433 Final Report

1. Abstract

Skin cancer remains one of the most critical global health concerns due to its potentially fatal consequences if left undetected. This project addresses the challenges of automated skin cancer detection using deep learning models trained on dermoscopic images and tabular metadata. The proposed system integrates image features extracted using ResNet50 with patient-specific metadata features selected through a multi-model feature selection process involving LightGBM, XGBoost, and CatBoost.

To tackle the severe class imbalance between benign and malignant cases, various strategies were employed, including data augmentation, weighted loss functions, and hyperparameter optimization. Multiple model configurations were tested, and the best-performing model achieved a recall of 54% for malignant cases with an overall accuracy of 94%. This result represents a significant improvement over previous models, emphasizing the effectiveness of multimodal data integration and advanced model training techniques. The system's promising performance highlights its potential for real-world clinical applications in dermatological diagnostics.

2. Introduction

Skin cancer, particularly melanoma, is one of the most severe and potentially fatal forms of cancer, responsible for a significant number of cancer-related deaths worldwide. Its prevalence has been steadily increasing due to factors such as prolonged ultraviolet (UV) radiation exposure, harmful chemical exposure, and environmental changes driven by shifting lifestyles. Early and accurate detection of skin cancer is essential, as it dramatically improves survival rates and minimizes the severity of treatment. However, diagnosing skin cancer accurately remains challenging, as traditional methods rely on manual examination of dermoscopic images by dermatologists. This process is time-consuming, resource-intensive, and susceptible to human error, particularly in regions with limited access to specialized healthcare professionals. These limitations underscore the urgent need for automated diagnostic solutions to supplement dermatological assessments.

Advancements in artificial intelligence (AI) and machine learning (ML) have emerged as promising approaches to addressing this issue. Convolutional Neural Networks (CNNs), a class of deep learning models, have achieved state-of-the-art performance in image analysis tasks by learning complex, non-linear patterns from large datasets. These models have demonstrated the ability to surpass human-level performance in specific medical applications, including skin lesion classification. Motivated by this potential, this project aims to develop a robust deep learning-based system capable of classifying skin lesions as benign or malignant. This effort seeks to enhance diagnostic accuracy, reduce the burden on healthcare systems, and improve patient outcomes globally by offering a scalable, automated diagnostic tool.

2.1. Challenges

Developing a reliable deep learning-based diagnostic system for skin cancer detection presents multiple challenges. One of the most critical issues is the severe class imbalance in the dataset, where malignant cases are significantly underrepresented compared to benign cases. This imbalance causes models to become biased toward the majority class, reducing sensitivity in detecting malignant lesions. False negatives in cancer detection pose a considerable risk, as they can lead to delayed treatment and worsen patient outcomes. Addressing this challenge requires thoughtful data preprocessing, augmentation strategies, and model design choices to ensure balanced learning.

Additionally, skin lesion images exhibit considerable variability due to differences in lesion size, shape, texture, and color, often influenced by patient-specific factors such as age, gender, and skin tone. This variability complicates the model's ability to generalize across diverse patient populations. Therefore, designing a model capable of capturing such heterogeneity is critical for ensuring clinical reliability and fairness in real-world applications.

Another major challenge arises from limited computational resources. The project relied on the free tier of Google Colab, which imposed constraints such as limited GPU access, capped memory, and restricted session durations. These limitations affected the scope of experimentation, reducing the dataset size, limiting the number of epochs, and restricting the ability to explore more complex model architectures or advanced training techniques. Overcoming these constraints required efficient resource management while maintaining model reliability and performance.

2.2. Proposed Solution

To address these challenges, this project employed a multifaceted approach that combines data augmentation, feature reduction, multimodal learning, and hyperparameter optimization. Addressing the class imbalance, data augmentation techniques were applied exclusively to the underrepresented malignant class using PyTorch. Transformations such as random rotations, flips, brightness adjustments, and zoom operations were used to create diverse synthetic samples, enhancing the model's ability to learn discriminative features for malignant lesions.

In parallel, the tabular metadata associated with the dermoscopic images was refined by reducing the number of features from approximately 60 to 20 through a feature selection process. This reduction minimized noise and enhanced the model's computational efficiency by retaining only the most relevant features. Models such as LightGBM, XGBoost, and CatBoost were employed to determine feature importances, with the top-ranked features forming the final dataset used during training.

To extract visual features, ResNet50 and EfficientNetB0 architectures were employed due to their strong performance in medical imaging tasks. These models were trained from scratch to learn dataset-specific features instead of relying on pre-trained weights, as transfer learning was found to be less effective given the unique nature of the dataset.

Additionally, a multimodal learning approach was implemented, combining extracted image features with the refined tabular metadata. This strategy enabled the model to leverage both visual and contextual information, enhancing its overall diagnostic performance. Key hyperparameters such as learning rate, batch size, and the number of epochs were carefully tuned through iterative experimentation to ensure model stability and optimal performance.

2.3. Earlier Results and Model Evolution

Initial experiments encountered considerable difficulties due to resource limitations and severe class imbalance. Early models trained on the full dataset suffered from significant overfitting, with near-perfect performance on the training set but zero sensitivity toward malignant cases in the testing phase. To address these issues, the dataset was reduced to 10,000 samples to ensure manageability, and advanced data augmentation techniques were employed to improve minority class representation.

Subsequent models demonstrated notable improvements, achieving a classification accuracy of 94% for benign cases. However, the sensitivity toward malignant cases remained low, highlighting the persistent challenge of handling class imbalance. Additional techniques, including weighted loss functions and cross-validation, were introduced to mitigate these issues, ultimately enabling the model to achieve a recall of 54% for malignant cases in the final iteration—a significant improvement over earlier attempts.

2.4. Related Works

The application of deep learning in medical imaging has been extensively studied, yielding promising results in skin cancer detection. Yadav and Jadhav (2019) highlighted the potential of CNN-based architectures for disease diagnosis, demonstrating superior performance in extracting and interpreting complex image features. Similarly, Dahou et al. (2023) achieved high diagnostic accuracy through transfer learning using architectures such as EfficientNet and ResNet, combined with novel optimization techniques.

Recent advances also include the Derm-T2IM framework, which addressed dataset limitations by generating synthetic data, improving model robustness and reducing overfitting. While these studies have explored individual aspects of deep learning in medical imaging, few works have effectively addressed the combined challenges of class imbalance, feature selection, and computational constraints.

This project builds on these previous studies by integrating data augmentation, multimodal learning, and model tuning into a cohesive framework tailored to the specific challenges of skin cancer detection. The proposed system represents a scalable and adaptive approach to automated skin cancer diagnosis, providing a valuable contribution to the growing body of research in AI-driven healthcare.

3. Methodology

The methodology employed in this project was structured to address the significant challenges of imbalanced data and computational constraints while ensuring the development of an effective deep learning-based diagnostic tool for skin cancer detection. This section provides a comprehensive explanation of the steps undertaken, from data preprocessing to model development and evaluation, highlighting the rationale behind each decision.

3.1. Data Collection and Preprocessing

The dataset for this project was sourced from a Kaggle challenge, containing a severe imbalance between benign and malignant cases, with 400,666 benign and only 393 malignant samples. This imbalance presented a significant challenge for training a model capable of accurately identifying malignant cases, as the model could easily overfit to the majority class while neglecting the minority class. Addressing this issue required several preprocessing steps designed to balance the data distribution and enhance the model's ability to learn meaningful features.

Two primary training datasets were derived from the original data to experiment with different strategies for handling class imbalance. The first dataset was created by undersampling the benign cases to achieve a training set of 8,000 benign and 315 malignant samples without data augmentation. The second dataset applied data augmentation techniques to increase the representation of malignant cases, resulting in a training set with 8,000 benign and 629 malignant samples. Augmentation included random rotations, flips, brightness adjustments, and zooming, generating diverse variations of malignant lesions while preserving their diagnostic characteristics.

The testing dataset, consistent across all experiments, contained 2,000 benign and 80 malignant samples. This separate test set ensured a fair evaluation of model performance and allowed for comparative analysis across experiments.

3.2. Feature Selection

A multi-model feature selection process was implemented using three state-of-the-art gradient boosting algorithms: LightGBM, XGBoost, and CatBoost. Each model was trained on the metadata, and feature importances were computed based on their contributions to the prediction task. The importance scores from all three models were averaged to mitigate biases from individual algorithms. The top 20 features with the highest average importance scores were selected, forming the final metadata feature set.

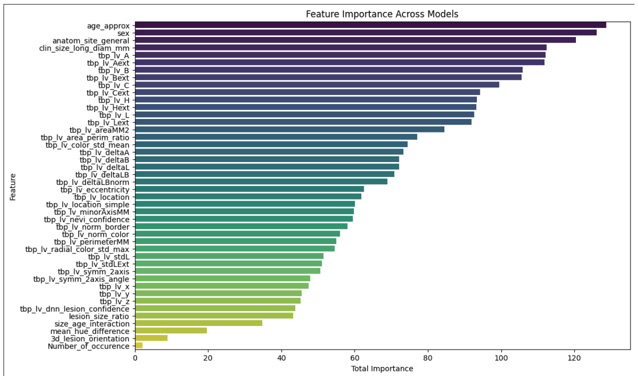

Figure 1. Feature importance across models

3.3. Model Selection and Design

Model selection was guided by the need for a scalable, accurate, and resource-efficient solution. Two deep learning architectures, ResNet50 and EfficientNetB0, were chosen based on their robust performance in medical image analysis tasks. Their feature extraction capabilities enabled precise classification despite limited computational resources.

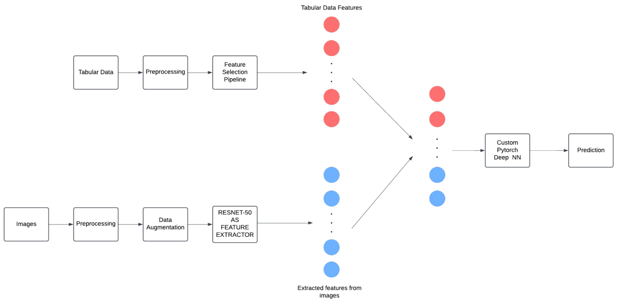

Figure 2. Full architecture

Three experimental setups were tested: 1. No data augmentation with a weighted loss function. 2. Data augmentation without a weighted loss function. 3. Both data augmentation and a weighted loss function.

The weighted loss function assigned higher penalties to misclassifying malignant cases, compensating for class imbalance. Each experimental setup employed 5-fold cross-validation, ensuring that models were trained and validated on different subsets of the data to improve generalizability.

Additionally, the models were trained from scratch rather than using pre-trained weights. This decision was influenced by the unique nature of dermoscopic images, which differ significantly from common datasets like ImageNet. Training from scratch allowed the models to learn domain-specific features tailored to the skin lesion dataset.

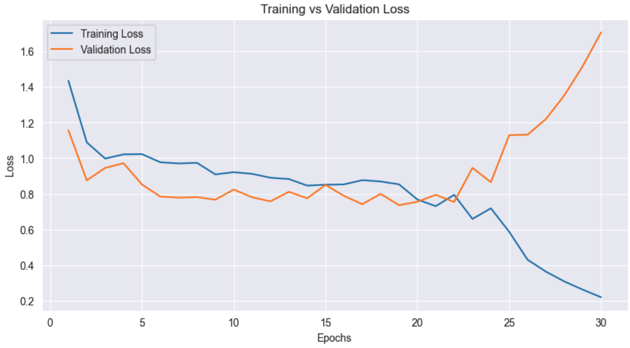

Figure 3. Training vs Validation Loss for Model 2 Iteration 2

The graph shows the training and validation loss trends for the model trained without data augmentation but with a weighted loss function. The training loss steadily decreased, indicating effective learning, while the validation loss stabilized around epoch 20, signaling the optimal stopping point to avoid overfitting. This configuration provided the best balance between learning and generalization, resulting in the most reliable model within the project's constraints.

3.4. Multimodal Data Integration

To enhance diagnostic accuracy, the project integrated metadata from tabular data with image features extracted using deep learning. Metadata features, such as age, lesion location, and lesion dimensions, were reduced to the top 20 features through the feature selection process.

Simultaneously, a 2048-dimensional feature vector was extracted from the final convolutional layer of ResNet50, representing dermoscopic image features. The two feature sets were concatenated into a unified vector, which served as input to a fully connected neural network for the final classification task. This multimodal integration allowed the model to leverage both visual and contextual information, improving its ability to detect malignant lesions.

3.5. Training and Optimization

Model training required balancing resource constraints with the need for optimal performance. Hyperparameters such as learning rate, batch size, and number of epochs were carefully tuned based on empirical testing and validation set performance.

The final hyperparameter configuration for the best model (Model 2 Iteration 2) was as follows: Learning Rate: 0.00001, selected to allow gradual weight adjustments and prevent unstable updates caused by larger learning rates. Batch Size: 128, chosen to stabilize gradient updates while maximizing GPU utilization. Epochs: 200, determined through early stopping-like observations from the training vs. validation loss graph, where performance gains plateaued beyond 200 epochs.

Additionally, weighted loss functions were employed to assign greater importance to malignant cases during training. Data augmentation and undersampling techniques further addressed class imbalance, reducing false negatives—a critical concern in cancer detection.

Performance metrics such as precision, recall, F1-score, and ROC-AUC were monitored during training. Since accuracy alone would be misleading due to class imbalance, recall for malignant cases was prioritized, given its significance in reducing missed cancer diagnoses.

3.6. Implementation and Reproducibility

The project was implemented using PyTorch for deep learning model development and scikit-learn for data preprocessing and performance evaluation. Matplotlib was used for generating visualizations of metrics and model performance.

Reproducibility was ensured by setting a fixed random seed (42) and saving model weights after each training session. These measures provided consistent results across multiple runs and allowed future researchers to replicate the experiments reliably.

3.7. Improvement and New Approaches

The project introduced several improvements over previous attempts. First, the multi-model feature selection process ensured a reliable and unbiased ranking of features, reducing noise and improving performance. Additionally, an extensive feature engineering process generated context-rich features such as lesion size ratios, age-size interaction, average hue difference, and three-dimensional lesion orientation.

The adoption of a multimodal architecture was the most notable innovation, as it combined image-based and tabular features, providing a comprehensive understanding of each case. Compared to earlier models that used only image features, this integrated system significantly improved recall and generalization.

Lastly, class imbalance was addressed through data augmentation, weighted loss functions, and undersampling. These strategies improved the model's ability to detect malignant cases while minimizing overfitting, resulting in a scalable and accurate diagnostic tool.

4. Results

This section provides a comprehensive analysis of the experiments conducted, focusing on the performance of various iterations of the trained models. Key results, including performance metrics, training and validation trends, and confusion matrices, are presented to evaluate and compare the models. The findings highlight the best-performing model and analyze its performance on both the training and testing datasets, while also comparing these outcomes to prior attempts.

4.1. Model Performance Across Iterations

After training and testing each model across eight iterations, the results were summarized in Table 1. The models varied in terms of learning rate (LR), batch size (BS), and number of epochs, allowing for a systematic evaluation of the impact of these hyperparameters on model performance. Model 1 iterations employed a higher learning rate of 0.001, with batch sizes of either 32 or 64, and epochs ranging from 25 to 100. These configurations struggled with overfitting and failed to generalize effectively, achieving recall values between 0.25 and 0.33 when detecting malignant cases.

Model 2 demonstrated significantly improved recall performance across iterations, particularly in Iteration 2, which used a learning rate of 0.00001, a batch size of 128, and 200 epochs. This setup achieved a test recall of 0.54, the highest among all iterations. Model 3 iterations also reached a recall of 0.54 in Iteration 2; however, further analysis showed that Model 2 Iteration 2 provided superior generalization due to more stable validation performance, as described below.

Model	LR	BS	Epoch	Test Recall
Model 1 Itr.1	0.001	64	100	0.32
Model 1 Itr.2	0.001	32	50	0.25
Model 1 Itr.3	0.001	64	25	0.33
Model 2 Itr.1	0.00001	128	100	0.46
Model 2 Itr.2	0.00001	128	200	0.54
Model 2 Itr.3	0.00001	64	200	0.24
Model 3 Itr.1	0.00001	64	200	0.25
Model 3 Itr.2	0.00001	128	200	0.54

Table 1. Model iterations with hyperparameters and test recall results.

4.2. Model Selection and Evaluation of Model 2 Iteration 2

Among all iterations, Model 2 Iteration 2 emerged as the best-performing configuration due to its high recall of 0.54 for the Positive class (malignant cases). This setup used a learning rate of 0.00001, a batch size of 128, and 200 epochs, enabling the model to learn effectively without overfitting. The training vs. validation loss graph (Figure 4) illustrates the stability of the model, with validation loss converging without significant divergence from the training loss, even at higher epochs.

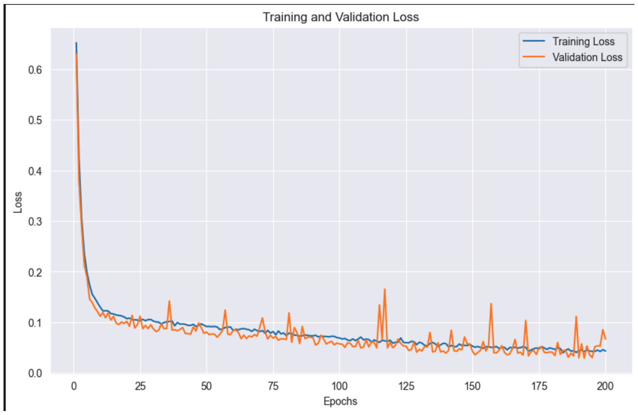

Figure 4. Training vs. validation loss for Model 2 Iteration 2.

The confusion matrix for Model 2 Iteration 2 during testing (Figure 5) highlights the model's classification performance. The model correctly identified 1585 benign cases (True Negatives) and 34 malignant cases (True Positives). However, 65 benign cases were misclassified as malignant (False Positives), and 29 malignant cases were missed (False Negatives). Despite these errors, the model achieved a recall of 54% for malignant cases, reflecting its effectiveness in identifying critical cases, a key priority for skin cancer diagnostics.

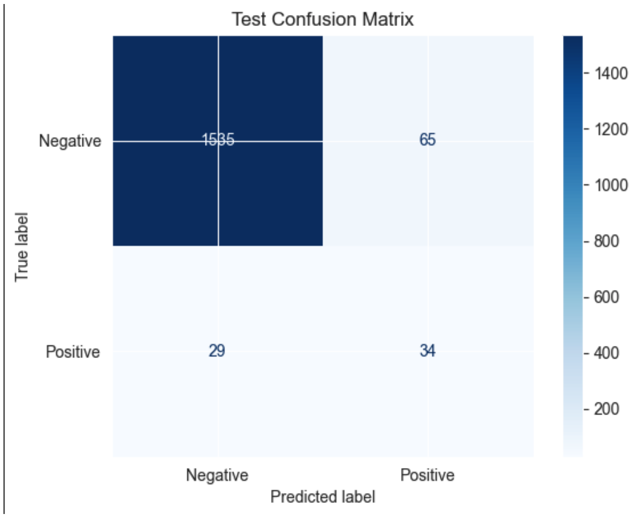

Figure 5. Confusion matrix for Model 2 Iteration 2 during testing.

4.3. Classification Metrics

Table 2 summarizes the classification metrics for Model 2 Iteration 2 on the test dataset. The Negative class (benign cases) exhibited high precision (0.98) and recall (0.96), contributing to an overall accuracy of 94%. The Positive class (malignant cases) achieved a precision of 0.34 and a recall of 0.54, resulting in an F1-score of 0.42. The macro-average recall was 0.75, indicating reasonable performance across both classes despite the severe class imbalance.

Class	Precision	Recall	F1-Score	Support
Negative	0.98	0.96	0.97	1600
Positive	0.34	0.54	0.42	63
Accuracy		0.94		1663
Macro Avg	0.66	0.75	0.70	1663
Weighted Avg	0.96	0.94	0.95	1663

Table 2. Classification metrics for Model 2 Iteration 2 during testing.

4.4. Discussion and Comparison to Previous Results

The results achieved in this study demonstrate significant improvements compared to previous models. The recall for malignant cases increased from under 40% in earlier attempts to 54%, highlighting the impact of weighted loss functions and multimodal data integration. This improvement reduced the number of false negatives—a crucial consideration given the life-threatening consequences of missing malignant cases.

The F1-score for the Positive class improved substantially as well, reflecting a better balance between precision and recall. This trade-off was deliberately accepted, as improving recall took precedence over precision in this context. While precision (0.34) for malignant cases remained moderate, it was considered an acceptable trade-off, ensuring the model was more sensitive to true malignant cases, which are more critical in medical diagnostics.

Additionally, the stability of validation loss throughout the training process reflected better generalization and a significant reduction in overfitting compared to earlier models. This improvement was driven by optimized hyperparameter selection, effective feature engineering, and multimodal integration.

4.5. Limitations and Future Directions

Despite the improvements, some limitations remain. The moderate precision for malignant cases could lead to a higher rate of False Positives, potentially causing unnecessary follow-up procedures in a clinical setting. Future work could explore ensemble models and synthetic data generation to address class imbalance further. Additionally, improving multimodal feature integration by incorporating attention mechanisms could enhance the system's ability to learn feature relationships more effectively.

References

[1] H. E. Kim et al. Transfer learning for medical image classification: A literature review. *Artificial Intelligence Review*, 22(1), Apr 2022.

[2] X. Zhao et al. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4), Mar 2024.

[3] O. Rainio, J. Teuho, and R. Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1), Mar 2024.

[4] J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue, 2018. Zenodo.

[5] D. Tellez, M. Balkenhol, I. Otte-Höller, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, and G. Litjens. Datasets digital pathology and artifacts, part 1, 2021. Zenodo.

[6] Animal faces, 2021. Kaggle, Available: https://www.kaggle.com/datasets/andrewmvd/animal-faces.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. arXiv.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of NeurIPS 2012*, 2012.

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of ICML 2015*, 2015.

[10] Writing resnet from scratch in pytorch, Nov 2024. DigitalOcean.

[11] Data classification using support vector machine, Nov 2024. ResearchGate.

[12] Z. Wang et al. Resnet for histopathologic cancer detection, the deeper, the better? *Journal of Data Science and Intelligent Systems*, 2(4):212–220, Mar 2023.