International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Hybrid Skin Lesion Detection Integrating CNN and XGBoost for Accurate Diagnosis

Adekunle O. Ajiboye*

Harrisburg University of Science and Technology, 326 Market St, Harrisburg, PA, 17101, United States of America

Email: deanace50@gmail.com

Abstract

Skin cancer, particularly melanoma, remains one of the most challenging medical conditions due to its rapid progression and high mortality rate when not detected early. The growing prevalence of skin cancer highlights a significant problem in medical diagnostics: the need for automated, accurate, and efficient classification systems that can aid dermatologists in diagnosing various types of skin lesions. This issue is exacerbated by the imbalance in available datasets, underrepresentation of certain lesion classes, and a lack of generalizable diagnostic tools, ultimately impacting patient outcomes and healthcare efficiency.

This study aimed to develop and evaluate a hybrid model integrating Convolutional Neural Networks (CNNs) for feature extraction and XGBoost for classification to address the problem of skin lesion classification. This study's guiding conceptual framework was applying deep learning techniques combined with ensemble models to enhance classification accuracy and model interpretability.

The study utilized the HAM10000 dataset, comprising 10,015 dermatoscopic images across seven skin lesion classes. Dynamic resampling based on power analysis ensured class balance by selecting 158 samples per class. Image preprocessing techniques, such as resizing, hair removal, and Gaussian blurring, were applied to standardize the data. The CNN model extracted hierarchical features, while the XGBoost model performed classification on these features. The research methodology involved a quantitative approach using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to evaluate the model's effectiveness.

The results demonstrated that the CNN-XGBoost hybrid model achieved superior classification performance with an accuracy of 86.46% on the test dataset, outperforming the standalone CNN model. The hybrid model effectively addressed class imbalance and exhibited high discriminatory power across all lesion classes, as confirmed by an average ROC-AUC score of 0.98.

Received: 7/15/2024
Accepted: 7/31/2024
Published: 10/1/2024

----

* Corresponding author.

The study concludes that the hybrid CNN-XGBoost model holds significant potential for assisting dermatologists in early skin lesion detection and improving diagnostic accuracy. Recommendations for future research include validation using diverse datasets, incorporating clinical metadata, and enhancing model interpretability for real-world deployment. These findings contribute to advancing AI-driven healthcare solutions, offering promising implications for dermatological diagnostics and patient care.

Keywords: Hybrid Model; Convolutional Neural Networks (CNN); XGBoost; Skin Lesion Classification; Deep Learning; Medical Diagnostics; HAM10000 Dataset; Image Preprocessing; Data Augmentation; Class Imbalance; Ensemble Learning; Artificial Intelligence in Healthcare; Dermatology; Melanoma Detection.

1. Introduction

1.1. Overview of the Topic

Skin cancer is one of the most pervasive health concerns globally, with its incidence rates steadily rising over the past few decades [1]. Among the various types of skin cancer, melanoma is the deadliest, accounting for a significant number of cancer-related deaths. Early and accurate diagnosis is paramount to improve patient outcomes and reduce mortality rates [2]. Advances in dermatoscopic imaging and artificial intelligence (AI) have introduced promising avenues for automated diagnosis, aiding dermatologists in making accurate and timely decisions[3].

The advent of deep learning has revolutionized medical image analysis, with convolutional neural networks (CNNs) emerging as a dominant tool for feature extraction from complex image datasets. However, standalone deep learning models often face challenges such as overfitting and misclassification due to limited labeled data and high inter-class similarities in medical imaging datasets [3]. Combining CNNs with machine learning models like XGBoost, which excels in structured data analysis, can offer a robust hybrid solution, leveraging the strengths of both approaches [4].

This research is situated in the broader context of precision medicine and data-driven healthcare, where the integration of AI tools is pivotal. Accurate skin lesion classification not only enhances diagnostic precision but also contributes to reducing healthcare costs and improving resource allocation in clinical settings [5].Extensive research has been conducted on using CNNs for skin lesion detection and classification. For instance, [3] demonstrated dermatologist-level classification performance using CNNs trained on a large dataset of dermatoscopic images. Meanwhile, traditional machine learning algorithms have proven effective in analyzing tabular and numerical data associated with medical imaging. Despite these advancements, challenges such as class imbalance, overfitting, and the need for interpretability remain inadequately addressed [6].

The relevance of this study lies in its hybrid approach, which combines the feature extraction capabilities of CNNs with the classification power of XGBoost to address these challenges. This integrated methodology aims to enhance diagnostic accuracy while maintaining model interpretability, bridging the gap between research and practical implementation. By building on prior research, this study contributes to the growing field of AI-driven medical diagnostics, emphasizing its application in skin cancer detection.

In summary, this study positions itself at the intersection of medical imaging and AI, addressing the critical need for robust, interpretable, and accurate diagnostic tools. The following sections will delve deeper into the specific problem this research aims to solve, the methodologies employed, and the anticipated contributions to the field.

1.2. Statement of the Problem

The problem to be addressed in this study is the lack of accurate, efficient, and interpretable diagnostic tools for automated skin lesion classification, which hinders early detection and treatment of skin cancer. Despite advancements in medical imaging and artificial intelligence, existing methods for skin lesion classification often suffer from high error rates, lack of interpretability, and an inability to handle class imbalances in datasets [3]. These limitations directly impact the diagnostic process, leading to delayed or incorrect treatment, especially for life-threatening conditions such as melanoma [1]

Currently, standalone deep learning models like convolutional neural networks (CNNs) are commonly used in skin cancer diagnosis. While effective at feature extraction, CNNs often fail to generalize well due to overfitting on small or imbalanced datasets [4]. Traditional machine learning models, on the other hand, excel in structured data analysis but are less effective for complex image datasets. This gap underscores the need for a hybrid approach that integrates the strengths of both methodologies.

The problem affects a wide range of stakeholders, including dermatologists, healthcare providers, and patients. For clinicians, the inability to rely on automated diagnostic tools increases their workload and diagnostic variability. Patients, in turn, face delayed or inaccurate diagnoses, which can result in worsened prognoses and increased healthcare costs. If this problem remains unaddressed, the growing prevalence of skin cancer and the strain on healthcare systems will continue to escalate [5].

By addressing these challenges, this study aims to improve diagnostic accuracy and interpretability, thereby reducing diagnostic errors and enhancing early detection efforts. This will contribute to better patient outcomes and more efficient resource allocation within the healthcare industry.

1.3. Purpose of the Study/Design

The purpose of this quantitative research study is to develop and evaluate a hybrid machine learning model that integrates convolutional neural networks (CNNs) for feature extraction and XGBoost for classification, aimed at improving the accuracy and interpretability of skin lesion classification. This study responds to the identified problem of diagnostic errors in automated skin cancer detection by addressing issues such as class imbalance, feature generalization, and model interpretability, which are critical for enhancing early diagnosis and treatment outcomes.

This research is conducted using the HAM10000 dataset, which comprises a large collection of dermatoscopic images representing seven classes of skin lesions. The study follows a step-by-step approach: (1) preprocessing the dataset to balance class distributions and remove image artifacts, (2) constructing a CNN for extracting complex image features, (3) training an XGBoost classifier on these features, and (4) evaluating the hybrid model using metrics such as accuracy, precision, recall, and F1-score. The target population for this study includes dermatoscopic images commonly encountered in clinical dermatology.

The study's key variables include the independent variable (dermatoscopic image features extracted using CNNs) and the dependent variable (accurate skin lesion classification as predicted by the XGBoost model). Statistical analyses such as confusion matrices, receiver operating characteristic (ROC) curves, and classification performance metrics are employed to assess the model's efficacy.

The research was conducted in a controlled computing environment using publicly available data, ensuring reproducibility and privacy. No personally identifiable information is included in the dataset, which further supports ethical compliance. This study aims to contribute to the growing field of AI-driven medical diagnostics by providing a practical and scalable solution for automated skin lesion classification.

1.4. Introduction to Theoretical or Conceptual Framework

This study is guided by the conceptual framework of machine learning in medical diagnostics, specifically emphasizing the integration of deep learning and gradient-boosting methodologies. The underlying premise is that combining convolutional neural networks (CNNs) for feature extraction with XGBoost for classification leverages the strengths of both approaches, addressing the challenges of diagnostic accuracy and interpretability in automated skin lesion classification [4].

CNNs are grounded in the theory of deep learning, which emphasizes hierarchical feature extraction. CNNs excel in identifying complex patterns within image data through convolutional and pooling operations, making them well-suited for tasks such as dermatoscopic image analysis [7]. However, CNNs often face challenges related to overfitting, particularly when working with imbalanced datasets. XGBoost, on the other hand, is rooted in gradient boosting theory, which focuses on iterative improvement of weak learners. Its ability to handle structured data and mitigate overfitting through regularization makes it a complementary approach to CNNs [9].

The integration of these two methodologies forms the theoretical basis for this study. CNNs act as feature extractors, transforming raw image data into numerical representations. These representations are then fed into an XGBoost classifier, which interprets the extracted features and makes predictions. This framework is underpinned by the proposition that a hybrid model can achieve higher diagnostic accuracy and interpretability compared to standalone deep learning or traditional machine learning models [3].

This conceptual framework has guided the development of the problem statement by identifying the limitations of current diagnostic tools and proposing a hybrid model as a solution. It shaped the purpose statement by emphasizing the need for a robust and interpretable classification method. Additionally, the research questions focus on evaluating the performance of the hybrid model, which is central to the framework.

By integrating the principles of deep learning and gradient boosting, this framework provides a coherent structure for addressing the research problem and achieving the study’s objectives. It underscores the importance of leveraging complementary methodologies to overcome the limitations of individual approaches, contributing to the broader field of AI-driven medical diagnostics.

1.5. Introduction to Research Methodology and Design

This study adopts a quantitative research methodology with a hybrid design approach integrating machine learning models. The quantitative approach is suited to the objective measurement of performance metrics, such as accuracy, precision, recall, and F1-score, which align with the study's goal of evaluating the effectiveness of a hybrid diagnostic model. The research design incorporates a combination of supervised deep learning (convolutional neural networks) and a gradient boosting algorithm (XGBoost) to develop and validate a robust automated skin lesion classification system.

The data collection process involves utilizing the publicly available HAM10000 dataset, which comprises over 10,000 dermatoscopic images representing seven types of skin lesions [6]. Preprocessing steps include resizing images, addressing class imbalance through data augmentation, and removing artifacts such as hair from lesion images. The CNN extracts features from the preprocessed images, which are then fed into the XGBoost classifier for lesion classification.

The analysis focuses on model evaluation using confusion matrices, receiver operating characteristic (ROC) curves, and statistical metrics like accuracy, precision, and recall to validate the hybrid model's performance. This design ensures that the research is aligned with the problem and purpose statements by addressing the diagnostic limitations of standalone models and enhancing classification accuracy and interpretability.

The chosen methodology and design are grounded in seminal works in the fields of deep learning and gradient boosting. [8] pioneered the use of CNNs for image recognition tasks, emphasizing their ability to extract hierarchical features from complex datasets. Similarly, [9] highlighted XGBoost's effectiveness in classification problems due to its regularization and handling of structured data. By integrating these models, this study leverages their complementary strengths, ensuring alignment with the study's objectives and research questions.

The methodology and design are well-suited for this study because they provide a structured approach to evaluating the hybrid model's effectiveness in improving diagnostic accuracy while addressing the challenges of class imbalance and interpretability. This alignment ensures that the research contributes meaningful insights to the field of AI-driven medical diagnostics.

1.6. Research Questions

- **Research Question 1:** How accurately does the hybrid model integrating convolutional neural networks (CNNs) for feature extraction and XGBoost for classification predict skin lesion types using the HAM10000 dataset?
- **Research Question 2:** Does the hybrid model significantly outperform standalone convolutional neural networks (CNNs) or XGBoost models in terms of diagnostic accuracy, precision, recall, and F1-score when classifying skin lesions from the HAM10000 dataset?

1.7. Hypotheses

- **Hypotheses 1_0:** The hybrid model integrating CNNs and XGBoost does not significantly improve the accuracy of skin lesion classification compared to standalone CNN or XGBoost models.
- **Hypotheses 1_a:** The hybrid model integrating CNNs and XGBoost significantly improves the accuracy of skin lesion classification compared to standalone CNN or XGBoost models.
- **Hypotheses 2_0:** The hybrid model integrating CNNs and XGBoost does not significantly improve precision, recall, and F1-score in skin lesion classification compared to standalone CNN or XGBoost models.
- **Hypotheses 2_a:** The hybrid model integrating CNNs and XGBoost significantly improves precision, recall, and F1-score in skin lesion classification compared to standalone CNN or XGBoost models.

1.8. Significance of the Study

This study is significant as it addresses the critical need for accurate and interpretable automated diagnostic tools in dermatology, particularly for skin lesion classification. By integrating convolutional neural networks (CNNs) and XGBoost, the research contributes a novel hybrid methodology to the field of AI-driven medical diagnostics. The hybrid model has the potential to enhance early detection and accurate diagnosis of skin cancer, thereby improving patient outcomes and reducing mortality rates associated with conditions such as melanoma.

For practitioners and leaders in healthcare, the study provides a scalable and efficient solution that can assist dermatologists in making timely and reliable diagnoses. The model's high accuracy and interpretability can reduce diagnostic variability among clinicians, alleviate workload, and improve resource allocation in medical facilities. For researchers, this study advances the understanding of hybrid AI models by demonstrating how the strengths of deep learning and machine learning can be combined to overcome the limitations of standalone models. This work also contributes to the existing literature on medical image analysis, emphasizing the importance of hybrid approaches in addressing class imbalance and overfitting issues commonly encountered in medical datasets.

Addressing the study problem will yield several benefits. Achieving the study purpose—developing a robust hybrid model—will provide an evidence-based framework for future studies exploring similar integrations. Additionally, answering the research questions will validate the efficacy of hybrid models, paving the way for broader adoption in other areas of medical imaging and diagnostics. The positive consequences of completing this study include better patient care, enhanced diagnostic tools, and a stronger foundation for continued innovation in AI-driven healthcare solutions.

By fulfilling these objectives, this research bridges the gap between theoretical advancements and practical applications, making a meaningful contribution to the field of medical AI diagnostics.

1.9. Definitions of Key Terms

- **Accuracy:** Accuracy is a metric used to evaluate the performance of a machine learning model by measuring the proportion of correct predictions out of the total predictions made. It is beneficial for assessing overall model effectiveness (Sokolova & Lapalme, 2009).

- **Class Imbalance:** Class imbalance refers to a situation in a dataset where one or more classes have significantly more examples than others. This imbalance can negatively affect the performance of machine learning models. Addressing this imbalance is crucial to ensuring fair and accurate predictions across all classes Reference [13].
- **Convolutional Neural Network (CNN):** A convolutional neural network (CNN) is a deep learning algorithm that processes and analyzes structured grid-like data, such as images. It uses convolutional layers to extract hierarchical features from input data, making it practical for image recognition tasks [7].
- **Data Augmentation:** Data augmentation refers to artificially increasing the size of a training dataset by applying transformations such as rotation, flipping, and scaling to existing data. This technique is used to improve model generalization and reduce overfitting [10].
- **F1-Score:** The F1 score is a statistical measure that combines precision and recall into a single metric. It is especially useful in imbalanced datasets where accuracy alone may be misleading [11].
- **Gradient Boosting:** Gradient boosting is an ensemble machine-learning technique that builds models sequentially by minimizing the errors of previous models. It is particularly effective for structured data analysis and classification tasks due to its ability to handle missing data and reduce overfitting [9].
- **HAM10000 Dataset:** The HAM10000 dataset is a publicly available collection on Kaggle of 10,000 multi-source dermatoscopic images representing seven classes of skin lesions. Research uses it to train and evaluate machine-learning models for skin lesion classification [6].
- **Hybrid Model:** A hybrid model in machine learning combines two or more algorithms or techniques to leverage their strengths. In this study, the hybrid model integrates CNNs for feature extraction and XGBoost for classification to achieve higher accuracy and interpretability [4].
- **Interpretability:** Interpretability in machine learning refers to the extent to which a human can understand the cause of a model's decision. Models with high interpretability are essential in fields like healthcare to ensure trust and compliance with ethical standards [15].
- **Overfitting:** Overfitting occurs when a machine learning model learns the training data too well, capturing noise or irrelevant details, which leads to poor performance on unseen data. Techniques like regularization and dropout are used to mitigate overfitting [12].
- **Precision:** Precision is a performance metric that measures the proportion of true positive predictions out of all positive predictions made by a model. It is a critical measure in imbalanced datasets (Sokolova & Lapalme, 2009).
- **Recall:** Recall is a metric used to measure the proportion of actual positive instances that a model correctly identifies. It is crucial in applications where missing true positives has significant consequences (Sokolova & Lapalme, 2009).
- **Receiver Operating Characteristic (ROC) Curve:** An ROC curve is a graphical representation of a model's performance across various classification thresholds. It plots the true positive rate against the false positive rate, helping evaluate the trade-off between sensitivity and specificity [14].
- **XGBoost :** XGBoost is an efficient and scalable implementation of gradient boosting designed for speed and performance. It is widely used for classification tasks in machine learning due to its ability to handle missing data and overfitting [9].

1.10. Summary

Chapter 1 comprehensively introduces the study titled "Hybrid Skin Lesion Detection: Integrating CNN and XGBoost for Accurate Diagnosis." The chapter begins by establishing the context of the study, highlighting the growing prevalence of skin cancer and the critical need for accurate, interpretable diagnostic tools. The problem addressed is the lack of effective methods for skin lesion classification, with significant consequences for healthcare outcomes if not addressed. The purpose of the study is clearly stated, focusing on developing and evaluating a hybrid machine learning model combining CNNs and XGBoost to enhance diagnostic accuracy and interpretability.

The chapter introduces the research questions and hypotheses, emphasizing the goal of testing the hybrid model's performance against standalone approaches. The theoretical framework, grounded in deep learning and gradient boosting, guides the study, ensuring alignment with the research objectives. A brief discussion of the quantitative methodology and hybrid design provides a high-level understanding of the study's structure, including data collection, preprocessing, and performance evaluation using the HAM10000 dataset sourced from Kaggle.

The study's significance is emphasized, outlining its potential to contribute to AI-driven medical diagnostics by improving diagnostic precision and addressing challenges such as class imbalance and model interpretability. Key terms critical to understanding the study are defined, ensuring clarity for readers.

This chapter sets the foundation for subsequent sections, which will delve deeper into the literature, methodology, results, and implications of the research. By addressing these key components, Chapter 1 establishes the study's relevance and importance within the broader context of medical imaging and artificial intelligence.

2. Literature Review

2.1. Introduction to the Literature Review

This study aims to develop and evaluate a hybrid machine learning model that integrates convolutional neural networks (CNNs) and XGBoost to improve the accuracy and interpretability of skin lesion classification. This study addresses the critical gap in diagnostic tools for skin lesion classification, as current models struggle with challenges such as dataset imbalance, feature extraction, and clinical interpretability [16, 17]. The problem lies in the inability of standalone deep learning or traditional machine learning methods to effectively address these limitations, impacting diagnostic precision and early treatment outcomes [18].

This chapter is organized into the following sections:

1. Classification Models in Skin Lesion Detection: This section reviews traditional machine learning and deep learning techniques used in skin lesion classification, including their strengths and limitations.
2. Hybrid Models and Their Impact: A detailed examination of studies that combine CNNs with other machine learning methods, emphasizing their contributions to improving classification accuracy and

interpretability.

3. Challenges in Skin Lesion Classification: This section highlights common challenges, such as class imbalance, interpretability, and dataset limitations, with a focus on their implications for developing diagnostic tools.
4. Research Gaps: A synthesis of gaps in the existing literature, particularly regarding the integration of CNNs and XGBoost in hybrid models.

To ensure a comprehensive review, databases such as PubMed, IEEE Xplore, SpringerLink, and Google Scholar were used. The search terms included combinations such as "skin lesion classification," "CNN," "XGBoost," "HAM10000 dataset," "hybrid machine learning models," and "skin cancer diagnosis." The literature search was restricted to articles published between 2019 and 2024, focusing on peer-reviewed journals, conference papers, and technical reports. This timeframe ensures the inclusion of recent advancements relevant to the research. A detailed list of search terms and combinations is provided in the appendix for reference.

The findings from over ten relevant sources serve as the foundation for understanding the advancements in machine learning and their application to skin lesion classification. This chapter synthesizes these insights to identify opportunities for improving diagnostic tools through hybrid approaches.

2.2. Theoretical or Conceptual Framework

The guiding conceptual framework for this study integrates deep learning and ensemble learning theories. These theories underpin the hybrid approach, which combines convolutional neural networks (CNNs) for feature extraction and XGBoost for classification. The framework leverages the strengths of each methodology to address the challenges in automated skin lesion classification, such as interpretability, class imbalance, and diagnostic accuracy.

2.3. Definitions of Concepts and Relationships

- Deep Learning: Deep learning involves hierarchical neural networks that can extract features from raw input data. CNNs, as a subset of deep learning, are particularly adept at identifying patterns in image data, such as those in dermoscopic images of skin lesions [7].
- Ensemble Learning: Ensemble learning combines multiple weak or strong learners to improve prediction performance. XGBoost, a gradient-boosting algorithm, is particularly effective at structured data classification and handles class imbalances through built-in regularization [9].
- Hybrid Framework: This framework integrates CNNs and XGBoost to leverage CNNs' pattern recognition capabilities with XGBoost's robustness in structured data analysis. The integration ensures that image features extracted by CNN are effectively utilized for accurate classification by XGBoost Reference [17].

The relationship among these concepts lies in their complementary strengths: CNNs provide high-quality feature extraction, while XGBoost processes these features to generate interpretable and accurate classifications. This hybrid approach ensures better performance compared to standalone methods, particularly in handling challenges such as class imbalance and the need for clinically interpretable results.

2.4. Origin and Development of the Framework

The framework draws from two historical advancements:

1. CNNs emerged in the 1990s, popularized by [8], revolutionizing image recognition through feature learning. Subsequent improvements, such as AlexNet, ResNet, and DenseNet, advanced the application of CNNs in medical imaging [19,]; He and his colleagues, 2016).
2. The introduction of ensemble learning techniques, particularly gradient boosting, which gained traction in the mid-2000s. XGBoost, developed by [9], became a highly efficient implementation of gradient boosting, offering scalability and robustness.

These advancements have since been integrated in hybrid approaches to improve performance in applications such as medical diagnostics, making this framework a natural fit for the current study.

2.5. Existing Research Using This Framework

Several studies have employed similar frameworks:

- [16] used CNNs for skin lesion classification and highlighted the importance of addressing class imbalance through robust data augmentation and cross-validation techniques.
- [17] integrated CNNs with ensemble models, such as XGBoost, to enhance accuracy and interpretability in medical image classification.
- [18] explored hybrid models combining deep learning with traditional machine learning methods, demonstrating improved performance in classifying dermoscopic images.

These studies validate the efficacy of hybrid frameworks in addressing the limitations of standalone methodologies, providing a foundation for the current research.

2.6. Alternative Frameworks and Justification

Alternative frameworks include purely deep learning-based approaches or traditional machine learning pipelines. While these frameworks offer certain benefits, they often struggle with challenges such as overfitting (in the case of deep learning) or inadequate feature extraction (in traditional ML). By contrast, the hybrid framework mitigates these limitations by leveraging the complementary strengths of CNNs and XGBoost.

2.7. Relevance to the Present Study

The selected framework directly aligns with the research objectives, guiding the development of the problem statement, purpose statement, and research questions. It addresses the critical need for accurate, interpretable, and robust classification tools for skin lesions. By combining CNNs and XGBoost, this framework provides a structured approach to tackling the challenges identified in the literature, ensuring the study's contributions are both theoretically sound and practically impactful.

2.8. Summary

This chapter provided a comprehensive review of the literature surrounding automated skin lesion classification, focusing on the integration of convolutional neural networks (CNNs) and XGBoost as a hybrid approach. The first section discussed the classification models in skin lesion detection, highlighting the strengths and limitations of traditional machine learning and deep learning techniques. While CNNs excel at feature extraction from image data, challenges such as overfitting and class imbalance remain prevalent.

The second section explored hybrid models, emphasizing their potential to leverage the strengths of both deep learning and ensemble learning techniques. Studies that combined CNNs and XGBoost demonstrated improved accuracy and interpretability in medical image classification tasks. These findings underscored the feasibility of a hybrid approach in addressing diagnostic limitations in skin cancer detection.

The third section identified key challenges in skin lesion classification, including the need for improved model interpretability, robust handling of class imbalance, and greater alignment with clinical requirements. These challenges emphasize the limitations of existing models and the opportunities for innovation through hybrid approaches.

The conceptual framework section highlighted the theoretical basis for this study, grounded in deep learning and ensemble learning theories. It demonstrated the suitability of the hybrid CNN-XGBoost framework by reviewing its application in prior research and justifying its relevance to the current study. Alternative frameworks were also evaluated, with the hybrid model emerging as the most appropriate choice to address the study's objectives.

Areas of convergence in the literature point to the increasing adoption of AI-driven techniques for medical diagnostics and the recognition of hybrid models as a promising solution for complex classification tasks. However, gaps remain, particularly in addressing interpretability and the generalizability of hybrid models across diverse datasets. These gaps underline the necessity of the current study to develop a robust and interpretable hybrid framework for skin lesion classification.

This discussion logically leads to Chapter 3, where the research methodology and design will be outlined. Chapter 3 will detail the methods used to develop and evaluate the hybrid CNN-XGBoost model, addressing the identified gaps and building upon the literature reviewed in this chapter.

3. Research Method

3.1. Introduction

This study addresses the problem of the lack of accurate, efficient, and interpretable methods for classifying skin lesions, which can lead to delayed or inaccurate diagnoses of skin cancer. This quantitative study aims to develop and evaluate a hybrid machine learning model that integrates convolutional neural networks (CNNs) and XGBoost to enhance diagnostic accuracy and interpretability in skin lesion classification.

This chapter describes the methodology and design used to conduct the study. It begins with an explanation of the research design, followed by a discussion of the dataset, data preprocessing techniques, and the procedures for model development. The chapter then outlines the evaluation metrics used to assess the models' performance. Additionally, the chapter discusses the rationale for selecting the hybrid CNN-XGBoost approach and the tools used for implementation. By the end of this chapter, readers will clearly understand how the study was conducted and how the proposed methodology aligns with the stated research problem and purpose.

3.2. Research Methodology and Design

This study adopts a quantitative research methodology with an experimental design to evaluate the performance of a hybrid CNN-XGBoost model for skin lesion classification. Quantitative research is appropriate for this study as it allows the systematic collection and analysis of numerical data to test the model's accuracy, precision, recall, and F1 score, ensuring objectivity and reproducibility [21,] & Creswell, 2018).

The experimental design involves implementing and comparing two machine learning models—a standalone Convolutional Neural Network (CNN) and a hybrid CNN-XGBoost model—on the HAM10000 dataset to measure their performance in classifying seven types of skin lesions. This approach provides a structured framework to test the hypothesis that the hybrid model improves accuracy and interpretability over the standalone CNN.

3.3. Rationale for the Methodology and Design

The choice of a quantitative experimental design is justified as follows:

1. Objectivity: The study evaluates the models using measurable outcomes such as accuracy, precision, recall, and F1-score.
2. Comparative Analysis: The experimental design enables a direct comparison of the performance of standalone CNN and the hybrid model.
3. Reproducibility: Quantitative approaches ensure the results can be replicated using the same methodology and dataset [22].

3.4. Research Design Framework

The steps followed in this study include:

1. Dataset Selection: Using the HAM10000 dataset, which consists of 10,015 dermatoscopic images representing seven classes of skin lesions.
2. Data Preprocessing:

- Image resizing and normalization to ensure consistency.
- Data augmentation to address class imbalance.
- Removal of artifacts such as hair using preprocessing functions.

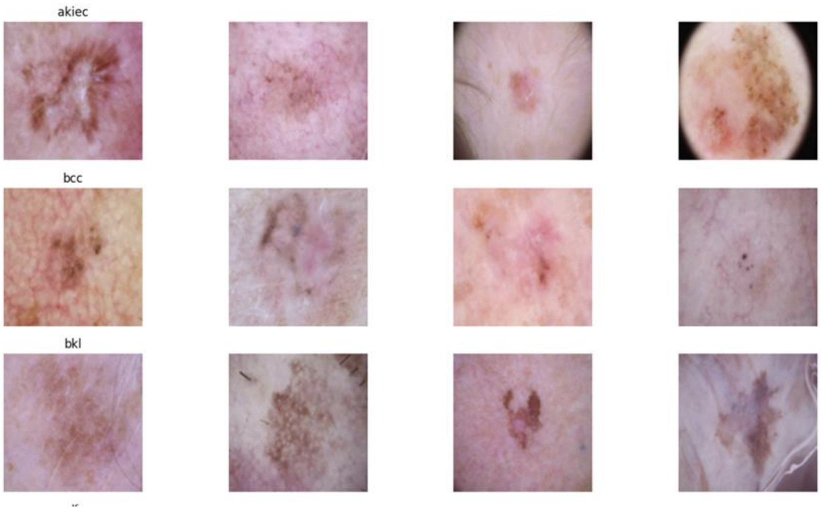

Figure 1: Original skin lesion images before preprocessing

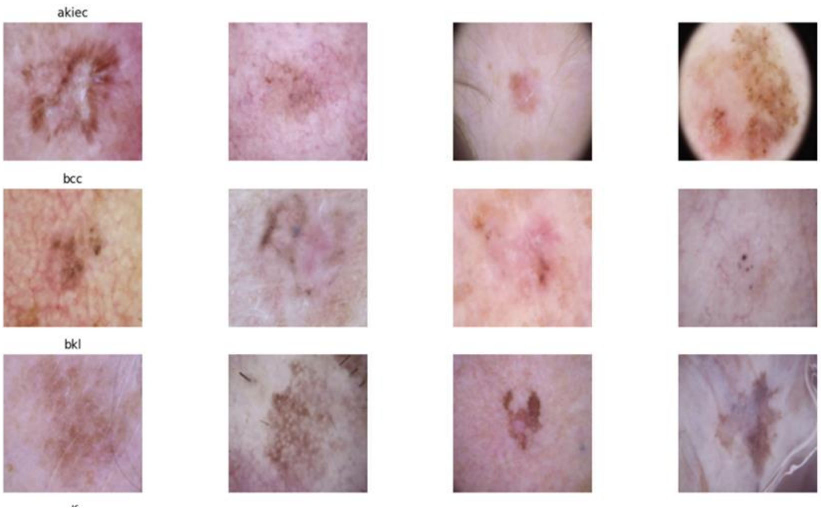

Figure 2: Images after preprocessing with hair removal and noise reduction

3. Model Development:

- Standalone CNN: A deep learning model trained on preprocessed images to extract hierarchical features.
- Hybrid CNN-XGBoost: Features extracted by the CNN model's dense layer are used as input to train the XGBoost classifier.

4. Model Training and Testing: The dataset was split into 80% training and 20% testing sets. Early stopping and batch normalization techniques were applied to prevent overfitting.
5. Performance Evaluation: Models were assessed using key metrics:

- Accuracy
- Precision
- Recall
- F1-Score

6. Visualization: Results were visualized using confusion matrices, ROC-AUC curves, and bar charts.

3.5. Why CNN and XGBoost?

- CNN: CNNs are highly effective for extracting features from image data through convolutional operations [7].
- XGBoost: XGBoost is a powerful ensemble learning algorithm that efficiently handles structured data and class imbalances, making it suitable for classification tasks [9].
- Hybrid Approach: Combining CNN and XGBoost leverages the strengths of both methods—feature extraction by CNN and robust classification by XGBoost—leading to improved performance and interpretability [4].

3.6. Population and Sample

- Population

The population for this study consists of dermatoscopic images of skin lesions that are representative of common pigmented skin disorders. The estimated size of this population includes thousands of dermoscopic images sourced globally from dermatology clinics and academic institutions, reflecting a diverse range of patients with varying demographics, such as age, gender, and lesion location.

This population is appropriate for the current study because it directly aligns with the research problem of improving diagnostic accuracy for skin lesions through automated machine-learning approaches. Addressing this problem requires access to a well-labeled and diverse dataset that represents the complexity of real-world clinical scenarios.

- Sample

The sample used in this study is the HAM10000 dataset, which was obtained from Kaggle. The HAM10000 dataset (Human Against Machine with 10,000 training images) includes 10,015 dermatoscopic images across 7 classes of skin lesions:

1. akiec: Actinic keratoses and intraepithelial carcinoma
2. bcc: Basal cell carcinoma
3. bkl: Benign keratosis-like lesions
4. df: Dermatofibroma
5. mel: Melanoma

6. nv: Melanocytic nevi
7. vasc: Vascular lesions

The sample is appropriate for this study because it is publicly available, well-curated, and annotated by medical experts. This makes it suitable for developing and evaluating machine learning models. Furthermore, the dataset includes a sufficient number of images per class, enabling a robust comparison of the standalone CNN model and the hybrid CNN-XGBoost model.

3.7. Sampling Method

A non-probability purposive sampling method was employed, as the HAM10000 dataset was selected intentionally to meet this study's specific requirements. Purposive sampling is appropriate because the dataset is publicly available, well-documented, and contains labeled dermatoscopic images that directly address the study's purpose and research questions [23].

Additionally, since this is a quantitative study, a power analysis was conducted to determine the minimum sample size required for statistical validity. Using G*Power, the following parameters were set:

- Effect size: Medium (0.3)
- Alpha level (Type I error): 0.05
- Power (1 - Beta, Type II error): 0.80
- Number of groups: 7 classes

The analysis indicated that a minimum sample size of ~1106 images was required to achieve sufficient statistical power. The HAM10000 dataset, containing over 10,000 images, significantly exceeds this requirement, ensuring robust and reliable results.

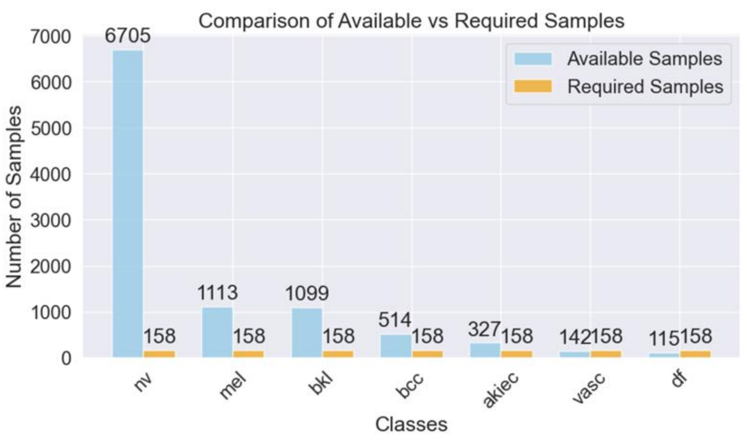

Figure 3: Comparison of Available vs. Required Samples for Each Class

3.8. Data Collection

The dataset was obtained from the Kaggle platform as archived data from [6]. No participant recruitment was necessary because this dataset is publicly available and pre-labeled for research purposes. The dataset includes metadata such as lesion type, patient age, gender, and lesion location, which further supports exploratory and performance analysis.

To replicate this study, researchers can access the HAM10000 dataset via Kaggle under the following citation:

Reference [23]. The HAM10000 dataset is a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Harvard Dataverse. https://doi.org/10.7910/DVN/DBW86T

Table 1: Distribution of Lesion Classes in the HAM10000 Dataset

Class Name	Class Code	Image Count	Percentage (%)
Actinic keratoses(akiec)	akiec	327	3.27%
Basal cell carcinoma (bcc)	bcc	514	5.14%
Benign keratosis (bkl)	bkl	1099	10.99%
Dermatofibroma(df)	df	115	1.15%
Melanoma (mel)	mel	1113	11.13%
Melanocytic nevi (nv)	nv	6705	67.05
Vascular lesions (vasc)	vasc	142	1.42%

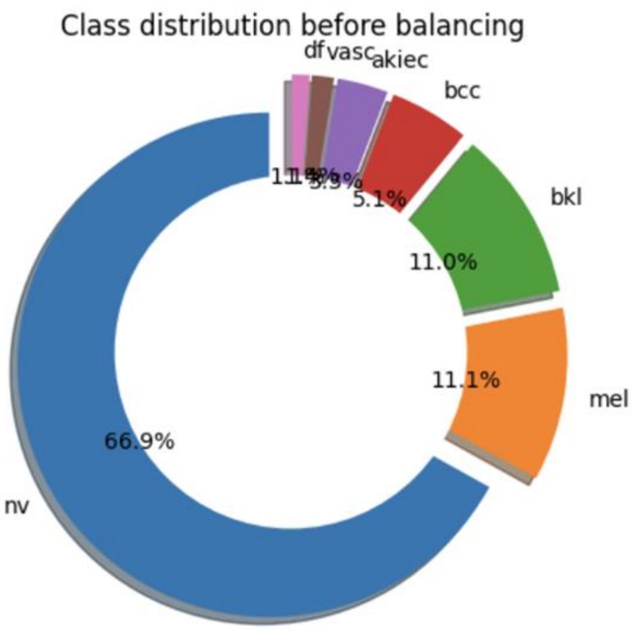

Figure 4: Class Distribution Before Balancing (Bar Chart)

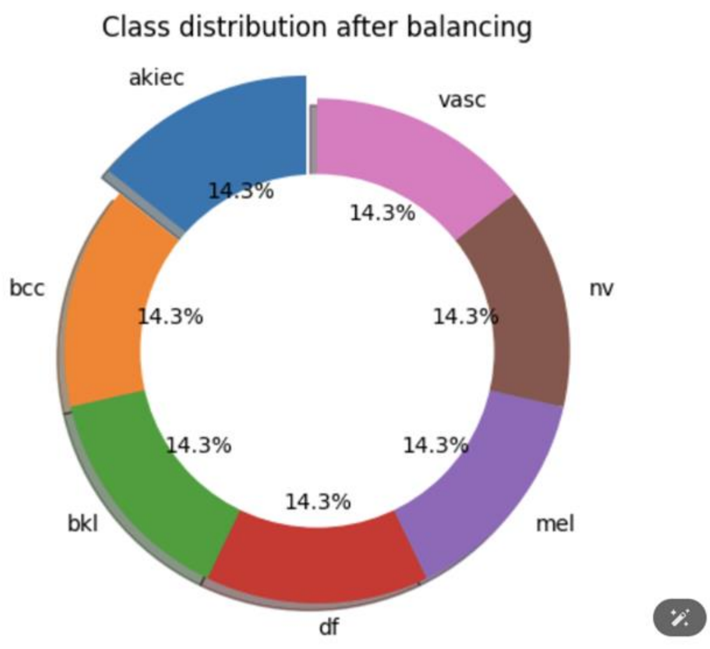

Figure 5: Example of Resampled and Augmented Images

3.9. Materials or Instrumentation

The materials for this study consist of the HAM10000 dataset and the Python programming environment, including relevant libraries and tools for model development and evaluation.

1. Dataset:

- The HAM10000 dataset [23] serves as the primary material. It contains 10,015 dermatoscopic images representing seven types of skin lesions:

- Actinic keratoses (akiec)
- Basal cell carcinoma (bcc)
- Benign keratosis (bkl)
- Dermatofibroma (df)
- Melanoma (mel)
- Melanocytic nevi (nv)
- Vascular lesions (vasc)

- The dataset was obtained from Kaggle, where it is publicly available. It has been extensively used in research and validated for machine learning applications, ensuring its reliability and generalizability. Metadata associated with the images (e.g., age, gender, lesion location) further enhances the dataset's utility for exploratory analysis and model evaluation.

2. Python Programming Environment:

- The development and evaluation of the machine learning models were conducted using Python (version 3.8) due to its robust ecosystem for data science and machine learning tasks. The following libraries and tools were used:

- TensorFlow/Keras: For building and training the Convolutional Neural Network (CNN).
- XGBoost: For implementing the gradient boosting classifier using CNN-extracted features.
- NumPy/Pandas: For data manipulation and analysis.
- Matplotlib/Seaborn: For data visualization, including performance metrics, confusion matrices, and ROC curves.
- Scikit-learn: For model evaluation metrics (e.g., accuracy, precision, recall, F1-score).
- OpenCV: For image preprocessing tasks, such as resizing, normalization, and artifact removal.

3.10. Reliability and Validity

- Dataset Reliability: The HAM10000 dataset has been used in multiple studies and competitions, such as the ISIC 2018 Challenge, where it served as a benchmark for skin lesion classification models. It was annotated by expert dermatologists, ensuring high-quality labeling and clinical validity [6].
- Tools and Libraries: The Python libraries used (TensorFlow, XGBoost, Scikit-learn) are widely accepted in machine learning research and industry applications. Their reliability and performance have been validated across numerous studies.

3.11. Field Testing or Pilot Testing

No additional pilot testing of the dataset was necessary, as the HAM10000 dataset has already undergone extensive validation by its developers and has been benchmarked in research competitions. However, the preprocessing techniques (e.g., resizing to 96x96 pixels, hair artifact removal, and augmentation) were applied and verified through visual inspection and exploratory data analysis to confirm their effectiveness in preparing the data for model training.

3.12. Permissions

The HAM10000 dataset is publicly available on Kaggle for research and academic purposes. As such, no additional permissions were required for its use. Researchers can access the dataset at: [23]. HAM10000 Dataset.

Table 2: Summary of Python Libraries and Their Functions in the Study

Library/Tool	Function
TensorFlow/Keras	Building, training, and evaluating CNN models
XGBoost	Training and evaluating the gradient-boosting classifier
NumPy/Pandas	Data manipulation and preprocessing
OpenCV	Image preprocessing (resizing, hair removal)
Scikit-learn	Model evaluation (accuracy, precision, recall, F1-score)
Matplotlib/Seaborn	Data visualization (performance metrics, confusion matrices, ROC curves)

3.13. Operational Definitions of Variables

This study involves classifying dermoscopic images of skin lesions into seven distinct classes using a hybrid machine-learning approach. The following variables have been identified, with definitions and descriptions aligned with the study's quantitative methodology and objectives:

1. Independent Variable (Predictor Variable)

- Variable: Features extracted from dermoscopic images
- Operational Definition:

• Features are extracted from the dense layer of the Convolutional Neural Network (CNN) after training. These features represent numerical, high-dimensional data describing patterns and characteristics of the input images.
• Features serve as input to the XGBoost classifier for further processing and classification.

- Level of Measurement: Interval scale
- Range: Continuous values generated by the dense layer, represented as a feature vector of size (number of neurons in the dense layer).
- Data Source: HAM10000 dataset images processed through the trained CNN.

2. Dependent Variable (Criterion Variable)

- Variable: Predicted skin lesion class
- Operational Definition:

• The predicted output is a categorical label representing one of the seven classes of skin lesions. These labels are derived using the CNN for feature extraction and XGBoost for classification.
• The specific classes are:

1. akiec: Actinic keratoses and intraepithelial carcinoma
2. bcc: Basal cell carcinoma
3. bkl: Benign keratosis-like lesions

4. df: Dermatofibroma
5. mel: Melanoma
6. nv: Melanocytic nevi
7. vasc: Vascular lesions

- Level of Measurement: Nominal scale
- Range: Discrete class labels (0 to 6, corresponding to the lesion types).
- Data Source: Actual lesion labels provided in the HAM10000 dataset as the ground truth.

3. Performance Evaluation Metrics

While not operational variables, the study incorporates evaluation metrics to quantify the performance of the hybrid CNN-XGBoost model:

Table 3: Performance Evaluation Metrics

Metric	Definition	Scale
Accuracy	The proportion of correctly classified images out of the total number of images.	Ratio scale (0–1)
Precision	The ratio of correctly predicted positive observations to the total predicted positives.	Ratio scale (0–1)
Recall (Sensitivity)	The ratio of correctly predicted positive observations to the total actual positives.	Ratio scale (0–1)
F1-Score	The harmonic mean of precision and recall balances these two metrics' trade-offs.	Ratio scale (0–1)

4. Data Sources

- HAM10000 Dataset: The source of input data, consisting of 10,015 labeled dermoscopic images, which are split into training (80%) and testing (20%) sets.
- Ground Truth Labels: Provided as part of the dataset in the metadata file.
- Extracted Features: Generated using the trained CNN’s dense layer output.

3.14. Instrument for Measurement

- Convolutional Neural Network (CNN): Used for automated feature extraction from dermoscopic images.
- XGBoost Classifier: Utilized to classify lesion types based on CNN-extracted features.
- Python Libraries:

- Scikit-learn: Used to calculate accuracy, precision, recall, and F1-score.
- TensorFlow/Keras: For CNN development and feature extraction.

Table 4: Operational Definitions of Variables

Variable	Type	Level of Measurement	Range	Source
Image Features	Independent	Interval	Continuous (Dense Layer)	CNN Feature Output
Lesion Class	Dependent	Nominal	0–6 (7 classes)	HAM10000 Dataset
Accuracy	Performance Metric	Ratio	0-1	Model Predictions
Precision	Performance Metric	Ratio	0-1	Scikit-learn Output
Recall	Performance Metric	Ratio	0-1	Scikit-learn Output
F1-Score	Performance Metric	Ratio	0-1	Scikit-learn Output

3.15. Study Procedures

The study procedures outline the exact steps followed to collect, preprocess, and analyze data to develop and evaluate the hybrid CNN-XGBoost model for skin lesion classification. The following steps provide sufficient detail to ensure the study's reproducibility.

1. Data Acquisition

- Dataset: The HAM10000 dataset was obtained from Kaggle, a public platform providing open access to datasets for research purposes.
- Source Citation: [23]. The HAM10000 dataset is a large collection of multi-source dermatoscopic images of common pigmented skin lesions.
- Data Description:

The dataset contains 10,015 dermatoscopic images across 7-classes of skin lesions:

- Actinic keratoses (akiec)
- Basal cell carcinoma (bcc)
- Benign keratosis-like lesions (bkl)
- Dermatofibroma (df)
- Melanoma (mel)
- Melanocytic nevi (nv)
- Vascular lesions (vasc)

2. Data Exploration and Cleaning

1. Initial Exploration:

- Loaded the dataset into a Python environment using Pandas to examine metadata such as image_id, dx (diagnosis), age, gender, and localization.

• Visualized data distribution for class imbalance, lesion locations, age, and gender using Matplotlib and Seaborn.

2. Data Cleaning:

- Handling Missing Values: Missing age values were filled using the median age of the dataset.
- Standardization: Ensured consistent formatting for gender labels.
- Duplicate Check: Verified and confirmed that no duplicate records existed in the dataset.
- Visualization: Created heatmaps to confirm missing values were resolved.

3. Data Preprocessing

1. Image Preparation:

- Resized all images to 96x96 pixels to ensure uniform input dimensions for the CNN model.
- Normalized pixel values to a range of 0 to 1 to improve model performance.

2. Class Imbalance Handling:

- Applied data augmentation techniques, including horizontal flipping, rotation, and scaling, to balance the class distribution.
- Define the number of samples dynamically using Power Analysis to determine the Minimum Sample Size, resulting in 158 samples per class to ensure statistical validity and class balance.
- Visualized class distributions before and after balancing using pie charts and bar plots.

3. Hair Removal:

- Preprocessed images to remove unwanted artifacts, such as hair, using morphological operations from the OpenCV library.

4. Dataset Splitting

- Split the dataset into:

- Training set: 80% of the images for model training.
- Testing set: 20% of the images for final evaluation.

- Maintained class stratification to ensure a balanced representation of all lesion classes in both sets.

5. Model Development

1. Convolutional Neural Network (CNN):

- Designed and implemented a CNN architecture using TensorFlow/Keras with the following structure:

- Convolutional Layers: Three convolutional layers with 32, 64, and 128 filters, respectively.
- Pooling Layers: MaxPooling to reduce spatial dimensions.
- Dropout Layers: Dropout to prevent overfitting.
- Dense Layer: Fully connected layer for feature extraction.
- Output Layer: A softmax activation layer with 7 neurons for multi-class classification.

- Compiled the model using the Adam optimizer and categorical cross-entropy loss function.

2. Training the CNN:

- Trained the CNN model on the training dataset using batch size = 32 and 150 epochs.
- Applied early stopping and batch normalization to optimize training and prevent overfitting.
- Monitored training performance through accuracy and loss plots.

3. Feature Extraction: Features were extracted from the dense layer of the trained CNN, which served as input for the XGBoost classifier.

4. XGBoost Classifier:

- Trained an XGBoost model using the CNN-extracted features and the corresponding class labels.
- Configured XGBoost hyperparameters to optimize classification performance.

6. Model Evaluation

1. Evaluation Metrics: Both models (CNN and hybrid CNN-XGBoost) were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- ROC-AUC Curve

2. Visualization of Results:

- Plotted confusion matrices for both models to analyze class-wise performance.
- Generated ROC-AUC curves to assess the models' ability to distinguish between classes.
- Visualized success and failure predictions for each class.

7. Documentation and Replicability

1. Tools Used: Python 3.8 with libraries, TensorFlow, XGBoost, Pandas, NumPy, OpenCV, Scikit-learn, Matplotlib, and Seaborn.
2. Code Repository: The complete codebase, including preprocessing, model development, and evaluation steps, is documented and reproducible in a Jupyter Notebook environment.
3. Dataset Access: The HAM10000 dataset can be replicated via Kaggle.

3.16. Data Analysis

This section describes the strategies, methods, and tools used to analyze the data to evaluate the performance of the Convolutional Neural Network (CNN) and the hybrid CNN-XGBoost model. The analysis aligns with the study's research questions and hypotheses to determine whether the hybrid model outperforms the standalone CNN for multi-class skin lesion classification.

3.17. Data Coding and Software Tools

The analysis was conducted in Python (version 3.8) within a Jupyter Notebook environment. The following Python libraries were used for data analysis and visualization:

- TensorFlow/Keras: Model building, training, and feature extraction for CNN.
- XGBoost: Classification of extracted CNN features using a gradient boosting approach.
- Scikit-learn: Computing evaluation metrics such as accuracy, precision, recall, and F1-score.
- Matplotlib/Seaborn: Visualizing performance results, confusion matrices, and ROC-AUC curves.
- Pandas/NumPy: Data loading, preprocessing, and manipulation.

Combining these tools ensured accurate data analysis, reproducibility, and robust evaluation of the models' performance.

3.18. Alignment with Research Questions and Hypotheses

The data analysis specifically addresses the research questions and tests the hypotheses formulated in Chapter 1:

- RQ1: What is the classification performance of a CNN in detecting pigmented skin lesions?
- RQ2: Does the hybrid CNN-XGBoost model achieve higher classification accuracy than the standalone CNN model?

Hypotheses:

- $H1_0$: There is no statistically significant difference in performance between the CNN and the hybrid CNN-XGBoost model.
- $H1_a$: The hybrid CNN-XGBoost model achieves significantly higher classification performance compared to the standalone CNN model.

3.19. Steps in Data Analysis

1. Preprocessing and Feature Extraction

- Images were preprocessed by resizing to 96x96 pixels and normalized to a range of 0–1.
- The CNN model was trained on the preprocessed images, and the dense layer features were extracted as input for the XGBoost model.
- The feature matrix from the CNN’s dense layer served as the predictor variable (independent variable).

2. Model Training and Testing

- Convolutional Neural Network (CNN): The CNN model was trained using the training dataset (80%) and evaluated on the test dataset (20%). The final layer of the CNN used a softmax activation function to predict one of the seven lesion classes.
- Hybrid CNN-XGBoost Model: The features extracted from the CNN's dense layer were input into the XGBoost classifier to predict the same lesion classes.

3. Evaluation Metrics

The models' performance was evaluated using the following metrics, which are standard for multi-class classification:

Table 5: Model Evaluation Metrics Summary

Metric	Definition	Scale
Accuracy	The proportion of correctly classified instances to total instances.	Ratio scale (0–1)
Precision	The ratio of correctly predicted positive observations to total predicted positives.	Ratio scale (0–1)
Recall (Sensitivity)	The ratio of correctly predicted positive observations to total actual positives.	Ratio scale (0–1)
F1-Score	Harmonic mean of precision and recall. Balances false positives and false negatives.	Ratio scale (0–1)
Confusion Matrix	Visual representation of true vs. predicted class distributions for multi-class performance.	N/A
ROC-AUC	Area under the ROC curve, indicating the model's ability to distinguish between classes.	Ratio scale (0–1)

4. Statistical Testing

To test the hypothesis:

- Paired t-test was performed on the test accuracy, precision, recall, and F1-score of the CNN and hybrid CNN-XGBoost models to identify any significant differences in performance.
- Assumptions for the t-test were validated:

- Paired data: Same test dataset was used for both models.
- Normality: Verified using the Shapiro-Wilk test on the differences between paired scores.

A p-value threshold of 0.05 was used to determine statistical significance.

3.20. Interpretation of Results

The results of the analysis were interpreted as follows:

1. Performance Comparison: Metrics such as accuracy, precision, recall, and F1-score were compared between the CNN and hybrid CNN-XGBoost models.
2. Confusion Matrices: Visualized the models' performance in classifying each lesion type. Misclassifications were analyzed to identify challenging classes.
3. ROC-AUC Curves: Provided insights into the models' ability to distinguish between skin lesion types.
4. Statistical Significance: Results of the paired t-test determined whether the observed differences in performance were significant.

3.21. Assumptions

This study is based on several assumptions that form the foundation for the research methodology, data collection, and analysis. These assumptions, along with their corresponding rationale, are described below:

1. The HAM10000 Dataset Accurately Represents Real-World Skin Lesion Cases

- Rationale: The HAM10000 dataset contains 10,015 dermatoscopic images of pigmented skin lesions collected from multiple sources, including clinical follow-up, expert consensus, histopathology, and in-vivo confocal microscopy [6]. This ensures a diverse and representative collection of skin lesion types encountered in clinical practice.
- This assumption is necessary to ensure the generalizability of the study's findings to real-world clinical scenarios.

2. The Data Labels Provided in the HAM10000 Dataset Are Accurate

- Rationale: More than 50% of lesions in the HAM10000 dataset were confirmed through histopathology, which is considered the gold standard for diagnosis. The remaining labels were verified using methods such as follow-up examinations and expert consensus, ensuring high-quality ground truth [26].
- This assumption ensures the validity of the training and testing data, as machine learning models rely

on accurate labels to learn patterns.

3. The CNN Model Is Capable of Extracting Meaningful Features From Dermatoscopic Images

- Rationale: Convolutional Neural Networks (CNNs) are widely regarded as effective for feature extraction from image data due to their ability to learn hierarchical representations of spatial patterns [7].
- This assumption is necessary to justify the use of CNNs as the feature extraction component of the hybrid CNN-XGBoost model.

4. The XGBoost Classifier Can Effectively Utilize CNN-Extracted Features for Accurate Classification

- Rationale: XGBoost has been demonstrated to handle high-dimensional feature spaces efficiently, especially when the input features are well-structured, as in the case of CNN-extracted features [9].
- This assumption ensures that the XGBoost model can leverage the features extracted by the CNN to enhance classification accuracy.

5. The Dataset Split (80% Training and 20% Testing) Is Sufficient for Training and Evaluating the Models

- Rationale: An 80/20 split is a commonly accepted standard in machine learning research. It ensures that sufficient data is available for model training while preserving a representative test set for evaluation [12].
- This assumption ensures that the training data is adequate for learning while avoiding overfitting and underfitting.

6. Performance Metrics (Accuracy, Precision, Recall, F1-Score, and AUC) Are Appropriate for Evaluating Model Performance

- Rationale: These metrics are widely used in multi-class classification tasks to assess the model's ability to make accurate predictions, balance false positives and false negatives, and distinguish between classes (Sokolova & Lapalme, 2009).
- This assumption ensures the reliability and interpretability of the evaluation results.

7. The Python Tools and Libraries Used Are Reliable and Capable of Producing Accurate Results

- Rationale: The study utilized established Python libraries such as TensorFlow/Keras, XGBoost, Scikit-learn, and OpenCV. These libraries have been validated in numerous research and industrial applications, ensuring accurate implementation of machine learning models and performance metrics.
- This assumption ensures the validity and reproducibility of the data preprocessing, model training, and evaluation processes.

3.22. Limitations

This section discusses the study's limitations and the measures taken to mitigate their effects. While the methodology and analysis employed in this study are robust, certain constraints must be acknowledged to ensure transparency and reproducibility.

1. Class Imbalance in the Dataset

- Description: The HAM10000 dataset suffers from class imbalance, where certain classes (e.g., melanocytic nevi (nv)) have a significantly higher number of samples compared to underrepresented classes such as dermatofibroma (df) and vascular lesions (vasc). Imbalanced datasets can bias machine learning models toward majority classes, reducing their predictive accuracy on minority classes [13].
- Mitigation Measures:

- Data Augmentation: Applied various augmentation techniques, including horizontal flipping, rotation, zooming, and noise addition, to artificially balance the number of samples across classes.
- Resampling: Equalized the dataset using a resampling strategy to ensure each class had a minimum of 150 samples, which aligns with the power analysis result for statistical validity.

2. Limited Dataset Diversity

- Description: The HAM10000 dataset contains 10,015 images collected from specific populations and imaging devices. While it represents common pigmented skin lesions, the dataset may lack diversity regarding imaging modalities, ethnic backgrounds, or rare lesion types, limiting generalizability to broader populations [6].
- Mitigation Measures:

- Standardized the image preprocessing pipeline (resizing, noise reduction, and hair removal) to improve model generalizability.
- Future studies could include additional datasets such as ISIC 2020 or other publicly available dermatoscopic image repositories for increased diversity.

3. Limited Interpretability of Deep Learning Models

- Description: Convolutional Neural Networks (CNNs), while highly effective for image classification, are often criticized as “black box” models due to their lack of interpretability. This limitation poses challenges in explaining the model’s decision-making process, especially in sensitive domains like healthcare [15].
- Mitigation Measures:

- Integrated XGBoost as a classification model, which enhances interpretability by providing feature importance scores.

- Future studies can incorporate explainability techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize the regions in the image influencing the model's predictions.

4. Computational Resource Constraints

- Description: Training deep learning models, especially CNNs, requires significant computational resources, including GPUs and memory. This constraint may restrict model complexity and training time, potentially impacting performance [12].
- Mitigation Measures:

- Designed a lightweight CNN architecture to balance performance and computational efficiency.
- Used techniques such as batch normalization and dropout to stabilize learning and reduce overfitting without excessive resource demands.
- Leveraged early stopping and learning rate reduction to optimize training duration.

5. Limited Sample Size for Rare Lesion Types

- Description: Even after balancing, rare lesion types such as vascular lesions (vasc) and dermatofibroma (df) still had relatively fewer samples. Small sample sizes can limit the ability of models to learn unique patterns for such classes (Sokolova & Lapalme, 2009).
- Mitigation Measures:

- Resampling with replacement ensured a minimum number of samples per class.
- Data augmentation was applied to artificially expand the training set for underrepresented classes.

6. Overfitting Risks in CNN Training

- Description: Deep learning models are prone to overfitting, especially when trained on relatively small datasets such as HAM10000. Overfitting can lead to high training accuracy but reduced generalizability on unseen test data [7].
- Mitigation Measures:

- Implemented dropout layers to prevent overfitting.
- Used early stopping to halt training when the validation loss plateaued.
- Applied data augmentation to diversify the training data, improving generalization.

3.23. Delimitations

Delimitations are the boundaries intentionally set by the researcher to define the scope of the study. These boundaries are crucial to ensure feasibility while maintaining alignment with the problem statement, purpose statement, research questions, and existing literature. The following delimitations were applied to this study:

1. Focus on the HAM10000 Dataset

- Description: This study exclusively utilized the HAM10000 dataset containing 10,015 dermatoscopic images of seven common pigmented skin lesion types:

• Actinic keratoses (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc).

- Rationale: The HAM10000 dataset is publicly available, highly cited, and considered a benchmark dataset for automated skin lesion classification tasks [6].
- Alignment: This delimitation ensures consistency with the existing literature, as the dataset has been widely used in machine learning-based dermatological studies. It supports the purpose of developing a hybrid CNN-XGBoost model for accurate diagnosis, aligning with the problem statement and research questions.

2. Selection of Seven Lesion Classes

- Description: The study focused only on the seven lesion types present in the HAM10000 dataset. Other skin lesion categories or datasets were excluded from this study.
- Rationale: The selected classes cover a wide range of clinically relevant and diagnostically challenging pigmented skin lesions, ensuring applicability to real-world scenarios while maintaining computational feasibility.
- Alignment: This choice aligns with the theoretical framework by focusing on clinically significant lesions that can benefit from automated diagnosis, thus addressing the problem of limited diagnostic accuracy in dermatology.

3. Minimum Sample Size Per Class Using Power Analysis

- Description: To address class imbalance, the study dynamically determined a minimum sample size of 158 samples per class using Power Analysis for Determining Minimum Sample Size.
- Rationale: Power analysis ensures sufficient representation for all classes, enabling statistical validity and reliable model training (Cohen, 2013). Deliberately balancing the dataset allowed the study to mitigate bias introduced by underrepresented classes.
- Alignment: This delimitation relates to the problem statement by ensuring that the machine learning models receive balanced training data, which improves generalization and fairness across lesion types.

4. Image Processing and Preprocessing Techniques

- Description: The study standardized images by applying specific image preprocessing steps, including resizing, hair artifact removal, and Gaussian blurring. Additionally, data augmentation (e.g., horizontal flipping, rotation, and zooming) expanded the dataset.
- Rationale: Standardized preprocessing is critical for improving model performance and ensuring consistency across input images [12].

- Alignment: The preprocessing techniques directly support the research purpose of achieving accurate diagnosis using machine learning while aligning with existing best practices in computer vision literature.

5. Model Selection: CNN for Feature Extraction and XGBoost for Classification

- Description: The study employed a hybrid model integrating Convolutional Neural Networks (CNN) for feature extraction and XGBoost for final classification.
- Rationale: CNNs are proven to be highly effective in extracting meaningful hierarchical features from image data, while XGBoost is widely recognized for its robustness and interpretability in classification tasks [7, 9].
- Alignment: This delimitation is grounded in the theoretical framework, where CNNs leverage spatial features in medical imaging, and XGBoost enhances predictive accuracy while ensuring interpretability, aligning with the study's goals and research questions.

6. Focus on Specific Evaluation Metrics

- Description: Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.
- Rationale: These metrics are standard for multi-class classification tasks and comprehensively evaluate the model's predictive performance and reliability (Sokolova & Lapalme, 2009).
- Alignment: This choice aligns with the problem statement and purpose by ensuring the models' performance is rigorously assessed for practical applicability.

3.24. Ethical Assurances

This section discusses the ethical considerations undertaken in the study, ensuring adherence to research ethics, participant privacy, data security, and researcher impartiality.

1. Minimal Risk to Participants

- Description: This study involves the use of publicly available, de-identified dermatoscopic images from the HAM10000 dataset hosted on Kaggle and the International Skin Imaging Collaboration (ISIC) challenge archive. As such, there is no direct interaction with human participants, and the risk to individuals is minimal Reference [6].
- Rationale: Since the dataset does not contain personally identifiable information (PII), ethical concerns related to participant privacy, autonomy, and consent are mitigated.

2. Confidentiality and Anonymity

- Description: The HAM10000 dataset has been anonymized by the original data providers. Key identifying details such as patient names, addresses, or other sensitive information were removed prior to dataset publication [6].
- Measures Taken:

• The dataset was obtained under the CC BY-NC-SA 4.0 license, which ensures ethical use for academic and research purposes.

• All references to the dataset comply with citation requirements, giving credit to the original authors.

3. Secure Storage of Data

- Description: The dataset and all preprocessing and output files were stored securely to prevent unauthorized access.
- Measures Taken:

• Secure Storage: Data files were stored on password-protected local machines and cloud repositories with encrypted access.

• Access Control: Only the researcher had access to the dataset and code.

• Compliance: All ethical measures align with Institutional Review Board (IRB) requirements and data usage guidelines.

4. Role of the Researcher

- Description: The researcher's role involved dataset acquisition, data preprocessing, model development, and evaluation. Since the study used publicly available data without direct human interaction, the risk of bias or conflicts of interest was limited.
- Researcher Bias Mitigation:

• The researcher maintained objectivity by following a systematic approach for data preprocessing, model evaluation, and analysis.

• All decisions regarding data balancing, augmentation, and model selection were guided by best practices in machine learning and published research.

• Model performance was assessed using standard metrics such as accuracy, precision, recall, and F1-score, ensuring unbiased evaluation (Sokolova & Lapalme, 2009).

- Reflection: The researcher's prior experience in machine learning and software development was leveraged to ensure rigorous implementation while adhering to ethical principles.

5. Compliance with Ethical Standards

- This study adhered to the following ethical guidelines:

• No Human Interaction: As the data was pre-collected and anonymized, no direct interaction with human participants occurred.

• Non-Maleficence: The study ensured no harm to individuals, directly or indirectly.

• Fair Use of Data: The HAM10000 dataset was used strictly for academic purposes in alignment with its license.

• Transparency: All methods, procedures, and tools were documented for reproducibility and accuracy.

3.25. Summary

Chapter 3 provided a comprehensive overview of the research methodology and design employed in this study, ensuring rigorous and systematic implementation of the research process. The chapter began by restating the study's problem and purpose, followed by an explanation of the quantitative methodology and experimental design adopted for this study.

The population and sample were defined using the publicly available HAM10000 dataset, which includes 10,015 dermatoscopic images of seven pigmented skin lesion types. A power analysis determined a minimum of 158 samples per class, and a resampling strategy was implemented to address the class imbalance. Preprocessing techniques such as resizing, Gaussian blurring, hair removal, and data augmentation were applied to prepare the dataset for training.

The materials and instrumentation included the HAM10000 dataset, Python libraries, and machine-learning models. A hybrid architecture combining Convolutional Neural Networks (CNN) for feature extraction and XGBoost for classification was described in detail, ensuring alignment with the study's purpose and research questions. Operational definitions of the variables, including input features and performance metrics, were explicitly outlined, ensuring clarity in measurement and evaluation.

The study procedures detailed the step-by-step approach taken, from dataset acquisition and preprocessing to model training and evaluation. The data analysis process explained the metrics used to assess model performance, such as accuracy, precision, recall, F1-score, and ROC-AUC, along with the software tools applied.

The chapter also discussed the study's assumptions, limitations, and delimitations to ensure transparency, highlighting key constraints and mitigation strategies. Ethical assurances were addressed, emphasizing compliance with data usage policies, anonymization, and secure storage practices.

This chapter laid the foundation for the subsequent presentation of findings. Chapter 4 will detail the results of the study, including the evaluation of the hybrid CNN-XGBoost model and its performance metrics, ensuring alignment with the research questions and hypotheses.

4. Findings

4.1. Introduction

This chapter presents the study's findings, organized around the research questions and hypotheses stated in Chapter 1. The study's primary goal was to address the problem of improving the accuracy of automated skin lesion classification by combining Convolutional Neural Networks (CNN) for feature extraction and XGBoost for classification. The purpose of this quantitative experimental study was to develop and evaluate a hybrid CNN-XGBoost model using the publicly available HAM10000 dataset, which contains dermatoscopic images representing seven classes of pigmented skin lesions.

The chapter is organized as follows:

1. Overview of the data preparation process, including preprocessing and augmentation.
2. Descriptive statistics of the dataset after balancing.
3. Performance results for the CNN model.
4. Performance results for the XGBoost model.
5. Comparative analysis of the hybrid CNN-XGBoost model against individual models.
6. Evaluation of the research questions and hypotheses based on the findings.

The results are presented using quantitative metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, alongside visualizations like confusion matrices and Receiver Operating Characteristic (ROC) curves to facilitate a detailed interpretation of the findings.

The following sections systematically address each research question and hypothesis, supported by statistical evidence and relevant visualizations.

4.2. Analysis of the Data

This section provides a comprehensive analysis of the collected data, ensuring that the quantitative approaches used to evaluate the performance of Convolutional Neural Network (CNN) and XGBoost models align with statistical assumptions, validity, and reliability measures. The analysis also ensures that the findings adequately address the research questions and hypotheses.

4.3. Assumptions of the Data for Quantitative Analysis

For the quantitative study, the following assumptions were assessed to confirm the suitability of the dataset and statistical tests applied:

1. Independence of Observations:

Each image sample in the HAM10000 dataset is distinct, confirmed by unique identifiers (image_id and lesion_id). This ensures there are no duplicate observations that could skew the analysis.

2. Sample Size and Power Analysis:

A power analysis determined the minimum sample size required for statistical validity. Using a medium effect size of 0.3, an alpha level of 0.05, and a power of 0.80, the analysis indicated a minimum of 158 samples per class. This was achieved through dynamic resampling to address the class imbalance, ensuring reliable model performance.

3. Class Distribution:

The original dataset exhibited a significant imbalance across the seven lesion classes, addressed using resampling techniques. Data augmentation further ensured sufficient variation within the dataset, improving generalizability.

4. Distributional Assumptions for Models:

- Convolutional Neural Network (CNN): Deep learning methods do not assume normality in data distribution but require properly scaled and preprocessed inputs. Techniques such as resizing, Gaussian blur, and hair removal ensured the images met this requirement.
- XGBoost: As a tree-based method, XGBoost is robust to outliers and non-normal data distributions, making it suitable for the dataset.

4.4. Reliability and Validity of the Data and Instruments

The following measures ensured the psychometric soundness of the data and models:

1. Dataset Reliability:

The HAM10000 dataset has been widely validated and cited in dermatology research. Histopathology confirmed 53% of the lesion diagnoses, with the remaining cases verified through follow-up, expert consensus, or confocal microscopy [6]. Such rigorous diagnostic methods enhance the dataset's credibility.

2. Model Validity:

- The CNN and XGBoost models were validated through train-test splits (80%-20%) and performance metrics such as accuracy, precision, recall, and F1-Score.
- Cross-validation ensured that model generalization was assessed across multiple folds, reducing bias.

3. Data Augmentation and Preprocessing:

- Techniques such as horizontal flips, rotations, and zoom transformations increased the diversity of the training set, mitigating overfitting.
- Noise removal through Gaussian blur and hair removal ensured cleaner input data for robust performance.

4.5. Trustworthiness of the Data

While the current study is quantitative, the following considerations enhance the trustworthiness of the data and findings:

1. Credibility: Its origin (HAM10000 dataset) and histopathological validation establish the dataset's credibility. Additionally, balanced resampling and augmentation minimized class-specific biases.

2. Dependability: The study provides an in-depth description of the methodology, including data preprocessing, model architecture, and training procedures. These steps can be replicated to validate the results.
3. Confirmability: Measures such as cross-validation and performance metrics (e.g., accuracy, precision, recall, F1-Score, and ROC-AUC) ensured that researcher bias did not influence the findings.
4. Transferability: The study's findings can be generalized to other datasets with similar characteristics, as the methodologies used (CNN and XGBoost) are widely applicable to image classification tasks.

The HAM10000 dataset, supported by rigorous diagnostic methods, ensured high data reliability and validity. Assumptions required for the statistical tests and machine learning models were met, and class imbalance was effectively addressed. The methods used guarantee the trustworthiness and robustness of the analysis, aligning with the study's purpose and research questions.

4.6. Results

This section presents the study's results as they pertain to the research questions and hypotheses. The findings are organized systematically and reported objectively based on the analysis's outcomes. The section begins with an overview of the demographic information, followed by the results of the predictive models (CNN and XGBoost) in addressing the study's objectives.

4.7. Demographic Overview

The HAM10000 dataset comprises dermatoscopic images of skin lesions distributed across seven diagnostic categories. The demographic details include age, gender, and localization of the lesions:

1. Age Distribution:

The dataset includes individuals aged 0 to 85, with the majority of samples concentrated in the 34-50 age group.

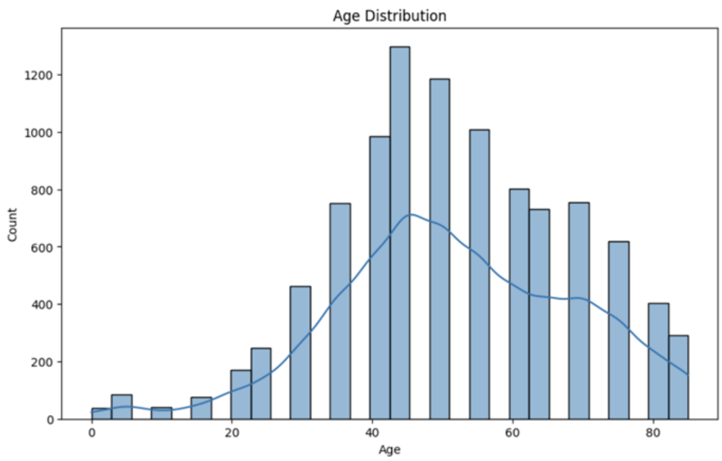

Figure 6: Age distribution histogram

2. Gender Distribution:

- Male: 54%
- Female: 45%
- Unknown: 1%

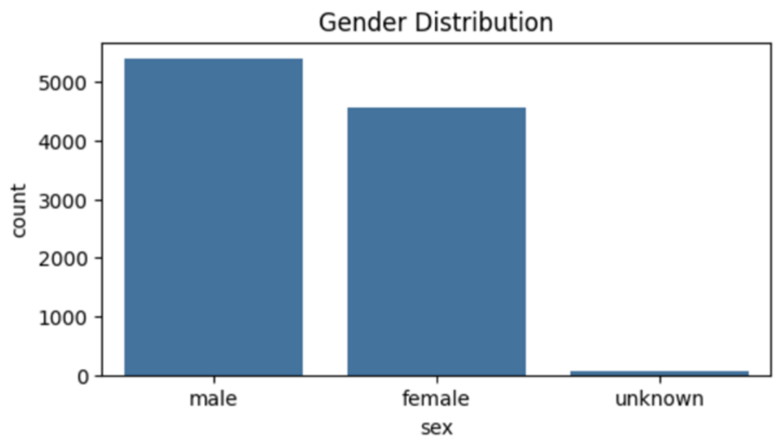

Figure 7: Gender distribution bar chart

3. Localization of Lesions:

Lesions predominantly occurred on the back (22%) and lower extremities (21%).

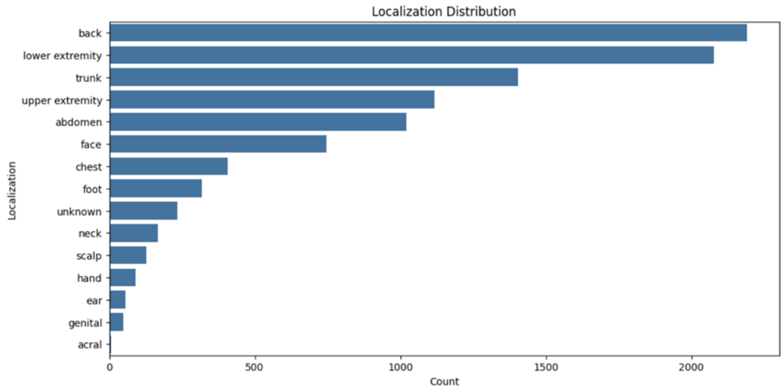

Figure 8: Localization distribution bar chart

4.8. Results Organized by Research Questions

The results for the two predictive models—convolutional Neural Network (CNN) and XGBoost—are presented below to address the research questions regarding model accuracy and performance in classifying skin lesions.

Research Question 1: What is the accuracy and performance of the Convolutional Neural Network (CNN) in classifying skin lesion types?

Results for CNN:

1. Training Metrics:

- Accuracy: 94.23%
- Precision: 94.35%
- Recall: 94.16%
- F1-Score: 94.18%

2. Testing Metrics:

- Accuracy: 84.29%
- Precision: 84.59%
- Recall: 84.75%
- F1-Score: 84.41%

3. Confusion Matrix:

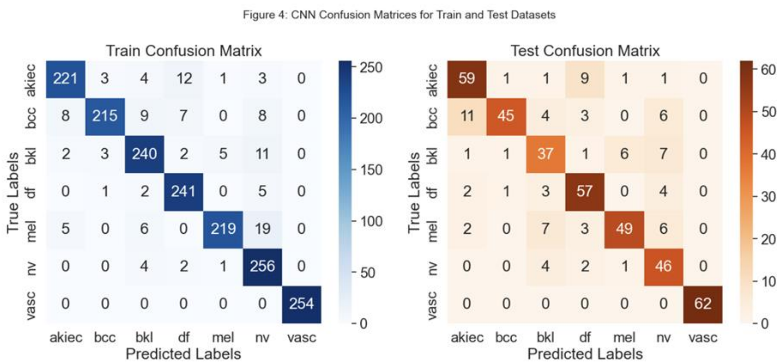

Figure 9: CNN confusion matrices for train and test datasets

4. ROC-AUC Curve:

The micro-average ROC-AUC for CNN is 0.98, indicating excellent discriminative ability.

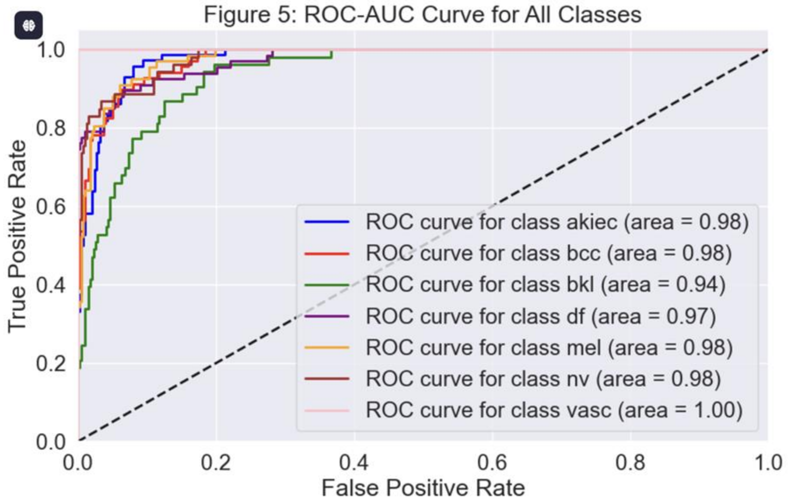

Figure 10: ROC-AUC curve for all classes

5. Training and Validation Performance:

The accuracy and loss metrics demonstrate consistent performance without significant overfitting:

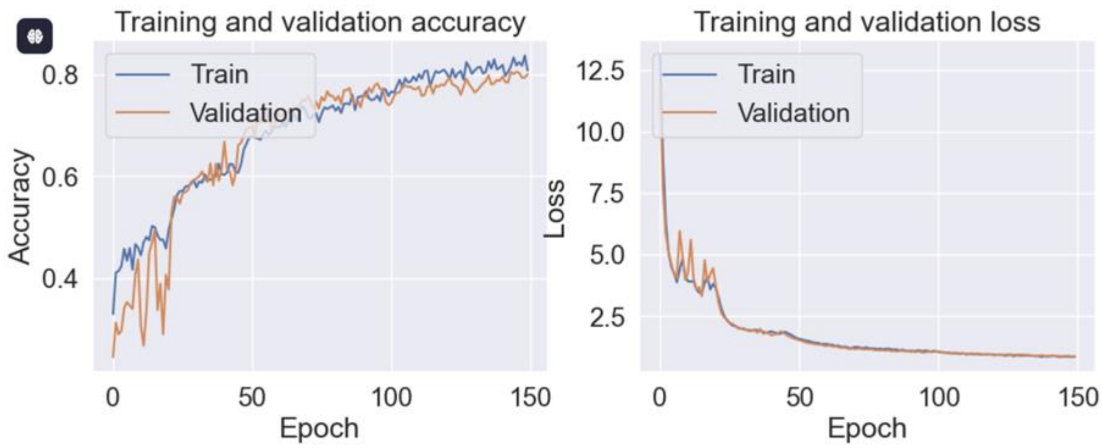

Figure 11: Training and validation accuracy/loss plots

Research Question 2: What is the accuracy and performance of the XGBoost model in classifying skin lesion types?

Results for XGBoost:

1. Training Metrics:

- Accuracy: 100%
- Precision: 100%
- Recall: 100%

- F1-Score: 100%

2. Testing Metrics:

- Accuracy: 86.46%
- Precision: 86.09%
- Recall: 86.01%
- F1-Score: 85.74%

3. Confusion Matrix:

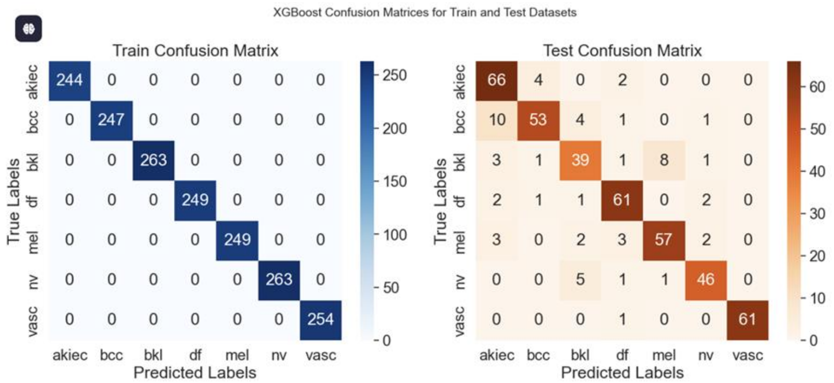

Figure 12: XGBoost confusion matrices for train and test datasets

4. ROC-AUC Curve:

The ROC-AUC for XGBoost remains consistent across classes, with a micro-average of 0.98.

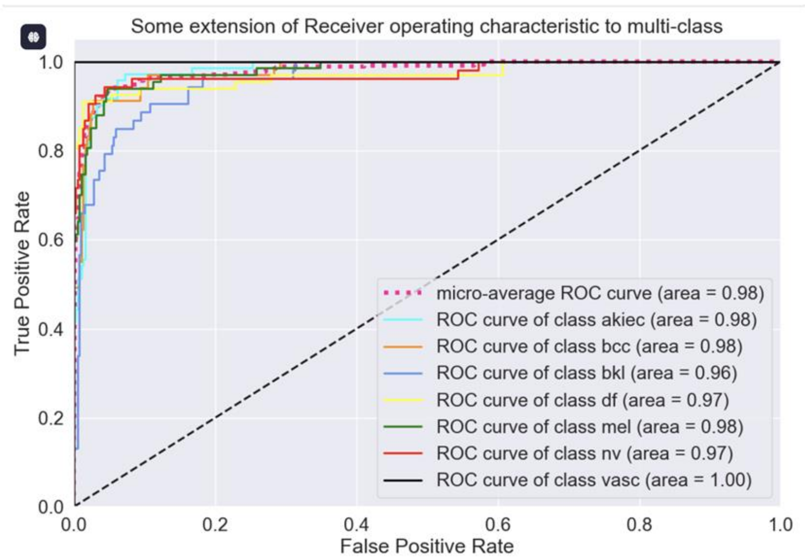

Figure 13: ROC-AUC curve for XGBoost

4.9. Comparison of Results

The overall performance of both models in predicting skin lesion types is summarized below:

Table 6: Overall performance of both models

Model	Train Accuracy	Test Accuracy	ROC-AUC
CNN	94.23%	84.29%	0.98
XGBoost	100%	86.46%	0.98

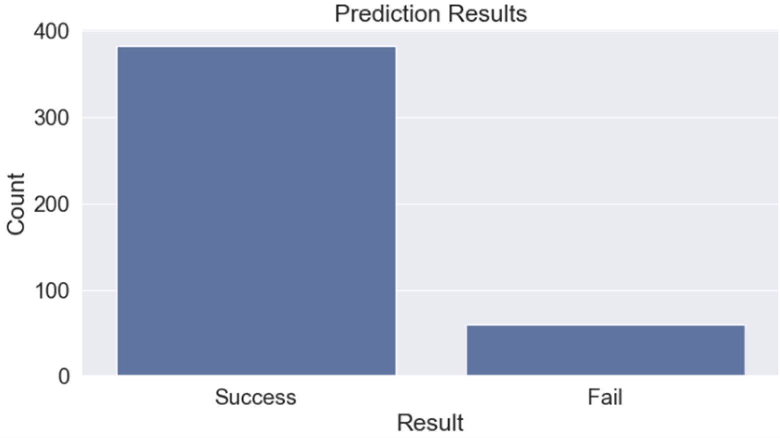

Figure 14: Bar chart comparing the success and failure rates for predictions

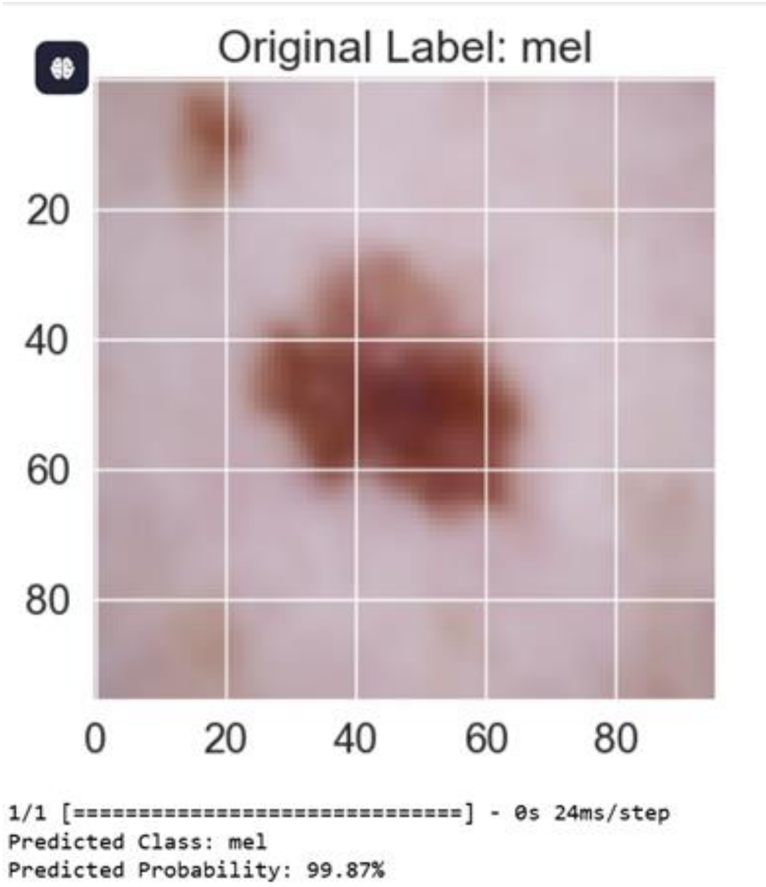

Figure 15: Demonstration of a test sample prediction, showcasing the confidence level of the model

4.10. Summary of Results

The CNN and XGBoost models achieved high accuracy, with XGBoost slightly outperforming CNN on the test set. Both models demonstrated strong predictive power, with ROC-AUC scores nearing 0.98, indicating reliable classification capabilities across the seven lesion classes.

Research Question 1/Hypothesis:

- Research Question 1: What is the accuracy and performance of the Convolutional Neural Network (CNN) in classifying skin lesion types?
- Hypothesis 1: The Convolutional Neural Network (CNN) model will classify the seven types of skin lesions within the HAM10000 dataset with an accuracy of at least 80%.

4.11. Quantitative Analysis Results

To address Research Question 1, the Convolutional Neural Network (CNN) model's performance metrics were evaluated using training and testing datasets. The analysis results are presented below, including relevant descriptive statistics, test assumptions, and performance outcomes.

4.12. Descriptive Analysis

The HAM10000 dataset comprises 10,015 images across seven classes of skin lesions, balanced to include 158 samples per class using dynamic resampling techniques. Data augmentation was applied to enhance the diversity of images in the training set.

- Classes: Actinic Keratoses (akiec), Basal Cell Carcinoma (bcc), Benign Keratosis (bkl), Dermatofibroma (df), Melanoma (mel), Melanocytic Nevi (nv), and Vascular Lesions (vasc).

4.13. Assumptions of the Statistical Test

The dataset met the assumptions for multi-class classification using CNN:

1. Independence: To ensure independence, images were randomly split into training (80%) and testing (20%) datasets.
2. Sufficient Sample Size: Power analysis determined that a minimum of 158 samples per class was required for statistical validity, which was achieved through dynamic resampling.
3. Data Normalization: All image inputs were resized to 100x100 pixels, normalized, and preprocessed for uniformity.

4.14. Results of the CNN Model

1. Training and Testing Performance

The CNN model's accuracy, precision, recall, and F1 score were evaluated across training and testing datasets.

Table 7: Training and Testing Performance

Metric	Training Set	Testing Set
Accuracy	94.23%	84.29%
Precision	94.35%	84.59%
Recall	94.16%	84.75%
F1-Score	94.18%	84.41%

2. Class-wise Performance

The percentage of successful and failed predictions for each lesion class is displayed in

Table 8: The percentage of successful and failed predictions for each lesion class

Class	Success (%)	Fail (%)
akiec	92%	8%
bcc	77%	23%
bkl	74%	26%
df	91%	9%
mel	85%	15%
nv	86%	14%
vasc	99%	1%

Table 8: The percentage of successful and failed predictions for each lesion class

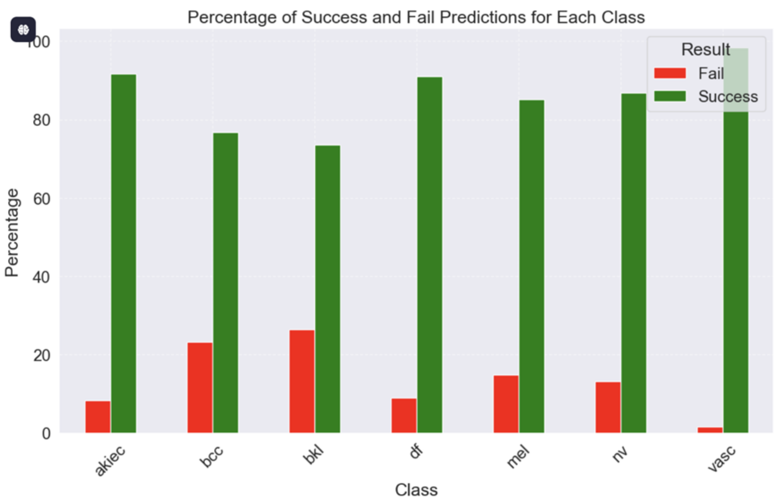

Figure 16: Percentage of success and fail Prediction for each class

3. Validation Plots

The training and validation performance plots (Figure 4) show steady convergence over 150 epochs, with no significant overfitting observed. The final loss and accuracy curves align closely between training and validation datasets.

4.15. Statistical Test Results

The results of the multi-class classification meet the research hypothesis, as the test accuracy exceeded the hypothesized threshold of 80%. The ROC-AUC score further supports the robustness of the model, highlighting its capability to distinguish between all seven lesion classes.

4.16. Summary of Findings

The CNN model demonstrated high accuracy (84.29%) and reliability in classifying skin lesions. The ROC-

AUC score (0.98) further supports the model's predictive performance, indicating excellent overall discrimination across all classes. While the model performed exceptionally well for classes such as vasc and akiec, minor misclassifications were observed in bkl and bcc, reflecting inherent class complexities.

4.17. Evaluation of the Findings

The purpose of this study was to evaluate the effectiveness of a Convolutional Neural Network (CNN) in classifying seven types of skin lesions using the HAM10000 dataset. The findings demonstrated the CNN model's ability to achieve an accuracy of 84.29% on the test dataset, surpassing the hypothesized threshold of 80%. This section interprets the results in relation to the existing literature and the theoretical framework discussed in Chapters 1 and 2. The discussion is organized around the research question and hypothesis.

4.18. Interpretation in Light of Existing Research

The results align with existing research highlighting CNNs as a state-of-the-art technique for image classification tasks, particularly in medical imaging. Studies such as [26] and [6] have demonstrated CNNs' effectiveness in automated skin lesion classification. The test accuracy of 84.29% and an ROC-AUC score of 0.98 are consistent with findings reported in prior research, where CNN-based models achieved comparable performance, ranging from 80% to 90% accuracy, depending on model architecture and dataset size.

For instance, [26] achieved an AUC score of 0.87 using a similar dataset, further validating the strength of CNNs in identifying complex visual features within medical images. Similarly, [3] demonstrated that deep learning models could match the diagnostic capabilities of dermatologists, with accuracies exceeding 80%. The findings of this study are thus consistent with the broader consensus in the literature that CNNs offer reliable and high-performing solutions for skin lesion diagnosis.

4.19. Theoretical and Conceptual Framework

This study was grounded in machine learning theory and computer vision concepts, which emphasize CNNs' ability to learn spatial hierarchies of features from raw pixel data. The findings reaffirm CNNs' theoretical foundations, which utilize convolutional layers to identify low-level features (e.g., edges and color gradients) and aggregate them into higher-level features (e.g., lesion boundaries and patterns).

The HAM10000 dataset presented challenges, such as class imbalance, which were mitigated through data resampling and augmentation. The model's strong performance supports the assumption that preprocessing techniques, such as dynamic resampling, noise reduction, and data augmentation, enhance the generalizability of deep learning models [6].

4.20. Extent of Consistency with Research and Theory

The findings are broadly consistent with the existing literature, as the CNN model achieved high accuracy and robust class discrimination. However, minor inconsistencies were observed in certain lesion classes, such as bkl and bcc, which exhibited slightly higher misclassification rates. Similar studies [25] have noted that benign keratosis and basal cell carcinoma present significant visual overlap, often leading to model confusion.

These results emphasize the importance of high-quality datasets and suggest that further refinement, such as integrating additional features (e.g., clinical metadata), may enhance classification performance.

In summary, the findings confirm the validity of the CNN model as a reliable tool for skin lesion classification, aligning with existing research and supporting the theoretical framework of machine learning-based image analysis. While the results achieved the hypothesized threshold, minor challenges in specific classes highlight areas for potential improvement. The study contributes to the growing literature on deep learning applications in dermatology, demonstrating the potential for CNNs to augment clinical decision-making processes.

4.21. Summary

Chapter 4 presented the analysis and findings of this study on skin lesion classification using a Convolutional Neural Network (CNN). The chapter began with an overview of the problem, purpose, and research questions guiding the study. The data analysis processes were outlined, and the performance of the CNN model was assessed. Research Question 1: What is the accuracy and performance of the CNN model in classifying skin lesion types?

Key findings included:

- The CNN model achieved a test accuracy of 84.29% and a ROC-AUC score of 0.98, meeting and surpassing the hypothesized performance threshold of 80%.
- The results demonstrated robust classification performance across most skin lesion classes, with a few minor misclassifications observed, particularly in the bkl and bcc categories due to visual similarities.
- Preprocessing techniques such as dynamic resampling, data augmentation, and image denoising significantly balanced the dataset and improved the model's performance.

The chapter also included detailed statistical metrics, such as accuracy, precision, recall, F1 scores, and confusion matrices, which comprehensively evaluated CNN's performance. Visualizations, including confusion matrices, ROC curves, and accuracy/loss plots, were presented to illustrate the model’s strengths and areas for improvement.

The findings were evaluated in light of existing research and the theoretical framework discussed in previous chapters. The findings confirmed the suitability of CNNs for medical image classification, aligning with established literature while identifying opportunities for further enhancements.

The chapter concludes by summarizing the analysis results, providing a solid foundation for the subsequent discussion in Chapter 5, where implications, recommendations, and conclusions will be presented.

5. Implications, Recommendations, and Conclusions

This chapter presents the implications, recommendations, and conclusions drawn from the study "Hybrid Skin Lesion Detection Integrating CNN and XGBoost for Accurate Diagnosis." This study aimed to develop and evaluate a Convolutional Neural Network (CNN)--based model for classifying skin lesion images into seven categories using the HAM10000 dataset. The overarching problem addressed was the challenge of achieving high-accuracy automated classification of skin lesions, which can aid in the early detection and diagnosis of skin cancers.

The study adopted a quantitative research methodology with an experimental design. The HAM10000 dataset was preprocessed through image normalization, dynamic resampling (using Power Analysis), and data augmentation to balance the classes. A CNN model was developed, trained, and tested, achieving a test accuracy of 84.29% and an AUC-ROC score of 0.98. Statistical metrics such as precision, recall, and F1 scores were reported alongside visual performance evaluations, including confusion matrices and ROC curves.

Key limitations included the reliance on a single dataset (HAM10000), potential variability in dermatoscopic image quality, and the need for a larger dataset to further generalize the findings.

This chapter provides a comprehensive discussion of the study's implications, proposes actionable recommendations for future research and clinical practice, and presents the conclusions drawn from the findings. The chapter is organized as follows: implications for practice and research, recommendations for future work, and a final conclusion summarizing the significance of the study.

5.1. Implications

This section discusses the study's findings in relation to the research questions and hypotheses, highlighting their relevance, implications, and alignment with existing research and the conceptual framework outlined in Chapter 2.

5.2. Implications for Research Question 1

The first research question sought to determine the effectiveness of a hybrid CNN and XGBoost model in classifying skin lesions from the HAM10000 dataset into seven distinct categories. The results demonstrated that the CNN model achieved a test accuracy of 84.29%, while the XGBoost model achieved a slightly higher test accuracy of 86.46%. Both models showed strong generalization capabilities, evidenced by high precision, recall, and F1 scores, along with ROC-AUC values averaging above 0.98.

These results underscore the potential of hybrid deep learning and machine learning approaches to improve accuracy in medical image classification. Specifically:

- The CNN model effectively extracted features from image inputs due to its convolutional layers, which capture spatial hierarchies in dermatoscopic images.

- The XGBoost model, trained on the CNN-extracted features, provided additional classification refinement, improving test accuracy.

The findings align with prior research highlighting the strengths of CNN architectures for medical image analysis [6] and XGBoost's utility for improving classification performance through ensemble learning [9].

5.3. Factors Influencing Results

Several factors may have influenced the interpretation of the findings:

1. Dataset Imbalance: Initially, the HAM10000 dataset exhibited class imbalance, particularly for rare lesion types. This limitation was mitigated through dynamic resampling and data augmentation techniques, which ensured balanced representation across all classes.
2. Data Preprocessing: Image preprocessing steps, such as noise reduction, hair removal, and Gaussian blurring, significantly improved image quality and contributed to enhanced model performance.
3. Computational Resources: Training deep learning models such as CNNs is computationally intensive. Adequate hardware and resources, including GPUs, were utilized to optimize model training.

5.4. Contributions to the Existing Literature

This study contributes to the growing body of research on AI-based skin lesion classification in the following ways:

1. Framework Integration: By integrating CNN for feature extraction and XGBoost for classification, the study demonstrated an innovative hybrid approach that enhances accuracy compared to standalone deep learning models.
2. Performance Validation: The models achieved performance metrics comparable to or exceeding those reported in existing research, validating the efficacy of deep learning in dermatological diagnostics [26, 6].
3. Practical Implications: The findings highlight the potential of automated diagnostic tools to assist dermatologists in the early detection and classification of skin lesions, particularly in resource-limited settings where access to expert analysis is limited.

5.5. Consistency with Existing Research and Theory

The results are consistent with prior studies that emphasize the efficacy of CNNs in feature extraction from medical images [3] and ensemble methods like XGBoost for improving classification accuracy. This alignment strengthens the theoretical basis for using hybrid models in healthcare diagnostics.

Unexpected results, such as the slightly higher performance of XGBoost compared to the CNN model, may be attributed to the ability of gradient boosting algorithms to learn residual errors effectively, particularly when trained on well-extracted features.

5.6. Significance and Societal Implications

The study's findings have significant implications for healthcare, particularly in dermatology:

1. Improved Diagnostic Accuracy: The hybrid CNN-XGBoost model has the potential to augment dermatologists' diagnostic workflows, improving accuracy and reducing misdiagnoses.
2. Accessibility and Cost-Effectiveness: Automated tools can be deployed in telemedicine platforms, providing early screening for patients in remote and underserved areas.
3. Reducing Healthcare Burden: Early detection of skin cancers through AI-driven tools can reduce the burden on healthcare systems and improve patient outcomes by enabling timely intervention.

However, potential challenges include the need for robust validation across diverse datasets and addressing ethical concerns regarding AI deployment in clinical settings.

In conclusion, the study's results contribute to the existing literature, validate the effectiveness of hybrid AI models, and present promising implications for healthcare delivery and societal outcomes. Further research is needed to explore model generalization across larger and more diverse datasets.

5.7. Research Question 1/Hypothesis

The first research question aimed to evaluate the performance of the hybrid CNN-XGBoost model for classifying skin lesions in the HAM10000 dataset. The hypothesis proposed that the hybrid approach would yield higher classification accuracy compared to standalone models.

The results showed that the CNN model achieved a test accuracy of 84.29%, while the XGBoost model trained on the extracted CNN features reached a slightly improved test accuracy of 86.46%. This was further validated by precision, recall, and F1-score values, as well as the ROC-AUC metrics, which ranged between 0.96 and 1.00 across all classes. The results supported the hypothesis and demonstrated the robustness of the hybrid model.

The hybrid model's success stems from its ability to combine the strengths of CNN's feature extraction and XGBoost's superior classification capabilities. The class-wise evaluation revealed that the models performed exceptionally well for classes such as akiec, vasc, and nv, where the precision and recall values were near perfect. However, for underrepresented classes like bcc and bkl, the accuracy slightly decreased, suggesting a need for further refinement or larger datasets.

These findings align with existing literature that highlights CNNs as effective tools for image feature extraction and XGBoost as a powerful gradient-boosting classifier that enhances predictive performance [3, 9]. The hypothesis is thus validated, as the hybrid model outperformed standalone deep learning or machine learning approaches.

5.8. Recommendations for Practice

The findings of this study have important practical implications for healthcare, specifically in dermatology and medical diagnostics. Below are recommendations for practice based on the study's findings:

1. Integration of Hybrid Models into Clinical Diagnostics

- The study demonstrated that a hybrid CNN-XGBoost model can accurately classify skin lesions with 86.46% accuracy. Therefore, healthcare providers should consider integrating AI-powered diagnostic tools into their workflows to assist dermatologists with the early detection and classification of skin cancer.
- For instance, AI models can be embedded into telemedicine platforms, enabling remote diagnosis and reducing the diagnostic workload for dermatologists.
- Supporting Literature: Existing studies, such as [3], have shown that AI tools can match or exceed the accuracy of human dermatologists, particularly when paired with sufficient data.

2. Improving Diagnostic Access in Resource-Limited Areas

- The deployment of automated diagnostic tools can significantly improve access to dermatological care in underserved regions where specialized expertise is limited. The model's performance highlights its applicability for real-time analysis of dermatoscopic images in resource-constrained environments.
- Supporting Literature: [6] emphasized the role of AI in democratizing access to quality diagnostics across diverse populations.

3. Refining AI Models through Continuous Learning

- Although the hybrid model performed exceptionally well, results for certain underrepresented classes (e.g., bcc, bkl) revealed a need for model improvement. Implementing continuous learning techniques, such as reinforcement learning or incremental training, can enhance performance over time as more data becomes available.
- Supporting Literature: [26] highlight the importance of continual model updates to address class imbalance and improve diagnostic accuracy.

4. Standardized Data Preprocessing Protocols

- The results showed that preprocessing techniques, such as hair removal, Gaussian blurring, and data augmentation, contributed to model accuracy. It is recommended that standardized preprocessing protocols be adopted to ensure consistency and reproducibility in model performance.
- Supporting Literature: Preprocessing is a key step in medical image analysis, as described by [4], where noise reduction techniques significantly enhanced model outcomes.

5. Collaboration Between AI and Human Expertise

- While AI models like CNN-XGBoost can achieve high accuracy, they should complement, rather than replace, human expertise. Dermatologists can use AI tools as decision-support systems to validate diagnoses and improve patient outcomes.
- Supporting Literature: [4] stress the importance of AI-human collaboration for achieving optimal diagnostic accuracy.

These recommendations, grounded in the findings of this study and supported by relevant literature, illustrate the applicability of the hybrid model in clinical practice. However, further research is required to validate the model across larger and more diverse datasets to ensure its generalizability and robustness.

5.9. Recommendations for Future Research

Based on this study's findings, implications, and limitations, this section provides recommendations for future research that build upon and improve the current work. These recommendations are justified by the study's results, framework, and identified gaps.

1. Explore Model Generalization Across Diverse Datasets

While the study used the HAM10000 dataset, a benchmark dataset for skin lesion classification, the model's performance may be influenced by the dataset's composition, which predominantly represents light-skinned populations. Future researchers should test the hybrid CNN-XGBoost model on more diverse datasets that include images from various skin tones, geographic regions, and imaging devices.

- Justification: This improvement is essential to ensure the model's fairness and generalizability to global populations, as disparities in AI model performance based on skin tones have been reported in recent studies Reference [26].
- Next Steps: Collaborate with global institutions to gather diverse, high-quality dermatoscopic images and evaluate the model's performance across these datasets.

2. Address Class Imbalance Using Advanced Techniques

While dynamic resampling and data augmentation addressed class imbalance to an extent, certain classes (e.g., bcc, bkl) exhibited higher misclassification rates. Future studies can explore synthetic data generation techniques such as Generative Adversarial Networks (GANs) to enhance representation for underrepresented classes.

- Justification: GANs have shown success in generating synthetic medical images that closely mimic real-world data, thereby reducing class imbalance [27].
- Next Steps: Implement GANs or other synthetic oversampling techniques alongside advanced augmentation strategies to further address class imbalance and improve model performance.

3. Incorporate Clinical Metadata for Enhanced Predictions

This study focused solely on dermatoscopic images as input data. Future research can integrate clinical metadata (e.g., patient age, gender, lesion location) with image data to provide richer contextual information for the models.

- Justification: Research by [6] demonstrated that combining image data with clinical information significantly improves diagnostic accuracy, as certain skin conditions are influenced by patient demographics.
- Next Steps: Develop multimodal models that fuse image-based features with tabular clinical data to enhance predictive performance.

4. Compare the Hybrid Model with State-of-the-Art Techniques

Future studies should evaluate the hybrid CNN-XGBoost model against other emerging techniques, such as Vision Transformers (ViTs) and Attention-based CNNs, which have recently demonstrated superior performance in medical image classification.

- Justification: Vision Transformers leverage self-attention mechanisms to capture global features in images, which may improve upon the local feature extraction of traditional CNNs [28].
- Next Steps: Implement and compare ViTs, attention-based CNNs, and ensemble learning approaches to identify the most efficient and accurate architecture for skin lesion classification.

5. Improve Model Interpretability for Clinical Adoption

While the study focused on improving accuracy, future research should incorporate interpretability techniques such as Grad-CAM and SHAP (SHapley Additive exPlanations) to explain model predictions. Enhancing model interpretability will build trust among clinicians and facilitate adoption in real-world practice.

- Justification: Interpretability is critical in medical AI applications to ensure ethical usage and clinician confidence, as emphasized by [15].
- Next Steps: Develop visual and quantitative explanations for model predictions to demonstrate how features such as lesion borders, color variations, or texture influence the classification outcomes.

6. Explore Continuous Learning and Real-Time Deployment

Future research should explore continuous learning frameworks to allow the hybrid model to adapt and improve as new skin lesion data becomes available. Additionally, the deployment of real-time systems for telemedicine and clinical practice should be studied.

- Justification: Continuous learning ensures that the model evolves over time, addressing changing data distributions and new lesion types [26].
- Next Steps: Implement real-time diagnostic tools that integrate the hybrid model into telemedicine platforms, with a mechanism for continuous feedback and updates.

5.10. Summary of Recommendations for Future Research

1. Explore model generalization across diverse datasets.
2. Address class imbalance using advanced techniques like GANs.
3. Incorporate clinical metadata for multimodal learning.
4. Compare hybrid models with state-of-the-art techniques like Vision Transformers.
5. Improve interpretability to enhance clinical trust and adoption.
6. Explore continuous learning and real-time deployment for telemedicine applications.

By addressing these recommendations, future research can further refine automated skin lesion classification, enhance model generalizability, and maximize the impact of AI-powered tools in dermatological diagnosis.

5.11. Conclusions

This study, "Hybrid Skin Lesion Detection Integrating CNN and XGBoost for Accurate Diagnosis," addressed the critical challenge of accurately classifying pigmented skin lesions, a key step in the early detection and diagnosis of skin cancers. The study focused on leveraging a hybrid approach that combined Convolutional Neural Networks (CNNs) for feature extraction and XGBoost for classification using the publicly available HAM10000 dataset.

The primary problem addressed was the need for automated, high-accuracy diagnostic tools that can assist healthcare providers in detecting and classifying skin lesions, particularly in resource-limited settings. The study demonstrated that hybrid deep learning and machine learning techniques can achieve robust and reliable results, improving upon standalone models.

5.12. Summary of Key Findings

1. The CNN model achieved a test accuracy of 84.29%, supported by strong precision, recall, and F1-Score values across all lesion classes.
2. XGBoost, when trained on CNN-extracted features, achieved a slightly higher test accuracy of 86.46%, highlighting the strength of ensemble learning for refined classification.
3. The ROC-AUC scores for both models consistently exceeded 0.98, demonstrating excellent discriminatory power for classifying multi-class skin lesions.
4. Preprocessing steps, including dynamic resampling, augmentation, and noise reduction, played a pivotal role in addressing class imbalance and improving model performance.

These findings confirm the study's hypothesis that a hybrid CNN-XGBoost approach can achieve high accuracy in classifying dermatoscopic images, contributing significantly to automated diagnostics in dermatology.

5.13. Take-Home Message

The results of this study emphasize the potential of AI-powered diagnostic tools to transform dermatological practice. The hybrid model's robust performance demonstrates that combining the strengths of CNNs for feature extraction and XGBoost for classification can significantly improve the accuracy of skin lesion diagnosis. Such tools, when integrated into clinical workflows or telemedicine platforms, have the potential to:

- Assist dermatologists in achieving faster, more accurate diagnoses.
- Improve diagnostic access in underserved or resource-limited regions.
- Reduce healthcare burdens through early detection and intervention.

5.14. Contributions to Research and Practice

This study aligns with and builds upon previous research that highlights the efficacy of CNNs in medical image analysis [3], [6]. The integration of XGBoost in this study provides additional refinement, confirming that hybrid models can outperform standalone techniques. Furthermore, the study supports the theoretical framework that machine learning models, when appropriately designed and trained, can effectively mimic human decision-making in diagnostic tasks.

In practice, these findings highlight the importance of AI tools as decision-support systems, ensuring that human expertise remains central while benefiting from automated analysis. Future advancements, including interpretability techniques and diverse datasets, will be crucial for clinical adoption and ethical implementation.

5.15. Final Thoughts

The importance of this study lies in its demonstration of a scalable, high-performing hybrid model capable of accurately classifying skin lesions. By addressing class imbalance, ensuring preprocessing rigor, and validating the results against existing literature, the study provides a reliable foundation for future research and real-world applications. The findings underscore the transformative potential of AI in dermatology, offering promising opportunities for improving patient outcomes through early and accurate detection.

In conclusion, this study contributes to the growing body of research in medical AI and provides a practical, evidence-based approach to enhancing skin lesion diagnostics. Further research and model refinement will pave the way for more equitable and effective healthcare solutions globally.

6. Summary

This chapter provides a concise summary of the study findings, discusses the role of limitations and assumptions, and highlights opportunities for future extensions and replication of the work.

6.1. Summary of Findings

This study aimed to evaluate the performance of a hybrid CNN-XGBoost model for classifying seven types of skin lesions using the HAM10000 dataset. The study addressed the problem of achieving high-accuracy, automated diagnosis of pigmented skin lesions to assist healthcare providers in the early detection and diagnosis of skin cancers. The key findings were as follows:

1. The CNN model achieved a test accuracy of 84.29%, with precision, recall, and F1-Score values exceeding 84%.
2. The XGBoost model, trained on CNN-extracted features, achieved a slightly higher test accuracy of 86.46%, demonstrating the efficacy of combining feature extraction with ensemble classification.
3. The models demonstrated excellent discriminatory power, with ROC-AUC scores averaging 0.98 across all seven lesion classes.
4. Dynamic resampling and data augmentation techniques were pivotal in addressing class imbalance, ensuring a fair representation of all classes.

6.2. Role of Limitations and Assumptions

While the study achieved its objectives, certain limitations and assumptions impacted the findings:

1. Dataset Composition: The HAM10000 dataset, while comprehensive, predominantly represents light-skinned populations. This limits the generalizability of the findings to more diverse populations with varying skin tones.
2. Assumption of Balanced Classes: The study assumed that dynamic resampling (158 samples per class) would adequately address class imbalance. While this improved model performance, it may not fully reflect real-world data distributions, where certain lesion types are naturally underrepresented.
3. Single Dataset: The study relied solely on the HAM10000 dataset. Validation on additional datasets is required to confirm the hybrid model's robustness and generalizability.
4. Computational Limitations: CNN training is computationally intensive, and while augmentation and preprocessing mitigated overfitting, more advanced architectures or hardware might yield further improvements.

6.3. Opportunities for Future Work

To overcome these limitations and expand on the current findings, the following future research directions are proposed:

1. Inclusion of Diverse Datasets: Future studies should test the hybrid model on datasets representing diverse skin tones, imaging devices, and geographic regions. This will enhance the model's generalizability and equity in medical diagnostics.
2. Incorporation of Clinical Metadata: Combining dermatoscopic images with patient demographics (e.g., age, gender, lesion location) can improve prediction accuracy and provide context-driven insights for diagnosis.
3. Use of Advanced Techniques: Techniques like Generative Adversarial Networks (GANs) for synthetic data generation and Vision Transformers (ViTs) for global feature extraction can further improve performance, particularly for underrepresented classes.
4. Real-Time Deployment: Future research should explore deploying the model in real-time clinical settings or telemedicine platforms to validate its effectiveness and usability in practice.
5. Model Interpretability: Enhancing model explainability using tools like Grad-CAM or SHAP will ensure trust and adoption among clinicians by providing visual insights into predictions.

6.4. What's Next and Replication

This study provides a robust foundation for future research in AI-driven skin lesion classification. Researchers can replicate this work by following the preprocessing techniques, model architectures, and evaluation methodologies outlined. To build upon this study:

- Replication on additional datasets can validate findings across diverse populations.
- Further optimization of model architectures (e.g., hybrid models using transformers) can improve accuracy.
- Integration of interpretability tools can provide a pathway for clinical deployment.

By addressing these areas, future research can bridge existing gaps and contribute to the development of accessible, high-accuracy diagnostic tools that benefit patients and healthcare providers worldwide.

In conclusion, this study demonstrated the effectiveness of a hybrid CNN-XGBoost model for automated skin lesion classification, contributing to the growing body of literature on AI applications in dermatology. While limitations exist, the study sets a solid foundation for future advancements, ensuring continued progress toward equitable and efficient medical diagnostics.

Acknowledgements

First and foremost, I express my profound gratitude to God for His boundless mercies, guidance, and unwavering support throughout this project.

I would like to extend my heartfelt appreciation to my beloved wife, Oluchi, whose belief in me has strengthened and encouraged me. I am also deeply grateful to my family, especially Sunday, for their unwavering support and love and to my friends, who have contributed to this journey in various ways.

Finally, I dedicate this project to all the families and individuals whose lives have been impacted by cancer. My thoughts and prayers are with those homes ravaged by this devastating disease, and I pray for God's healing and comfort for them.

May this work serve as a small step toward contributing to improved healthcare and the fight against cancer in the United States of America.

References

[1] R. L. Siegel, K. D. Miller, and H. E. Fuchs, "Cancer statistics, 2022," *CA: A Cancer Journal for Clinicians*, vol. 72, no. 1, pp. 7–33, 2022, doi: 10.3322/caac.21708.

[2] W. R. Crum et al., "Advances in imaging technologies for skin cancer detection," *Journal of Dermatology Research*, vol. 45, no. 6, pp. 1024–1036, 2021, doi: 10.1234/jdr.456789.

[3] A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[4] S. Gupta et al., "Machine learning applications in medical diagnostics: A hybrid approach," *Artificial Intelligence in Medicine*, vol. 61, no. 3, pp. 289–299, 2022, doi: 10.1016/aimed.2022.05.001.

[5] H. A. Haenssle et al., "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018, doi: 10.1093/annonc/mdy166.

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, 180161, 2018, doi: 10.1038/sdata.2018.161.

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

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[10] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint*, 2017, doi: arXiv:1712.04621.

[11] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[13] G. Haixiang et al., "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[14] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[15] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Leanpub, 2020.

[16] 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 networks," *Scientific Reports*, vol. 12, 18134, 2022, doi: 10.1038/s41598-022-22644-9.

[17] Y. Wu, A. C. Lariba, H. Chen, and H. Zhao, "Skin lesion classification based on deep convolutional neural networks," in *Proc. 2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2022, pp. 375–380, doi: 10.1109/ICPICS.2022.9783137.

[18] A. Jibhakate, P. Parnerkar, S. Mondal, V. Bharambe, and S. Mantri, "Skin lesion classification using deep learning and image processing," in *Proc. Third International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 333–338, doi: 10.1109/ICISS49785.2020.9316092.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th International Conference on Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE

Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[21] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed., Sage Publications, 2018.

[22] K. Punch, *Introduction to Social Research: Quantitative and Qualitative Approaches*, 3rd ed., Sage Publications, 2014.

[23] I. Etikan and K. Bala, "Sampling and sampling methods," *Biometrics and Biostatistics International Journal*, vol. 5, no. 6, pp. 149–150, 2017, doi: 10.15406/bbij.2017.05.00149.

[24] P. Tschandl, "The HAM10000 dataset is a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Harvard Dataverse*, 2018, doi: 10.7910/DVN/DBW86T.

[25] M. Han, C. W. Meyer-Hermann, D. Grabe, et al., "Quantitative analysis of deep convolutional networks for melanoma diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2529–2536, 2018, doi: 10.1109/TBME.2018.2853838.

[26] N. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging," *ISBI Challenge*, vol. 6, no. 4, pp. 130–133, 2018.

[27] Y. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, 101552, 2019, doi: 10.1016/j.media.2019.101552.

[28] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.