dubal, Sankirtan Bhatt, Chaitanya Joglekar, and Sonali Patil. Skin cancer detection and classification. In *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 1–6, 2017.

[10] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *CVPR*, pages 4700–4708, 2017.

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *NeurIPS*, 30, 2017.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] A. Murugan, S.H. Nair, and K.P.S. Kumar. Detection of skin cancer using svm, random forest and knn classifiers. *Journal of Medical Systems*, 43(269), 2019.

[16] Maura Gillis Kivanc Kose Walter Reade Nicholas Kurtansky, Veronica Rotemberg and Ashley Chow. All isic data 20240629. `https://www.kaggle.com/datasets/tomooinubushi/all-isic-data-20240629?select=metadata.csv`, 2024.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CVPR*, pages 2818–2826, 2016.

[19] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, pages 6105–6114, 2019.

[20] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018.

[21] M. Vidya and Maya V. Karki. Skin cancer detection using machine learning techniques. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–5, 2020.