Skin Diseases Classification with Machine Learning and Deep Learning Techniques: A Systematic Review

Amina Aboulmira[1], Hamid Hrimech[2], Mohamed Lachgar[3]
LAMSAD Laboratory, ENSA, Hassan First University, Berrechid, Morocco[1,2]
LTI Laboratory, ENSA, Chouaib Doukkali University, El Jadida, Morocco[3]
L2IS Laboratory, Faculty of Science and Technology, University Cadi Ayyad, Marrakech, Morocco[3]
Higher Normal School, Department of Computer Science, University Cadi Ayyad, Marrakech, Morocco[3]

***Abstract*—Skin cancer is one of the most prevalent types of cancer worldwide, and its early detection is crucial for improving patient outcomes. Artificial Intelligence (AI) has shown significant promise in assisting dermatologists with accurate and efficient diagnosis through automated skin disease classification. This systematic review aims to provide a comprehensive overview of the various AI techniques employed for skin disease classification, focusing on their effectiveness across different datasets and methodologies. A total of 220 articles were initially identified from databases such as Scopus and IEEE Xplore. After removing duplicates and conducting a title and abstract screening, 213 studies were assessed for eligibility based on predefined criteria such as study relevance, clarity of results, and innovative AI approaches. Following full-text review, 56 studies were included in the final analysis. These studies were categorized based on the AI techniques used, including Convolutional Neural Networks (CNNs), Transformer-based models, hybrid models combining CNNs with other techniques, Generative Adversarial Networks (GANs), and ensemble learning approaches. The review highlights that the ISIC dataset and its variations are the most commonly used data sources, owing to their extensive and diverse collection of dermoscopic images. The results indicate that CNN-based models remain the most widely adopted and effective approach for skin disease classification, with several hybrid and Transformer-based models also demonstrating high accuracy and specificity. Despite the advancements, challenges such as dataset variability, the need for more diverse training data, and the lack of interpretability in AI models persist. This review provides insights into current trends and identifies future directions for research, emphasizing the importance of integrating AI into clinical practice for improved skin disease management.**

Keywords*—Skin Disease Classification; Artificial Intelligence (AI); Convolutional Neural Networks (CNNs); Transformer-based Models; Generative Adversarial Networks (GANs); ensemble learning; hybrid models; ISIC dataset; dermatology; machine learning; deep learning; skin cancer detection; dermoscopic images; medical imaging; systematic review

I. Introduction

Skin diseases encompass a broad spectrum of conditions that affect the skin, which is the largest organ of the human body. These conditions can vary greatly in severity and presentation, ranging from benign issues like acne to life-threatening diseases such as melanoma. Accurate diagnosis, classification, and segmentation of skin diseases are critical as they directly influence treatment plans, patient outcomes, and overall healthcare efficiency. Misdiagnosis or delayed diagnosis can lead to severe consequences, including unnecessary treatments or the progression of the disease to more advanced stages [1], [2].

Common skin diseases that are frequently the focus of classification and segmentation efforts in medical research include melanoma, psoriasis, and eczema. Melanoma, in particular, is a malignant tumor of melanocytes and is one of the most serious forms of skin cancer due to its high potential for metastasis [3]. Psoriasis is a chronic inflammatory skin condition characterized by the rapid growth of skin cells, leading to thick, red, scaly patches [4]. Eczema, also known as atopic dermatitis, is a chronic condition that causes inflamed, itchy, cracked, and rough skin [5].

Traditionally, the diagnosis of skin diseases has relied heavily on clinical examinations performed by dermatologists. This process typically involves visual inspection, often aided by tools like dermoscopy, which provides magnified images of the skin, allowing for better visualization of structures beneath the skin surface [6]. In cases where the visual inspection is inconclusive, a biopsy may be performed, wherein a sample of the skin is taken for histopathological examination under a microscope [7].

While these traditional methods are well-established and widely used, they are not without limitations. Human error is a significant concern, as the accuracy of diagnosis can vary depending on the dermatologist's experience and expertise. Studies have shown variability in diagnostic accuracy, even among experienced dermatologists [8]. Additionally, the manual segmentation of skin lesions, which is crucial for treatment planning, is time-consuming and labor-intensive. This process often involves delineating the borders of the lesion manually, which is not only subjective but also prone to variability [9].

Given these challenges, there is a growing interest in the application of artificial intelligence (AI) to improve the accuracy, efficiency, and consistency of skin disease diagnosis and segmentation. A significant body of research has emerged that integrates AI techniques into the detection and segmentation of skin diseases, demonstrating promising results in enhancing diagnostic processes. This paper proposes a systematic review focusing on the use of AI for the classification and segmentation of skin diseases, aiming to consolidate current findings, identify prevailing methodologies, and highlight existing knowledge gaps within this rapidly evolving field. By systematically analyzing the literature, this review will provide valuable insights into the effectiveness of AI in dermatology, outline the strengths and limitations of current approaches, and suggest potential avenues for future research.

The remainder of this paper is organized as follows: The Theoretical Background section provides a comprehensive overview of the foundational concepts and existing literature related to AI-driven skin disease classification. This is followed by the Methodology section, where the criteria and procedures used to select and analyze the studies are detailed. The Results section presents the findings from the systematic review, including bibliometric analyses and the evaluation of methodologies. In the Discussion section, the implications of the findings are explored, and the strengths and limitations of current approaches are assessed. Finally, the paper concludes with Final Considerations, summarizing the key takeaways and suggesting directions for future research in this rapidly evolving field.

II. Theoretical Background

A. From Classic to Modern Dermatology

Dermatology has undergone a profound transformation from its early days of visual inspection and basic histopathological analysis to the incorporation of advanced imaging technologies and digital tools. Traditionally, dermatologists relied heavily on their clinical expertise, using simple tools like magnifying glasses to examine skin lesions, and performing biopsies followed by histopathological analysis to diagnose complex conditions. Histopathological images, derived from these biopsies, provided detailed views of skin tissue at the cellular level, becoming a cornerstone of accurate diagnosis, particularly for skin cancers [10].

Over time, the limitations of these conventional methods, including their invasiveness and the potential for diagnostic variability, drove the development of more sophisticated tools. The advent of dermoscopy revolutionized non-invasive examination by enabling magnified visualization of subsurface skin structures, greatly enhancing the accuracy of initial assessments [11]. This was followed by the introduction of digital dermoscopy, confocal microscopy, optical coherence tomography (OCT), and multispectral imaging, each offering unique insights into different aspects of skin anatomy and pathology [12], [13]. These technological advancements have not only improved traditional diagnostic practices but have also set the stage for the integration of artificial intelligence, which leverages these diverse imaging modalities to further revolutionize dermatological care.

These diverse imaging modalities—ranging from high-resolution dermoscopic images to detailed histopathological and confocal microscopy images—have provided a wealth of data that is now being harnessed by artificial intelligence (AI) to further advance the field. AI systems, trained on these extensive image datasets, are capable of analyzing and interpreting complex patterns within the skin, leading to more precise and efficient diagnoses [14], [15]. This integration of AI with a wide variety of imaging techniques is not only enhancing diagnostic performance but also paving the way for more personalized and effective treatment strategies in dermatology.

B. Artificial Intelligence and Skin Disease Classification

Building on the advancements in imaging technologies, the integration of artificial intelligence into dermatology marks a significant leap forward in skin disease classification. AI, through machine learning and deep learning algorithms, leverages the vast array of imaging data—from clinical and dermoscopic images to histopathological and confocal microscopy images—to identify subtle patterns and features that might escape human detection. This capability allows AI to offer unprecedented accuracy and efficiency in the diagnosis and classification of skin conditions, setting the stage for more personalized and precise dermatological care [16].

Artificial Intelligence (AI) is a broad field that includes machine learning (ML) and deep learning (DL), both of which have become crucial in healthcare, particularly in dermatology. ML involves algorithms that learn from data to make predictions, while DL, a more advanced subset, uses multi-layered neural networks to automatically extract features from complex datasets. These AI techniques are particularly well-suited for analyzing skin images, enabling more accurate and efficient diagnosis of dermatological conditions by identifying patterns that may be difficult for human clinicians to detect [17], [14], [16].

Machine learning (ML) initially made significant strides in dermatology by enabling the automated classification of skin lesions based on manually extracted features. Techniques such as support vector machines (SVMs), decision trees, and random forests were effectively used to analyze images and identify patterns that distinguish between benign and malignant conditions [14]. One notable success was the use of SVMs in melanoma detection, where these models achieved high accuracy by focusing on key features like color, texture, and shape [16]. Despite these advancements, ML models often required extensive feature engineering, relying on expert knowledge to select the most relevant attributes. This limitation, combined with the models' relatively shallow architecture, made it challenging to handle the complex and high-dimensional data typical in dermatology, which paved the way for the adoption of deep learning (DL).

Machine learning was the first wave of AI to make a substantial impact in dermatology. By training algorithms on datasets of labeled skin images, ML models have been able to assist in diagnosing various skin conditions. For example, support vector machines (SVMs), decision trees, and ensemble methods like random forests have been utilized to classify skin lesions based on features extracted manually or through basic automated processes [14]. These models improved diagnostic consistency and reduced human error, particularly in distinguishing between benign and malignant lesions. However, the effectiveness of ML in dermatology was often limited by the need for extensive feature engineering and the relatively shallow nature of these models, which struggled to capture the complexity of high-dimensional image data.

The introduction of deep learning marked a significant leap forward for AI in dermatology. Deep learning, particularly through convolutional neural networks (CNNs), overcame many of the limitations of traditional ML models by automatically learning hierarchical features directly from raw image data. CNNs, with their ability to process and analyze large amounts of image data, have been particularly effective in dermatology, where they are used to classify skin diseases with unprecedented accuracy [17]. These models have been trained on vast datasets of clinical, dermoscopic, and histopathological images, enabling them to recognize subtle patterns and features that might be missed by human clinicians or simpler algorithms.

As AI continues to evolve, its applications in dermatology are expected to expand, addressing current challenges such as model interpretability and data bias, while opening new avenues for personalized and accessible skin care.

III. Related Work

The application of machine learning (ML) and deep learning (DL) techniques in the classification of skin diseases has garnered significant attention in recent years, leading to the development of various models aimed at improving diagnostic accuracy and efficiency. Traditional ML methods such as Support Vector Machines (SVM), Random Forests (RF), and k-Nearest Neighbors (k-NN) have been extensively utilized for skin disease classification tasks. For instance, [14] employed SVM in conjunction with handcrafted features for melanoma detection, achieving competitive performance against dermatologists. However, these approaches are often limited by their dependence on feature engineering, which can be both time-consuming and reliant on domain expertise.

With the advent of DL, particularly Convolutional Neural Networks (CNNs), there has been a paradigm shift in the approach to skin disease classification. CNNs have the inherent ability to automatically learn hierarchical feature representations from raw image data, eliminating the need for manual feature extraction.

Over the past decade, several systematic reviews have been conducted to assess the efficacy of machine learning (ML) and deep learning (DL) techniques in the classification of skin diseases. These reviews have provided valuable insights into the trends, challenges, and future directions in this rapidly evolving field. The author in [18] conducted one of the early comprehensive reviews, focusing on the application of CNNs in dermatology. This review highlighted the increasing adoption of deep learning models over traditional machine learning approaches due to their superior accuracy and ability to process raw image data without extensive preprocessing.

Further advancing the field, [19] provided a thorough review of dermatological image analysis using both machine learning and deep learning techniques. The review emphasized the significant advancements in CNN architectures, such as ResNet, AlexNet, VGG.., and their impact on improving diagnostic accuracy for various skin conditions, including common cases and rare ones. However, the authors also noted the limitations related to the interpretability of these models and the challenges posed by imbalanced datasets.

[20] conducted a meta-analysis focused specifically on the performance comparison between human dermatologists and DL models. Their review concluded that DL models, particularly those based on CNNs, have reached a level of performance comparable to that of expert dermatologists, especially in the detection of malignant melanoma. This finding was corroborated by the review conducted by [21], which compared multiple DL models and found that ensemble methods often outperform individual models in terms of accuracy and robustness.

A more recent review by [22] explored the integration of advanced techniques such as transfer learning and generative adversarial networks (GANs) into dermatological applications. The authors highlighted that while these techniques offer promising avenues to overcome the challenges of limited labeled data and improve model generalizability, there is still a need for more standardized evaluation protocols and larger, more diverse datasets to fully realize their potential.

In addition to these, [23] reviewed the ethical and regulatory considerations associated with the deployment of ML and DL models in clinical settings. Their work underscores the importance of ensuring model transparency, patient data privacy, and the need for rigorous clinical validation before these models can be widely adopted in practice.

These reviews collectively illustrate the rapid advancements and the ongoing challenges in applying ML and DL techniques to skin disease classification. They provide a foundation for future research, particularly in addressing issues related to model interpretability, dataset bias, and the ethical implications of AI in dermatology.

In contrast, this systematic review aims to fill a specific gap in the literature by focusing on the progression from traditional ML techniques to advanced DL models. Process was to systematically search, extract, and analyze studies that detail the exact methodologies and techniques used in both ML and DL for skin disease classification. This approach allows us to map the evolution of these techniques, highlighting how deep learning, particularly CNNs, Transformers, and hybrid models, has been increasingly adopted and refined over time.

A notable particularity of this review is the attention given to hybrid models that combine both machine learning and deep learning techniques. This focus is crucial, as hybrid models represent a significant trend in the literature, often outperforming their pure ML or DL counterparts by leveraging the strengths of both approaches. These models' architectures and performances were meticulously documented, thereby offering insights into their potential for future research and clinical applications, by tracing the trajectory from early adoption phases to the more recent innovations, such as Vision Transformers (ViTs) and Generative Adversarial Networks (GANs), providing a clear picture of the technological advancements and their impact on classification performance.

This systematic review stands out in its detailed exploration of the transition from ML to DL in skin disease classification. It not only contributes to the existing body of knowledge but also serves as a valuable resource for researchers aiming to further advance the field.

IV. Methodology

A. Aim and Scope Definition

The primary objective of this study is to systematically review and synthesize the existing literature on the application of artificial intelligence (AI), including machine learning (ML) and deep learning (DL), in the classification of skin diseases. The study aims to explore how these AI techniques have been utilized in analyzing various types of dermatological images, including clinical, dermoscopic, and histopathological images, to enhance diagnostic accuracy and efficiency. The scope of the review includes original research articles, and review papers published in peer-reviewed journals. The study focuses on skin diseases such as melanoma, psoriasis, and acne, with no specific geographical or linguistic restrictions, although only English-language publications are included.

B. Data Collection

1) Retrieval database: The literature for this systematic review was retrieved from several databases known for their comprehensive coverage of medical and technological research. The primary databases used include IEEE Xplore and Scopus. These databases were selected due to their relevance to the fields of dermatology, artificial intelligence, and medical imaging, ensuring that a broad range of studies could be captured.

2) Keyword selection strategy: A strategic keyword selection process was employed to identify relevant studies. Keywords were chosen to reflect the core concepts of the study: artificial intelligence, machine learning, deep learning, and skin disease classification. Specific search terms included combinations of the following: “AI in dermatology”, “deep learning”, “machine learning”, “classification”. Boolean operators (AND, OR, NOT) were used to refine the search and ensure comprehensive coverage of the literature.

Research query for scopus = (TITLE-ABS-KEY (“skin diseases” OR “dermatological disorders” OR “skin conditions”) AND (“machine learning” OR “deep learning”) AND (“classification”) AND TITLE-ABS-KEY (“segmentation”) AND PUBYEAR ¿ 2020 AND PUBYEAR ¡ 2024 AND [LIMIT-TO (LANGUAGE, “English”)

Research query for IEEE Xplore = (“All Metadata”: “skin diseases” OR “All Metadata”: “dermatological disorders” OR “All Metadata”: “skin conditions” OR “All Metadata”: “skin lesions”) AND (“All Metadata”: “machine learning” OR “All Metadata”: “deep learning”) AND (“All Metadata”: “classification”) NOT (“All Metadata”: “segmentation”)] (Table I).

C. Inclusion and Exclusion Criteria

1) Inclusion criteria: To ensure the relevance and quality of the included studies, the following inclusion criteria were applied:

- The study focuses on the application of AI (ML or DL) in the classification of skin diseases.
- The study is published in a peer-reviewed journal.
- The study provides sufficient data for analysis, including details of the AI methods used and the types of images analyzed.
- The study is written in English.

TABLE I. RESEARCH QUESTIONS OF THE STUDY

No.	Research Question (RQ)
RQ1	What are the most commonly used artificial intelligence models for the classification of skin diseases?
RQ2	How effective are these AI models in accurately classifying various skin diseases?
RQ3	How do AI-based methods for skin disease classification compare to traditional diagnostic methods ?
RQ4	What datasets are commonly used to train and validate AI models for skin disease classification, and what are their characteristics?
RQ5	What are the potential clinical implications of integrating AI models into the diagnosis and treatment planning of skin diseases?
RQ6	What are the main challenges and limitations associated with the use of AI models in the classification of skin diseases?
RQ7	What future research directions are needed to improve the performance and clinical applicability of AI models in skin disease classification?

2) Exclusion criteria: Studies were excluded based on the following criteria:

- The study does not focus on dermatology or AI applications in skin disease classification.
- The study is a conference abstract, editorial, letter, or opinion piece with no original research data.
- The study lacks methodological rigor, as determined by the quality assessment process.
- The study does not utilize one of the following datasets: “ISIC”, “HAM10000”, “PH2”, “Dermnet” or “Derm7pt”.
- The study is about the segmentation of skin diseases.
- The study is not available in English.

In this systematic review, a total of 220 articles were initially identified through comprehensive database searches in Scopus and IEEE Xplore. After the removal of 7 duplicate articles, 213 unique records remained for screening.

The initial screening process led to no exclusions at this stage. All 213 records were then assessed for eligibility. Subsequently, 109 articles were excluded due to various reasons, such as lack of clear results reporting (11 articles), non-novel approaches (11 articles), irrelevant content (15 articles), or falling out of the scope of the review (14 articles). This left 104 articles for further retrieval attempts, out of which 48 were not retrieved. Ultimately, 56 articles were included in the final review and deemed suitable for qualitative synthesis (Fig. 1).

3) Software tools: The data analysis was performed using a combination of software tools, including Microsoft Excel for data management and descriptive analysis, R for statistical analysis, and Bibliometrix with Biblioshiny package for qualitative analysis. These tools were chosen based on their functionality, ease of use, and ability to handle large datasets effectively.

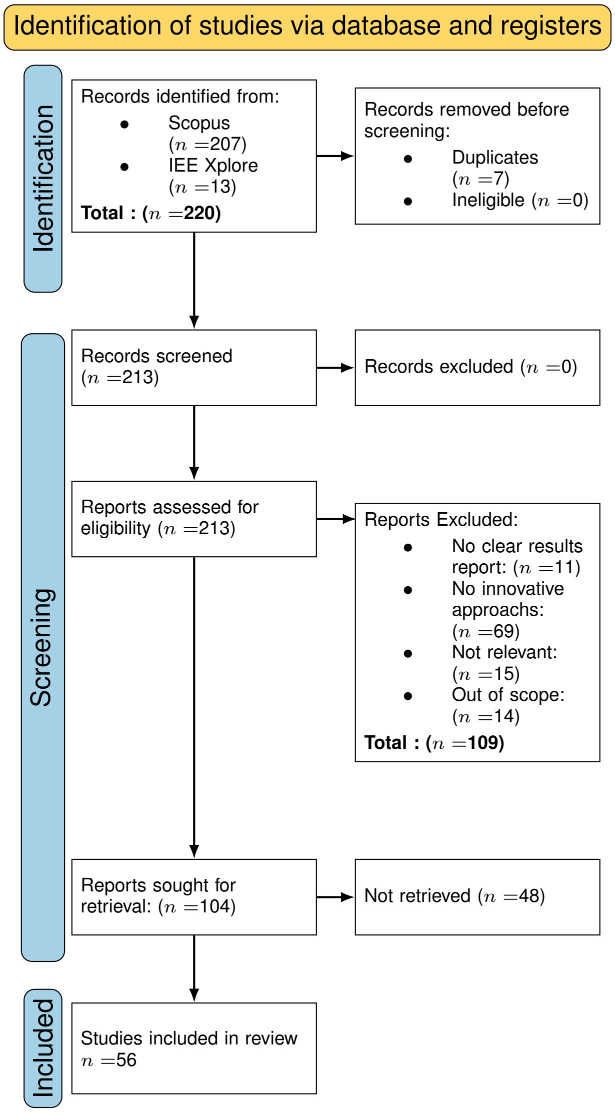

Fig. 1. Prisma diagram.

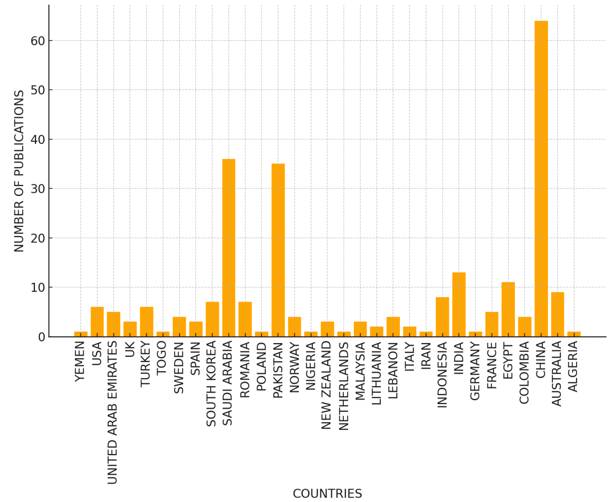

Fig. 2. Distribution of selected papers by countries.

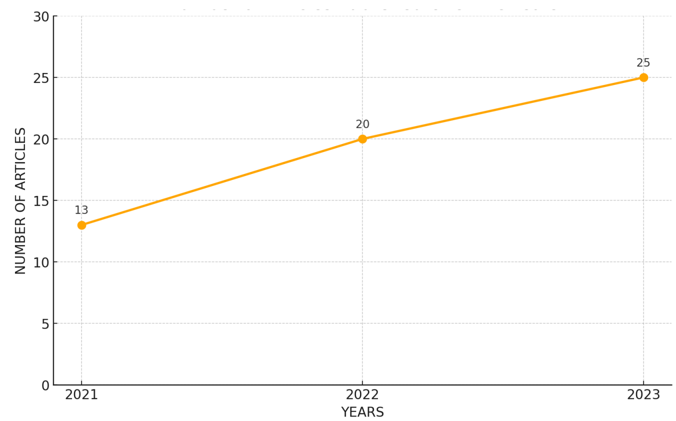

Fig. 3. Number of articles by years of publication.

V. Results

Retained articles related to the classification of skin diseases using machine learning and deep learning tools were published across 31 different countries.

Fig. 2 illustrates the distribution of included works by countries of publication. The countries representing the origin of the greatest number of publications are China with 64 articles, followed by Pakistan with 35 articles, and Saudi Arabia with 36 articles. Additionally, India and Indonesia contributed significantly with 13 and 8 articles, respectively. Together, China, Pakistan, Saudi Arabia, India, and Indonesia account for over 80% of the papers included in this study, with a total of 56 papers.

Considering publication dates of selected articles, a large part of the retained documents was published during the period 2020 to 2023, with a total of 45 papers. Fig. 3 below represents the curve of the publication's evolution per years between 2021 and 2023.

The bibliometric analysis also includes a thematic map in Fig. 4, which offers a visual representation of the key research themes within the field of AI-driven skin disease classification. The map categorizes themes based on their relevance (centrality) and development degree (density), providing a clear overview of the research landscape. This analysis highlights "deep learning" and "dermatology" as highly relevant but still developing themes, suggesting ongoing growth in these areas. Conversely, "human melanoma" appears as a well-established motor theme, indicating its foundational role in the field. Emerging or declining themes such as "optimization algorithms" are also identified, pointing to areas where future research may be necessary.

To further explore the thematic connections within the literature, Fig. 5 presents a co-occurrence network of keywords. This visualization highlights the central themes in AI-driven skin disease classification, with "deep learning" and "skin cancer" emerging as the most connected and frequent terms. The network also reveals the relationships between various methodologies, such as "convolutional neural networks" and "transfer learning", underscoring their significance in this research domain.

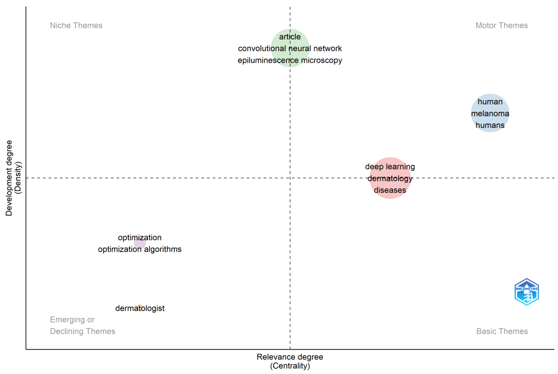

Fig. 4. Thematic map of research in AI-driven skin disease classification.

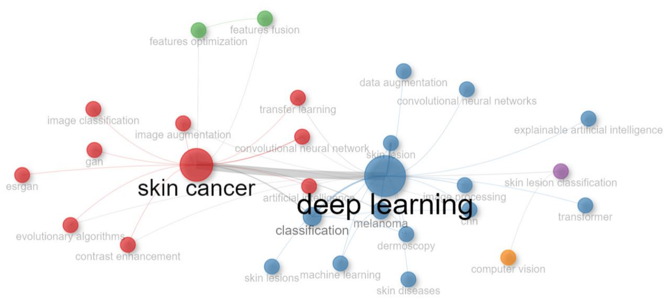

Fig. 5. Co-occurrence network of keywords in AI-driven skin disease classification literature. This network visualizes the relationships between key terms, with larger nodes indicating more frequently occurring keywords. The network highlights the centrality of terms like “deep learning” and “skin cancer”, revealing their strong connections with related concepts such as “image augmentation”, “transfer learning”, and “convolutional neural networks”.

RQ1: What are the most commonly used artificial intelligence models for the classification of skin diseases?

The qualitative analyse of the obtained results from the researches papers provided, indicates that Convolutional Neural Networks (CNNs) are the most frequently employed AI models for skin disease classification. CNNs appear in various forms, including standalone models, hybrid approaches (e.g. CNN combined with SVM, XGBoost, or Random Forest), and enhanced architectures (e.g. Inception-V3, EfficientNets). Additionally, Transformer-based models and Vision Transformers (ViTs) are increasingly utilized, reflecting a shift towards more complex, attention-based architectures. Hybrid models integrating multiple techniques, such as CNNs with transformers or Extreme Learning Machines (ELM), are also prevalent, suggesting an ongoing effort to enhance model performance through combining strengths of different methods.

RQ2: How effective are these AI models in accurately classifying various skin diseases?

The reviewed studies report high accuracy rates across different AI models, with CNN-based approaches often achieving accuracy levels above 90%. Specific examples include:

- A deep learning-based auto-encoder with an accuracy of 98.7%.
- A CNN combined with XGBoost and SVM showing 97.85% accuracy.
- Hybrid CNN-ELM models achieving up to 96.7% accuracy.

Sensitivity and specificity metrics are also robust, frequently exceeding 90%. For example, the sensitivity for CNN models ranges from 88.3% to 98.46%, while specificity ranges from 90% to 100%. These results indicate that AI models, particularly CNNs and hybrid approaches, are effective in accurately classifying skin diseases, often outperforming traditional diagnostic approaches.

RQ3: How do AI-based methods for skin disease classification compare to traditional diagnostic methods?

The analyse of the provided researches reveals that deep learning-based methods, particularly Convolutional Neural Networks (CNNs), consistently outperform traditional machine learning techniques in skin disease classification across various metrics. Deep learning models achieve higher accuracy, often exceeding 94%, with some models reaching up to 96.83% ([24], [25]). In contrast, traditional machine learning methods, including Support Vector Machines (SVM) and Random Forests, typically demonstrate accuracy within the 90% to 97% range.

Sensitivity and specificity are also higher in deep learning models. For instance, [26] reports a sensitivity of 98% and specificity of 98.1% using a Lightweight CNN with dynamic-sized kernels and ReLU/leaky ReLU activations, while machine learning approaches generally show slightly lower performance in these metrics.

Moreover, deep learning models benefit from data augmentation techniques, such as Generative Adversarial Networks (GANs), which enhance their generalization capabilities, especially in imbalanced datasets ([27]). Traditional machine learning methods, although effective, often require more extensive feature engineering and do not leverage data augmentation as effectively as deep learning approaches.

Hybrid models that combine deep learning with machine learning methods, such as CNNs with SVM or Random-Forest offer a balanced approach, leveraging the strengths of both techniques ([28], [29]). However, these hybrids still typically fall short of purely deep learning-based methods in terms of overall performance.

In terms of model interpretability, traditional machine learning models, particularly decision trees, provide more straightforward explanations. However, the integration of Explainable AI (XAI) techniques with CNNs has begun to address the “black box” nature of deep learning models, enhancing their transparency ([30]).

While machine learning models still hold value in scenarios requiring interpretability, deep learning approaches, particularly when augmented with hybrid techniques, represent the most effective tools for skin disease diagnostics.

RQ4: What datasets are commonly used to train and validate AI models for skin disease classification, and what are their characteristics?

Several datasets emerge as commonly utilized benchmarks for training and validation purposes for skin disease classification task. The following are the key datasets frequently employed in these studies, along with their characteristics:

- ISIC (International Skin Imaging Collaboration): This dataset is a comprehensive repository of dermoscopic images that has been widely adopted due to its diversity and scale [31]. The ISIC dataset includes several variations, such as ISIC2018, ISIC2019, ISIC2020, ISIC2017, and ISIC2008, each corresponding to different years of challenge submissions. These variations contain images that differ in terms of disease types, resolutions, and annotations, providing a robust foundation for model development (Fig. 6).

- HAM10000: The HAM10000 dataset [32] contains a large collection of dermatoscopic images with an emphasis on the most common pigmented skin lesions. It is particularly valued for its balanced representation of multiple classes of skin diseases, making it a reliable resource for training classifiers that need to generalize across various conditions.

- PH2: Although smaller, the PH2 dataset is a key resource that provides high-quality dermoscopic images specifically curated for the assessment of melanocytic and non-melanocytic skin lesions [33]. Its limited class diversity is counterbalanced by the precise annotations and image quality, making it ideal for specialized classification tasks.

- Derm7pt: This dataset focuses on a seven-point checklist system for melanoma detection, providing a structured approach to training models in clinical decision-making scenarios [34].

- PAD-UFES-20: This dataset includes images collected from a specific demographic, aiding in the development of models that are more inclusive and adaptable to different population groups [35].

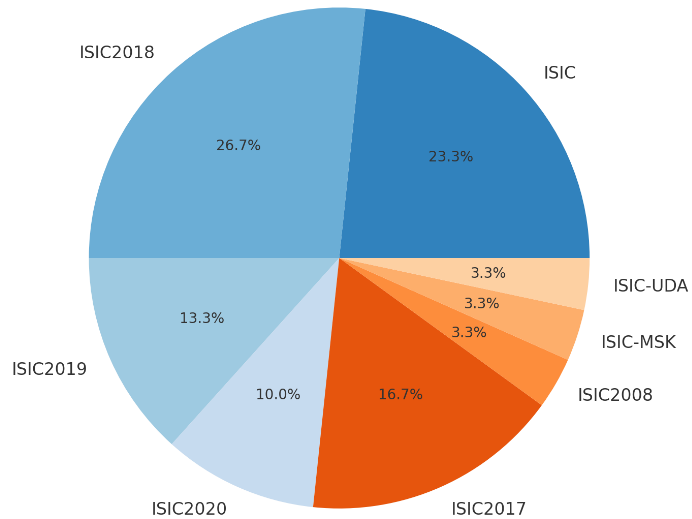

Fig. 6. Usage distribution of ISIC dataset variants across studies.

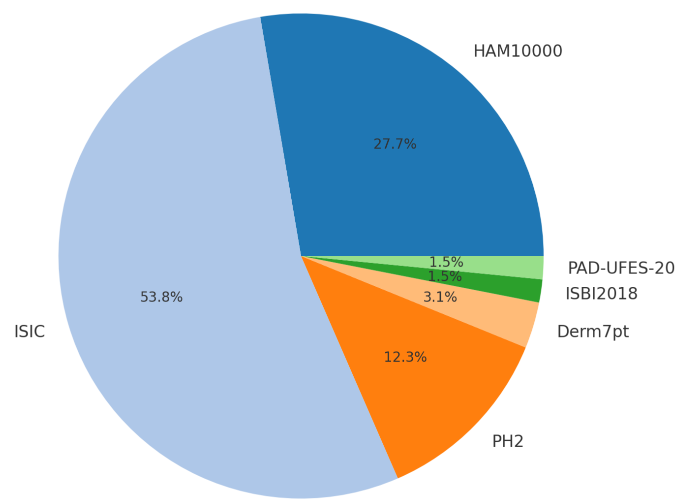

Fig. 7. Percentage of Dataset Usage in the Selected Studies.

Fig. 7 summarizes the usage distribution of these datasets across the reviewed studies, highlighting the dominance of the ISIC and its variants, followed by the widespread adoption of HAM10000 and other datasets.

RQ5: What are the potential clinical implications of integrating AI models into the diagnosis and treatment planning of skin diseases?

The high accuracy and specificity of AI models indicate significant potential for their integration into clinical practice. AI can assist dermatologists in diagnosing skin diseases more quickly and accurately, reducing the time required for analysis and potentially improving patient outcomes. Moreover, the precision of AI models could help in early detection of malignant lesions, leading to more timely interventions. However, the clinical integration of these models also requires careful consideration of ethical implications, including patient consent, data privacy, and the potential for algorithmic bias. The integration of AI into treatment planning could also extend to personalized medicine, where AI-driven insights help tailor treatment strategies to individual patient profiles.

RQ6: What are the main challenges and limitations associated with the use of AI models in the classification of skin diseases?

Despite the promising results, several challenges and limitations persist. One significant challenge is the generalizability of AI models across different populations and clinical settings. Most models are trained on specific datasets, which may not fully represent the diversity of skin types and conditions found in the broader population. Data quality and annotation is another concern; the accuracy of AI models heavily depends on the quality of the input data, and misannotations or inconsistencies can lead to reduced model performance. Moreover, computational complexity and the need for high-end hardware for training and deploying complex models like transformers could limit their accessibility in resource-constrained environments. Additionally, there is a need for explainability in AI models to ensure that clinicians can trust and understand the decisions made by these systems. Finally, there is limited research on how these models can be effectively integrated into real-world clinical workflows. Understanding the challenges of AI adoption in everyday clinical practice, including clinician trust, usability, and workflow compatibility, is essential for the successful implementation of AI in dermatology [36].

RQ7: What future research directions are needed to improve the performance and clinical applicability of AI models in skin disease classification?

Future research in AI for skin disease classification should prioritize several key areas to enhance both performance and clinical applicability. First, increasing dataset diversity and size is critical for improving model generalization, particularly for underrepresented skin conditions and demographic groups, ensuring broader applicability across diverse populations. Second, cross-validation in varied clinical settings is necessary to assess the robustness and consistency of AI models, which is essential for their practical integration into healthcare environments. Third, improving the explainability of AI models is vital for building trust among healthcare providers, enabling transparent and verifiable diagnoses that can encourage clinical adoption.

Additionally, research should focus on reducing the computational complexity of AI models, making them more accessible and deployable, especially in resource-limited settings. The development of multi-modal diagnostic systems that integrate visual data with patient-specific information, such as demographics and medical history, is another crucial area. These systems could enhance diagnostic accuracy and lead to more personalized treatments.

Moreover, most current research focuses on static images and one-time predictions. There is a gap in studies that investigate the application of AI in longitudinal analyses, where the progression of skin diseases is tracked over time. Such studies are crucial for developing AI tools that can not only diagnose but also monitor disease progression and treatment response [37], [38].

VI. Discussion

The analysis of methodologies employed across the studies included in this systematic review reveals distinct patterns in the application of various AI techniques for skin disease classification. To provide a more nuanced understanding of the approaches utilized by different researchers, the studies have been categorized based on the general technique employed, including Machine Learning techniques (e.g. XGBoost, SVM), Generative Adversarial Networks (GANs), Ensemble Learning approaches, Multi-modal methods, Transformer-based models, and Hybrid models combining Convolutional Neural Networks (CNNs) with other techniques.

The following sections summarize the methodologies applied within each of these categories, highlighting the specific strategies used in feature extraction, model training, and validation. This organization elucidates the strengths and limitations inherent to each approach, offering a clearer perspective on the current state of AI-driven skin disease classification. By structuring the analysis in this manner, a direct comparison between different techniques is facilitated, providing insights into potential future directions for research in this rapidly evolving field.

A. Studies Using Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) have been explored for their ability to generate synthetic data and improve classification tasks in skin disease detection.

Two notable studies have employed Generative Adversarial Networks (GANs) to enhance skin lesion classification. Abdelhalim et al. [39] combined GANs with Convolutional Neural Networks (CNNs) in their study, utilizing the ISIC dataset. In this approach, GANs were employed to generate synthetic dermoscopic images, which were subsequently used to train the CNN. This method effectively improved the model's robustness by augmenting the training data.

Similarly, Su et al. [27] adopted a GAN-CNN hybrid model for skin lesion classification. In their study, the GAN was used to generate additional training samples, while the CNN handled the classification task. This methodology was validated using the HAM10000 dataset, and the results demonstrated improved accuracy, attributed to the enriched dataset provided by the GAN.

B. Studies Using Ensemble Learning Approaches

Ensemble learning approaches combine multiple models to improve classification accuracy and robustness. The following table provides an overview of studies that utilized ensemble learning techniques in skin disease classification.

Several studies have utilized ensemble learning approaches to enhance the classification of skin lesions. Popescu et al. [40] proposed an ensemble learning methodology that involved combining multiple models to classify skin lesions using the ISIC dataset. In this approach, different base models were trained separately, and their predictions were subsequently combined through majority voting or weighted averaging, leading to higher accuracy.

Similarly, Thurnhofer-Hemsi et al. [41] applied an ensemble approach to improve the accuracy of skin lesion classification on the HAM10000 dataset. This method focused on reducing the variance and bias inherent in single models by integrating the strengths of multiple classifiers, ultimately enhancing the overall classification performance.

C. Studies Using Multi-modal Approaches

Multi-modal techniques have been explored in several studies to enhance the accuracy of skin disease classification by integrating various data sources. Fu et al. [42] utilized a graph nodes-based approach that incorporated multi-modal data from different sources, including the 7-point checklist, ISIC2017, and ISIC2018 datasets. This methodology focused on combining different types of data to improve the classification of skin lesions.

Zhang et al. [43] introduced TFormer, a throughout fusion transformer model for multi-modal skin disease classification. The transformer model served as the feature extraction backbone, fusing both image and metadata, which significantly enhanced the accuracy of classification.

Roge et al. [28] employed a deep ensemble learning approach that integrated demographic information with image data to improve classification performance. This study combined the use of a demographic machine with standard imaging techniques, leading to superior results in classification accuracy.

Cai et al. [44] applied a multi-modal transformer model that fused images and metadata, leveraging the strengths of Vision Transformer (ViT) in handling multi-modal data. This approach provided a notable boost in classification accuracy by effectively combining different data sources.

Nguyen et al. [45] utilized a combination of deep learning models (DenseNet) and traditional machine learning algorithms (SVM) to classify skin lesions on imbalanced datasets. By incorporating both image data and patient metadata, this approach improved model robustness and accuracy.

Yin et al. [46] introduced the MetaNet module, which employs multi-modal data from both images and metadata to enhance skin tumor classification. The comprehensive analysis enabled by this multi-modal approach resulted in improved classification performance.

D. Studies Using Transformer-Based Models

Transformer-based models, originally developed for natural language processing, have been adapted for image classification tasks due to their ability to capture long-range dependencies in data.

Several studies have explored the use of Transformer-based models for skin lesion classification, showcasing the versatility and effectiveness of this approach. Ding et al. [47] employed a Vision Transformer (ViT) model on the ISIC2018, ISIC2017, and PH2 datasets. The methodology capitalized on the Transformer's ability to process image patches, which provided an effective solution for skin lesion classification.

Zhang et al. [43] proposed a Transformer-based model tailored for the Derm7pt dataset. Their approach emphasized the model's capability to learn global contextual information, which is crucial for accurately classifying skin lesions.

Abbas et al. [48] utilized a Transformer-based model to classify dermoscopic images from the ISIC dataset. The study highlighted the use of self-attention mechanisms inherent in Transformers, which allowed the model to capture intricate patterns in the images, thereby improving classification performance.

Cai et al. [44] applied a Vision Transformer (ViT) to the ISIC2018 dataset, focusing on the model's ability to process images in a patch-wise manner. Their study demonstrated the effectiveness of the Transformer architecture in handling large-scale image data, resulting in enhanced classification accuracy.

Yang et al. [49] also employed a Transformer-based model, specifically for the HAM10000 dataset. This study leveraged the self-attention mechanism of the Transformer to capture the intricate details of skin lesions, thereby improving classification accuracy.

Aladhadh et al. [50] used a Transformer-based model to classify images from the HAM10000 dataset. Their focus was on enhancing the model's sensitivity and specificity by taking advantage of the deep attention mechanisms inherent in Transformer architectures.

E. Studies Using Hybrid Models (CNN with Other Techniques)

Hybrid models that combine CNNs with other techniques, such as machine learning algorithms or other deep learning methods, have shown potential in improving classification accuracy.

Hybrid models that combine Convolutional Neural Networks (CNNs) with other techniques have been widely explored for improving the classification of skin lesions. Ma et al. [24] proposed a hybrid approach that integrated CNN with Random Forest for the classification of skin lesions using the HAM10000 dataset. In this study, the CNN was responsible for deep feature extraction, which were then classified by the Random Forest, combining the strengths of both methods.

Similarly, Roge et al. [28] developed a CNN-Random Forest hybrid model for the ISIC dataset (HAM10000). Their methodology also involved using CNN for initial feature extraction, followed by Random Forest for final classification, which enhanced both the accuracy and robustness of the model.

Nie et al. [51] explored a hybrid model that combined CNN with Transformer architecture, specifically for skin lesion classification using the ISIC2018 dataset. In this study, the CNN component handled initial feature extraction, while the Transformer captured global dependencies in the image data, leading to improved classification performance.

Li et al. [52] also utilized a hybrid CNN-SVM approach for the ISIC2019 dataset. In this study, the CNN was used for initial feature extraction, with SVM refining the classification process to enhance the accuracy of multi-class skin lesion classification.

Tahir et al. [29] proposed a hybrid model that combined CNN with XGBoost and Support Vector Machine (SVM) for classifying skin lesions in the ISIC2017 dataset. The CNN was utilized to extract relevant features, with XGBoost and SVM models performing the classification, leveraging the strengths of both deep learning and traditional machine learning techniques.

Khan et al [53] proposes a fully automated system for multiclass skin lesion localization and classification using deep learning. To address class imbalance in the HAM10000, ISBI2018, and ISBI2019 datasets, the approach fine-tunes a pre-trained DarkNet19 model and fuses visualized images using a High-Frequency approach with a Multilayered Feed-Forward Neural Network (HFaFFNN). Two additional models, DarkNet-53 and NasNet-Mobile, are trained using transfer learning on localized lesion images. Features are fused using the parallel max entropy correlation (PMEC) technique, and the entropy-kurtosis controlled whale optimization (EKWO) algorithm selects the most discriminant features. The model achieves accuracies of 95.8%, 97.1%, and 85.35% on the HAM10000, ISBI2018, and ISBI2019 datasets, respectively. Zhou et Arandian. [54] introduce a new computer-aided skin cancer diagnosis method that combines deep learning with the Wildebeest Herd Optimization (WHO) algorithm. Inception CNN is used for feature extraction, followed by the WHO algorithm for feature selection to reduce analysis complexity. The method was tested on the ISIC-2008 dataset and compared with three other algorithms, demonstrating superior results.

Annaby et al. [55] proposes a melanoma detection approach that combines graph-theoretic representations with conventional image features to improve detection performance. Superpixels from dermoscopic images are used as graph nodes, and edges connect adjacent superpixels based on feature descriptor distances. Features are extracted from both weighted and unweighted graph models in the vertex and spectral domains, as well as from color, geometry, and texture. Various classifiers were trained on these feature combinations using ISIC datasets. The proposed system demonstrated significant improvements in accuracy, AUC, specificity, and sensitivity in detecting melanoma. Saeed et al. [56] employed a hybrid model that combined CNN with SVM for the classification of skin lesions using the ISIC2019 and ISIC2020 datasets. In this approach, the CNN handled feature extraction, while the SVM classifier provided enhanced decision boundaries, leading to improved model performance.

Afza et al. [57] introduced a hybrid CNN-Extreme Learning Machine (ELM) model for skin lesion classification using the ISIC dataset (HAM10000 and ISIC2018). In this study, CNN was used for feature extraction, while the ELM classifier processed these features, achieving high accuracy and specificity.

Nivedha et al. [58]focuses on melanoma detection using a novel computer-aided method combining the African Gorilla Troops Optimizer (AGTO) algorithm and Faster Region Convolutional Neural Networks (Faster R-CNN). The AGTO algorithm selects the most valuable features, and Faster R-CNN performs the classification. Applied to the ISIC-2020 skin cancer dataset, the proposed model achieves an accuracy of 98.55%, outperforming four existing approaches.

Damarla et al. [59] proposes an automated skin cancer classification system using a Deep Convolutional Neural Network (DCNN) for multiclass classification of dermoscopy images. The system employs transfer learning for feature extraction, followed by feature selection using metaheuristic algorithms like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Gorilla Troop Optimization (GTO). A two-level classification is then applied to optimize and reduce features. Tested on the HAM10000 dataset, the model achieved 93.58% accuracy, outperforming state-of-the-art techniques and showing high scalability.

Finally, Attique Khan et al. [60] developed a two-stream deep neural network framework for multiclass skin cancer classification. The first stream uses a fusion-based contrast enhancement technique and DenseNet201 to extract and optimize features, while the second stream extracts features from a fine-tuned MobileNetV2. The most discriminant features from both streams are fused using a novel multimax correlation method and classified with an extreme learning machine. The classification of skin lesions was done using a mixed datasets (HAM10000, ISBI2018, ISIC2019).

F. Studies Using CNN-Based Models

Convolutional Neural Networks (CNNs) are a popular choice for image classification tasks due to their ability to automatically learn and extract features from images. The following table summarizes studies that employed CNN-based models for skin disease classification.

Numerous studies have explored the use of Convolutional Neural Networks (CNNs) for the classification of skin lesions, showcasing the versatility and effectiveness of these models. Alassaf et al. [25] utilized a deep learning-based Auto-Encoder model on the ISIC dataset, focusing on unsupervised feature learning. The model was trained to classify skin lesions based on these features, achieving high accuracy and specificity.

Abayomi et al. [61] proposes a novel data augmentation technique based on the covariant Synthetic Minority Oversampling Technique (SMOTE) to address data scarcity and class imbalance in melanoma detection. The augmented images, generated from the PH² dataset, are used to train the SqueezeNet deep learning model. In binary classification, the model achieved 92.18% accuracy, 80.77% sensitivity, and 95.1% specificity. In multiclass classification, it improved melanoma detection with 89.2% sensitivity and 96.2% specificity for atypical nevus detection, outperforming state-of-the-art methods.

Akram et al. [62] employed a CNN to classify dermoscopic images from the PH2, ISIC-MSK, and ISIC-UDA datasets. Their methodology centered on using CNN architectures for feature extraction and classification, resulting in high accuracy and sensitivity.

Calderon et al. [63] presents a bilinear CNN approach for classifying seven skin lesion types with high accuracy and low computational cost. The framework includes data augmentation to address class imbalance, transfer learning, and fine-tuning using ResNet50 and VGG16 architectures. Tested on the HAM10000 dataset, the model achieved a 2.7% improvement over the state-of-the-art.

Khan et al. [64] enhanced a deep learning architecture with the Entropy-NDOEM algorithm for multiclass skin lesion classification. The process used include contrast enhancement, fine-tuning EfficientNetB0 and DarkNet19, feature extraction and selection using Entropy-NDOELM, feature fusion, and classification using an extreme learning machine with diverse datasets such as HAM10000, ISIC2018, and ISIC2019 datasets. The overall methodology leads to an overall improvement in model performance.

Rasel et al. [65] presents in their study, a deep learning-based system using a Convolutional Neural Network (CNN) for melanoma detection to improve early diagnosis. The research focuses on how different nonlinear activation functions affect CNN performance on limited dermoscopic image datasets. The proposed model, using the parameterized Leaky ReLU function, achieved 97.5% accuracy, 98% precision, and 98% sensitivity in classifying skin lesions into three classes. Experiments were conducted on the PH2 and ISIC datasets, demonstrating that this approach outperforms other activation functions for melanoma recognition.

Foahom et al. in their study [66] addresse the challenge of class imbalance in skin lesion (SL) detection by proposing an end-to-end decoupled training method for long-tailed skin lesion classification. The approach utilizes a novel loss function (Lf) for initial training to improve feature representation and a weighted variant of Lf to enhance robustness against class imbalance. Tested on the ISIC 2018 dataset, the model outperformed existing methods for SL detection by at least 2%, demonstrating its effectiveness in handling class imbalance.

Aldhyani et al. [26] proposes a lightweight deep learning model using dynamic-sized kernels for the accurate classification of skin lesions. The model incorporates both ReLU and leaky ReLU activation functions to enhance performance while maintaining a low number of trainable parameters. Tested on the HAM10000 dataset, the model achieved an impressive accuracy of 97.85%, outperforming several state-of-the-art models. The results demonstrate the model's efficiency in classifying various types of skin lesions.

Shen et al. [67] proposes a high-performance data augmentation strategy designed to improve skin cancer classification accuracy, particularly for use in low-resource settings. The strategy, which can be combined with any model in a plug-and-play mode, optimizes data augmentation with minimal computational cost. Using EfficientNets as a baseline, the model achieved strong performance on multiple datasets, including a BACC of 0.853 on HAM10000 and an AUC of 0.909 on ISIC 2017.

Kaur et al. [68] proposes an automated melanoma classifier using a deep convolutional neural network (DCNN) to classify malignant and benign melanoma from dermoscopic images. The DCNN is designed to efficiently extract features across multiple layers, optimizing filter selection, network depth, and hyperparameters to create a lightweight, less complex model. Tested on the ISIC 2016, 2017, and 2020 datasets, the model achieved accuracies of 81.41%, 88.23%, and 90.42%, respectively, outperforming other state-of-the-art methods and offering an efficient solution for melanoma diagnosis.

Dillshad et al. [69] focused on optimizing a CNN model for the classification of skin lesions in the HAM10000 dataset. The methodology based on a variance-controlled Marine Predator methodology optimizes feature selection and achieves high sensitivity and specificity in detecting various types of skin lesions.

Nugroho et al. [70] applied a CNN using the Inception-V3 architecture to the ISIC2019 dataset. Their approach emphasized feature extraction using the Inception-V3 model, followed by classification, which led to high accuracy in skin lesion detection. Mehmood et al. [71] proposed SBXception, a modified deep learning model based on the Xception architecture, designed for efficient skin lesion classification using the HAM10000 dataset. This methodology aim to enhance the original Xception model by making it shallower (reducing depth) and broader (increasing width), leading to fewer parameters and faster training times, achieving a high accuracy of 96.97% on the test set.

Mukadam et al. [72] used a CNN model to classify images from the HAM10000 dataset, with a focus on improving the CNN architecture to enhance accuracy and specificity in skin lesion classification.

Dahou et al. [73] proposes a robust skin cancer detection framework using a pre-trained MobileNetV3 architecture for feature extraction. The extracted features are optimized through a modified Hunger Games Search (HGS) algorithm, combining Particle Swarm Optimization and Dynamic-Opposite Learning (DOLHGS), to select the most relevant features. The model was tested on the ISIC-2016 and PH2 datasets and has improved classification accuracy across these datasets.

Finally, Supriyanto et al. [74] developed a CNN-based model for the classification of skin lesions using the HAM10000 dataset. The study focused on refining the CNN architecture to improve the sensitivity and specificity of the classification results.

G. Studies Using Machine Learning Techniques

Machine learning techniques such as XGBoost and SVM have been widely used for skin disease classification due to their robustness in handling complex datasets. The following table summarizes studies that employed these techniques, highlighting their methodologies.

Various studies have utilized machine learning techniques to improve the classification of skin lesions, often integrating multiple approaches to enhance performance. Khater et al. [75] employed XGBoost on selected features derived from dermoscopic images in the PH2 dataset. Their methodology focused on enhancing feature extraction through preprocessing and used explainable AI techniques, as SHAP to ensure the interpretability of the results, leading to improved classification accuracy.

Ahmed et al. [76] integrated Convolutional Neural Networks (CNNs) with Support Vector Machine (SVM) and Artificial Neural Network (ANN) models to analyze dermoscopic images from the HAM10000 and PH2 datasets. Their approach utilized MobileNet and ResNet101 architectures for feature extraction, followed by classification with SVM and ANN, achieving high accuracy across multiple metrics.

Ilkin et al. [77] proposed a combination of SVM with feature extraction from mixed datasets, including PH2 and ISIC. Their study focused on optimizing the feature set prior to SVM classification, aiming to improve the model's ability to differentiate between various skin lesion classes.

Finally, Pitchiah et al. [78] introduced a hybrid model combining K-Nearest Neighbors (KNN) with Random Forest and SVM. This model was validated on the PH2 dataset and aimed to balance sensitivity and precision by using ensemble methods to enhance classification performance.

These methodologies explanations serve as a focal point for the discussion. The detailed breakdown presented here assists researchers and practitioners in the identification of prevailing trends and gaps in the literature, and understanding which methodologies are most promising for clinical application and where further innovation may be required.

TABLE II. Summary of Techniques and Performance in Systematic Review Studies

Ref.	Year	Technique	Data Type	Classes Number	Performance
[25]	2023	Deep learning-based Auto-Encoder	Image dataset (ISIC)	7	Accuracy: 96.83%, Sensitivity: 96.57%, Specificity: 97.83%
[76]	2023	CNN combined with SVM/ANN	HAM10000	7	HAM10000: Accuracy: 98.4%, Sensitivity: 94.46%, Specificity: 99.43%, AUC: 97.53%
			PH2	3	PH2 : Accuracy: 100%, Sensitivity: 100%, Specificity: 100%, AUC: 100%
[24]	2023	CNN integrated with Random Forest	HAM10000	7	Accuracy: 94.96%, Sensitivity: 93.74%, Specificity: 93.16%, F1-score: 93.24%
[75]	2023	XGBoost applied to selected features	PH2	3	Accuracy: 94%, AUC: 99.47%
[64]	2023	CNN enhanced with ELM (Extreme Learning Machine)	HAM10000	7	Accuracy : 95.7%
			ISIC2018	7	Accuracy : 96.3%
			ISIC2019	8	Accuracy : 94.8%
[47]	2023	Vision Transformer (ViT) model	ISIC2018	8	Accuracy: 93.2%, Specificity: 92.2%, AUC: 97.7%
			ISIC2017	2	Accuracy: 89.5%, Specificity: 94.7%, AUC: 96.2%
			PH2	2	Accuracy: 91.4%, Specificity: 88.5%, AUC: 96.3
[62]	2023	Deep models with entropy-controlled optimization for feature selection	PH2	3	Accuracy: 98.89%, Specificity: 98.9%, Sensivity: 98%
			ISIC-MSK	Various	Accuracy: 99.01%, Specificity: 99.4%, Sensivity: 98.5%
			ISIC-UDA	3	Accuracy: 99.09%, Specificity: 99.4%, Sensivity: 98.6%
[43]	2023	Transformer-based model	Image dataset (Derm7pt)	2	Accuracy: 80.03%
[51]	2023	Hybrid CNN and Transformer model	Image dataset (ISIC2018)	7	F1 score: 87.37%, Sensitivity: 88.13%, Specificity: 88.29%
[29]	2023	DSCC_Net model with SMOTE Tomek	ISIC2020, Derm-IS, HAM10000	Various	Accuracy: 94.17%, Sensitivity: 94.28%, Specificity: 93.76%, F1-score: 93.93%
[58]	2023	Region-based CNN (RCNN)	Image dataset (ISIC2020)	2	Accuracy: 98.55%, Sensitivity: 96.92%, Specificity: 98.11%, Precision: 98.34%
[48]	2023	Transformer-based model	Image dataset (Personalized ISIC)	9	Accuracy: 95.6%, Sensitivity: 96.7%, Specificity: 95%
[69]	2023	Optimization-aided deep learning with MobileNetV2, NasNet, and Marine Predator	Image dataset (HAM10000)	7	Accuracy: 94.4%, Sensitivity: 94.4%, Specificity: 94.4%
[70]	2023	Convolutional Neural Network (CNN) using Inception-V3	ISIC dataset (ISIC2019)	7	Accuracy: 96.4%, AUC: 0.98
[28]	2023	Hybrid CNN with Random Forest	HAM10000	7	AUC: 87.6%
[72]	2023	Custom CNN with ESRGAN preprocessing	HAM10000	7	Accuracy: 98.9%
[73]	2023	CNN with Hunger Games Search	PH2	3	Accuracy: 96.43%
			ISIC2016	2	Accuracy: 88.19%
[44]	2023	Multi-modal Vision Transformer (ViT) model	ISIC dataset (ISIC2018)	7	Accuracy: 93.81%, Sensitivity: 90.14%, Specificity: 98.36%, F1-score: 90.13 %

Ref.	Year	Technique	Data Type	Classes Number	Performance
[49]	2023	Transformer-based model	HAM10000	7	Accuracy: 94.1%
[56]	2023	Hybrid CNN with SVM	ISIC2019	8	Accuracy: 96%
			ISIC2020	2	Accuracy: 92%
[71]	2023	SbXception: an enhanced Xception architecture	HAM10000	7	Accuracy: 96.97%, Sensitivity: 95.43%, Specificity: 85.34%
[74]	2023	Two-stage image augmentation with GAN and CNN models	HAM10000	7	Accuracy: 96.9%, Sensitivity: 96.87%, Specificity: 97.07%, F1-score: 96.97%
[52]	2022	Hybrid CNN with SVM	ISIC dataset (ISIC2019)	8	Accuracy: 98.76%, Sensitivity: 98.4%, Specificity: 99.81%
[59]	2022	A deep learning system using transfer learning and meta-heuristic optimization	HAM10000	7	Accuracy: 93.58%
[67]	2022	Convolutional Neural Network (CNN) using EfficientNets	HAM10000	7	Accuracy: 95.8%, Sensitivity: 85.3%, Specificity: 97.9%, AUC: 0.975
[45]	2022	Convolutional Neural Network with Soft-attention	HAM10000	7	Accuracy: 90%, Sensitivity: 81%, FA-score: 81%, AUC: 0.99
[63]	2021	Convolutional Neural Network (CNN) using bilinear approach	HAM10000	7	Accuracy: 93.21%, Sensitivity: 93%, Specificity: 92.92% , AUC: 0.98
[61]	2021	Convolutional Neural Network using SMOTE Oversampling Technique	Image dataset (PH2)	3	Accuracy: 92.18%, Sensitivity: 80.77%, Specificity: 95.1%, F1-score: 80.84%
[55]	2021	Graph nodes in CNN architecture	ISIC datasets	7	Accuracy: 97.4%, Sensitivity: 100%, Specificity: 95.16%, AUC: 99.91%
[68]	2022	Deep Convolutional Neural Network (DCNN)	ISIC2016	2	Accuracy: 81.41%
			ISIC2017	2	Accuracy: 88.23%
			ISIC2020	2	Accuracy: 90.42%
[65]	2022	Convolutional Neural Network (CNN) Leaky ReLU function	Mixed dataset (PH2, ISIC)	2,7	Accuracy: 97.5%, Precision: 98.0%, Sensitivity: 98.0%
[77]	2021	Support Vector Machine (SVM)	PH2	3	Accuracy: 97.5%, Sensitivity: 93.75%, Specificity: 100%, AUC: 97%
			ISIC	3	Accuracy: 97.56%, Sensitivity: 97.94%, Specificity: 97.07%, AUC: 98%
[57]	2022	Hybrid CNN-ELM (Extreme Learning Machine)	HAM1000	7	Accuracy: 93.4%, Precision: 93.10%
			ISIC2018	7	Accuracy: 94.36%, Precision: 94.08%
[46]	2022	Hybrid CNN with DenseNet-169	Mixed dataset (PAD-UFES-20, ISIC 2019)	7	Accuracy: 81.4%
[66]	2022	Convolutional Neural Network (CNN)	ISIC dataset (ISIC2018)	7	B.Accuracy: 88%
[54]	2021	CNN combined with Meta-Heuristic methods	ISIC dataset (ISIC2008)	Not defined	Accuracy: 96%, Sensitivity: 96%, Specificity: 95%
[26]	2022	Lightweight CNN with dynamic-sized kernels and ReLU/leaky ReLU activations	HAM10000	7	Accuracy: 97.8%, Sensitivity: 98%, Specificity: 98.1%, F1-score: 98%
[39]	2021	GAN combined with CNN	ISIC dataset (ISIC2018)	7	Accuracy: 70.1%
[78]	2022	K-Nearest Neighbors (KNN) combined with Random Forest (RF) and SVM	Image dataset (PH2)	3	Accuracy: 94.81%

Ref.	Year	Technique	Data Type	Classes Number	Performance
[53]	2021	CNN with PMEC feature fusion and EKWO optimization	HAM10000	7	Accuracy: 95.8%
			ISIC2018	7	Accuracy: 97.1%
			ISBI2019	8	Accuracy: 85.35%
[79]	2022	Convolutional Neural Network (CNN)	HAM10000	7	Accuracy: 91%, F1-score: 88.1%, ROC-AUC: 95%
[50]	2022	Transformer-based model	HAM10000	7	Accuracy: 96.14%, Sensitivity: 96.5%, Specificity: 96%, F1-score:97%
[80]	2021	CNN combined with Machine Learning (ML) techniques	HAM10000	7	Accuracy: 91.7%
[30]	2022	Explainable AI (XAI) approach using CNN	ISIC dataset (ISIC2019)	8	Accuracy: 94.47%, Sensitivity: 94.01%, Specificity: 93.57%, F1-score: 94.45%
[81]	2021	Neural Network (NN)	ISIC dataset (ISIC2017)	2	Precision: 94%, Sensitivity: 93%, Specificity: 91%
[82]	2021	Feedforward Neural Network combined with Artificial Neural Network	ISIC2018	7	Accuracy: 90%, Sensitivity: 89.37%, Specificity: 97.84%
			PH2	3	Accuracy: 95.8%, Sensitivity: 95.64%, Specificity: 98.21%
[83]	2023	CLCM-net model with layer-wise weight constraints	ISIC2018	7	Accuracy: 94.42%
			ISIC2019	8	Accuracy: 95.8%
			Combined dataset	-	Accuracy: 93%
[84]	2022	Capsule Network	HAM10000	7	Accuracy: 96.49%
[85]	2021	Neural Network (NN)	7-Point	2	Accuracy: 95.42%, Sensitivity: 98.01%, Specificity: 94.4%
			Med-Node	2	Accuracy: 94.71%, Sensitivity: 96.42%, Specificity: 87.5%
			PH2	3	Accuracy: 94.88%, Sensitivity: 100%, Specificity: 85.62%
[60]	2022	Two-stream neural network with feature fusion and ELM	HAM10000	7	Accuracy: 96.5%
			ISBI2018	7	Accuracy: 98%
			ISIC2019	8	Accuracy: 89%
[40]	2022	Ensemble Learning approach	HAM10000	7	Validation Accuracy: 86.71%
[86]	2022	StochasticProgressive Instance Learning	ISIC2017	2	Accuracy: 88%, AUC: 98.3
			ISIC2018	7	Accuracy: 89.4%, AUC: 92.9
[42]	2022	Graph nodes-based approach	7-point	2	AUC: 83.6%
[87]	2021	CNN combined with ANN	ISIC dataset	5	Balanced Accuracy: 92.34%, Sensitivity: 87.10%, Specificity: 94.19% , AUC-ROC: 97.10%
[41]	2021	Ensemble Learning approach	HAM10000	7	Accuracy: 83.5%, Sensitivity: 65.6%, Specificity: 95.4%
[27]	2024	Generative Adversarial Network (GAN) with CNN	HAM10000	7	Accuracy: 98.23%, Sensitivity: 88.85%, Specificity: 98.34%, F1-score:89.48

The varying performance of AI models across different datasets can be attributed to the unique characteristics of each dataset, such as size, number of classes, image diversity, and data quality. Larger datasets like ISIC2018 and ISIC2019, which contain thousands of high-resolution dermoscopic images, generally allow AI models to perform well due to the availability of diverse training data. However, the number of classes also plays a crucial role. For example, studies using ISIC2019, which contains 8 classes, reported lower accuracy (e.g. [64] achieved 94.8% accuracy) compared to ISIC2018 (96.3% accuracy), which has fewer classes and is less complex. The increased difficulty in distinguishing between more skin disease types likely accounts for this discrepancy. Conversely, smaller datasets like PH2, with only 200 images and 3 classes, often result in higher accuracy within the dataset, as seen in [62] with an accuracy of 98.89%. However, such models may struggle with generalizability when applied to larger, more complex datasets. Similarly, HAM10000 provides a diverse dataset with 7 classes, leading to good overall performance in most models (e.g., 95.7% accuracy in [64]), although class imbalance in some categories can affect results. Models tested on clinical image datasets like Derm7pt, which include more background noise and variability in image conditions, often report lower performance compared to dermoscopic datasets, as these models must handle more variability in input data. This analysis suggests that certain algorithms, particularly CNNs and hybrid models, are better suited to high-quality, diverse dermoscopic datasets, while challenges like class imbalance, dataset size, and image quality play a significant role in the comparative results across different datasets.

The systematic review of recent studies on AI models for skin disease classification, as detailed in Table II and categorized in the paragraphs above, reveals several critical trends, patterns, and challenges that underscore the current state of research in this domain. To provide a nuanced understanding of these approaches, Fig. 8 presents the distribution of the AI techniques utilized across the studies. This figure highlights the prevalence of different methodologies, offering insights into the current state of AI-driven skin disease classification.

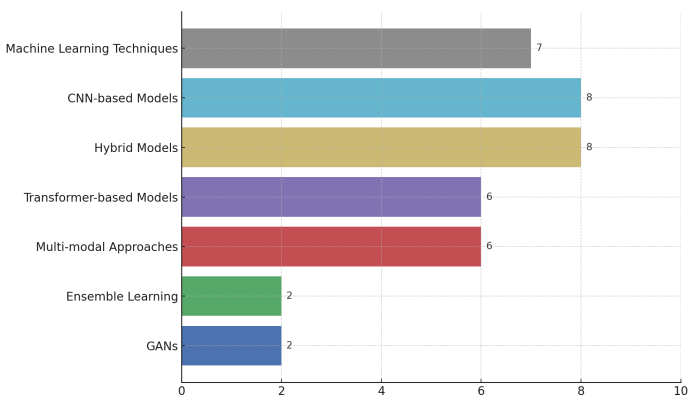

Fig. 8. Trends in AI techniques for skin disease classification (2021-2023).

A significant trend observed across the reviewed literature is the dominance of Convolutional Neural Networks (CNNs) and their variations, including hybrid models combining CNNs with other machine learning techniques such as Support Vector Machines (SVMs), Extreme Learning Machines (ELMs), and Random Forests. This reflects the strong performance of CNNs in image recognition tasks, which are central to dermatological diagnosis. To further analyze the methodologies within CNN-based models and their variations, Fig. 9 presents the usage frequency of various pre-trained CNN models across the studies. This figure provides a clear overview of which pre-trained models are most commonly employed in skin disease classification.

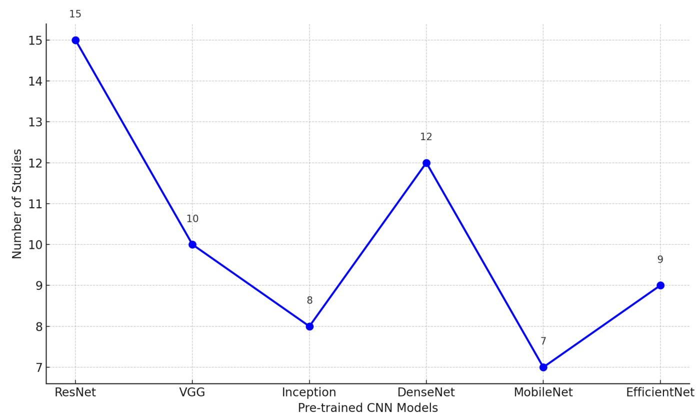

Fig. 9. Usage frequency of Pre-trained CNN models in skin disease classification studies.

More recently, the emergence of Transformer-based models and Vision Transformers (ViTs) suggests a growing interest in leveraging these advanced architectures, which have demonstrated exceptional performance in natural language processing and are now being adapted for medical imaging.

The reviewed studies span from 2021 to 2023, indicating a rapid evolution of techniques over a relatively short period. Early studies primarily focused on straightforward CNN architectures, while later studies have increasingly explored more complex hybrid models and the integration of transformers. This evolution suggests a shift towards more sophisticated, multi-faceted approaches aimed at improving model accuracy and generalization across diverse datasets. Fig. 10 illustrates the increasing trend in the use of Hybrid Models and Transformers over time.

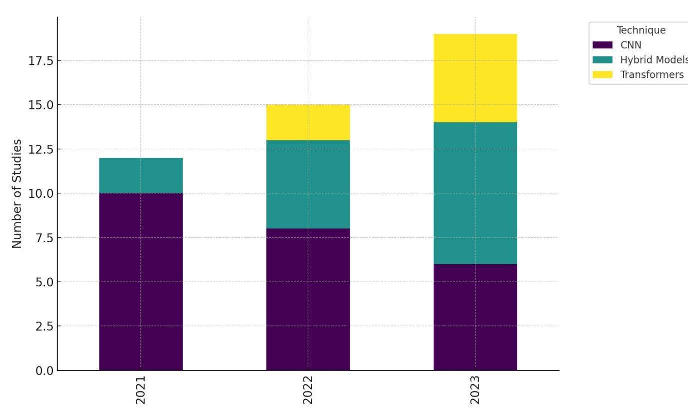

Fig. 10. Trends in AI techniques for skin disease classification (2021-2023).

Additionally, there is a noticeable increase in the application of Auto-Encoders, Generative Adversarial Networks (GANs), and ensemble methods, reflecting the ongoing effort to enhance model robustness and address challenges such as data scarcity and imbalance. The development of GANs combined with CNNs, as seen in recent studies, highlights a trend towards generating synthetic data to improve training outcomes, which is particularly valuable in medical fields where annotated data can be limited.

The predominant use of datasets such as HAM10000, ISIC (International Skin Imaging Collaboration), and PH2 across the studies indicates a reliance on these well-established, publicly available image datasets. The consistent use of these datasets underscores their role as benchmarks in the field. However, it also highlights a potential limitation in terms of dataset diversity, as most studies focus on the same data sources, which may not fully represent the variety of skin conditions encountered in clinical practice.

In terms of class distribution, most studies focus on classifying a limited number of conditions, with a particular emphasis on melanoma and non-melanoma skin cancers. This trend reflects the clinical importance of accurately diagnosing these conditions but also points to a gap in research focused on rarer or less visually distinct skin diseases, which are underrepresented in current models.

The reported performance metrics, including accuracy, sensitivity, specificity, and precision, show high variability across studies, with accuracy ranging from 88.19% to 98.7%. This variability can be attributed to differences in model complexity, data preprocessing methods, and the inherent difficulty of the classification tasks. Notably, hybrid models and those incorporating transformers generally report higher performance, suggesting that these more complex models may offer advantages in handling the nuances of skin disease classification. Fig. 11 provides a comparative analysis of the performance distribution across different model types.

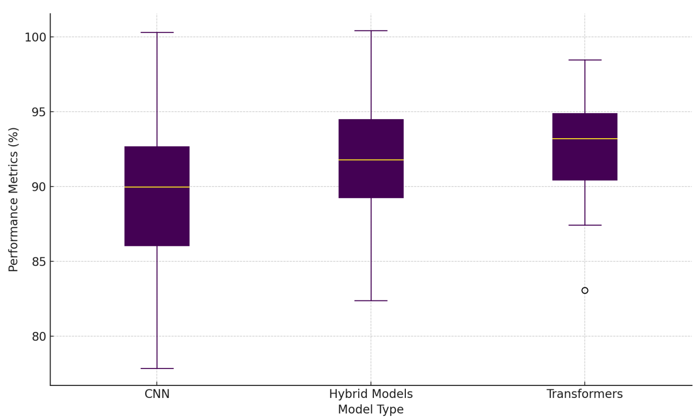

Fig. 11. Performance distribution across different models.

However, it is important to note that high performance on well-curated datasets does not necessarily translate to clinical effectiveness. The consistent reporting of high specificity across models is encouraging, as it suggests a strong ability to correctly identify negative cases, which is critical in avoiding unnecessary interventions. Nonetheless, the lower sensitivity scores reported in some studies, particularly those involving simpler CNN architectures, indicate a potential risk of missed diagnoses, which could have serious clinical implications.

Despite the progress made, several challenges remain. The reliance on a few datasets raises concerns about the generalizability of these models to broader, more diverse patient populations. Additionally, while the integration of multimodal data (e.g., combining images with patient history) is increasingly being recognized as essential for improving diagnostic accuracy, few studies have fully implemented this approach.

Moreover, the computational complexity of advanced models such as transformers and GANs may limit their deployment in resource-constrained settings, highlighting the need for research focused on optimizing these models for real-world clinical environments. Fig. 12 compares the performance metrics of hybrid models versus pure deep learning models, emphasizing the potential advantages of hybrid approaches.

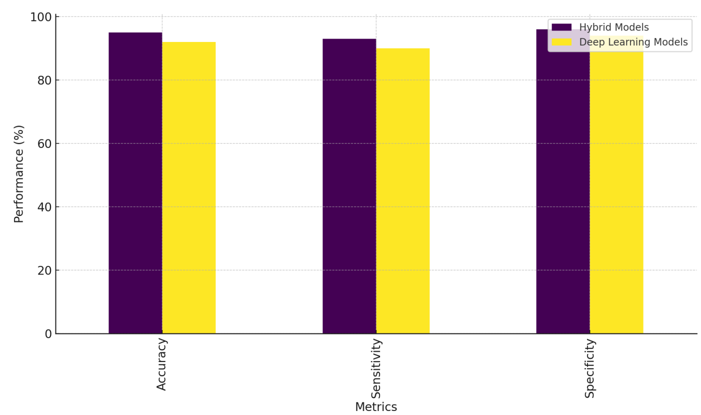

Fig. 12. Performance comparison: Hybrid models vs pure deep learning.

The general metrics commonly used to evaluate the performance of AI models in skin disease classification include Accuracy, Sensitivity, Specificity, Precision, F1-Score, and Area Under the Curve (AUC). These metrics are crucial in assessing the effectiveness of the models in correctly identifying and classifying skin conditions.

1. Accuracy: The proportion of correctly classified instances (both true positives and true negatives) among the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Sensitivity (Recall or True Positive Rate): The proportion of actual positives that are correctly identified by the model.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

3. Specificity (True Negative Rate): The proportion of actual negatives that are correctly identified by the model.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

4. Precision (Positive Predictive Value): The proportion of positive predictions that are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

5. F1-Score: The harmonic mean of Precision and Sensitivity, providing a balance between the two.

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

6. Area Under the Curve (AUC): The area under the Receiver Operating Characteristic (ROC) curve, representing the model's ability to discriminate between classes.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx$$

VII. Conclusion

This systematic review provides a focused analysis of AI methodologies in skin disease classification, highlighting the growing adoption of advanced techniques like GANs, Transformer models, and multi-modal approaches. While CNNs remain a dominant tool, their performance is often enhanced by hybrid and ensemble learning methods, demonstrating a trend towards more complex model architectures.

However, this review also identifies several key challenges that need to be addressed. The lack of standardization across studies and the limited application of multi-modal approaches restrict the generalizability of current models. Additionally, the datasets used in most studies lack diversity in terms of patient demographics, including under-representation of different skin tones and rare skin diseases, which limits the models' applicability across broader populations. Moreover, the explainability of AI models remains a critical barrier to their integration into clinical practice. Ensuring that AI-driven diagnostic tools provide transparent and interpretable outputs for clinicians is crucial for their adoption in real-world settings.

Future research should prioritize the development of standardized protocols and benchmarking methods to enable meaningful comparisons between different AI models. The creation of more diverse and representative datasets is essential to improve the generalization of AI models and their applicability in real-world clinical environments. Furthermore, addressing the ethical implications of AI use in dermatology is vital, particularly in relation to bias mitigation, ensuring that AI technologies perform equitably across all patient groups. There is also a growing need to investigate advanced techniques such as federated learning, which could enhance collaboration between institutions while preserving patient privacy, thereby improving model generalizability without compromising data security.

Finally, real-world validation through clinical trials and large-scale implementation studies is necessary to evaluate the practical utility and reliability of AI-driven skin disease classification tools. Collaboration between AI researchers and clinicians will be crucial in translating these models from research into clinical practice.

References

[1] H. Kutzner, T. B. Jutzi, D. Krahl, E. I. Krieghoff-Henning, M. V. Heppt, A. Hekler, M. Schmitt, R. C. Maron, S. Fröhling, C. von Kalle *et al.*, "Overdiagnosis of melanoma–causes, consequences and solutions," *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, vol. 18, no. 11, pp. 1236–1243, 2020.

[2] A. A. Marghoob, L. Changchien, J. DeFazio, W. C. Dessio, J. Malvehy, I. Zalaudek, A. C. Halpern, and A. Scope, "The most common challenges in melanoma diagnosis and how to avoid them," *Australasian Journal of Dermatology*, vol. 50, no. 1, pp. 1–13, 2009.

[3] C. Karimkhani, A. C. Green, T. Nijsten, M. A. Weinstock, R. P. Dellavalle, M. Naghavi, and C. Fitzmaurice, "The global burden of melanoma: results from the global burden of disease study 2015," *British Journal of Dermatology*, vol. 177, no. 1, pp. 134–140, 2017.

[4] C. E. Griffiths, A. W. Armstrong, J. E. Gudjonsson, and J. N. Barker, "Psoriasis," *The Lancet*, vol. 397, no. 10281, pp. 1301–1315, 2021.

[5] W. Peng and N. Novak, "Pathogenesis of atopic dermatitis," *Clinical & Experimental Allergy*, vol. 45, no. 3, pp. 566–574, 2015.

[6] L. Thomas and S. Puig, "Dermoscopy, digital dermoscopy and other diagnostic tools in the early detection of melanoma and follow-up of high-risk skin cancer patients." *Acta dermato-venereologica*, vol. 97, 2017.

[7] R. A. Shellenberger, F. Fayyaz, Z. Sako, M. Schaeffer, K. Tawagi, C. Scheidel, and M. Nabhan, "Impact of biopsy technique on clinically important outcomes for cutaneous melanoma: a systematic review and meta-analysis," *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, vol. 4, no. 4, pp. 373–383, 2020.

[8] P. Carli, D. Massi, V. De Giorgi, and B. Giannotti, "Reliability and inter-observer agreement of dermoscopic diagnosis of melanocytic lesions," *Journal of the American Academy of Dermatology*, vol. 48, no. 6, pp. 878–883, 2003.

[9] D. A. Gutman, N. C. Codella, M. E. Celebi, B. Helba, M. A. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016 hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.

[10] N. Noroozi and A. Zakerolhosseini, "Differential diagnosis of squamous cell carcinoma in situ using skin histopathological images," *Computers in biology and medicine*, vol. 70, pp. 23–39, 2016.

[11] H. D. Heibel, L. Hooey, and C. J. Cockerell, "A review of noninvasive techniques for skin cancer detection in dermatology," *American journal of clinical dermatology*, vol. 21, no. 4, pp. 513–524, 2020.

[12] S. L. Schneider, I. Kohli, I. H. Hamzavi, M. L. Council, A. M. Rossi, and D. M. Ozog, "Emerging imaging technologies in dermatology: Part ii: Applications and limitations," *Journal of the American Academy of Dermatology*, vol. 80, no. 4, pp. 1121–1131, 2019.

[13] S. R. Jartarkar, A. Patil, U. Wollina, M. H. Gold, H. Stege, S. Grabbe, and M. Goldust, "New diagnostic and imaging technologies in dermatology," *Journal of cosmetic dermatology*, vol. 20, no. 12, pp. 3782–3787, 2021.

[14] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, 2017.

[15] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Computers in biology and medicine*, vol. 127, p. 104065, 2020.

[16] M. A. Thaajwer and U. P. Ishanka, "Melanoma skin cancer detection using image processing and machine learning techniques," in *2020 2nd international conference on advancements in computing (ICAC)*, vol. 1. IEEE, 2020, pp. 363–368.

[17] T. Shanthi, R. Sabeenian, and R. Anand, "Automatic diagnosis of skin diseases using convolution neural network," *Microprocessors and Microsystems*, vol. 76, p. 103074, 2020.

[18] T. J. Brinker, A. Hekler, A. H. Enk, C. Berking, S. Haferkamp, A. Hauschild, K. Ghoreschi, M. Weichenthal, J. Klode, D. Schadendorf *et al.*, "Skin cancer classification using convolutional neural networks: systematic review," *Journal of the American Academy of Dermatology*, vol. 79, no. 6, pp. 1233–1242, 2018.

[19] J. Zhang, F. Zhong, K. He, M. Ji, S. Li, and C. Li, "Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: A review," *Diagnostics*, vol. 13, no. 23, p. 3506, 2023.

[20] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[21] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, J. Malvehy, S. Puig, C. Rosendahl, and H. Kittler, "Human-computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, 2020.

[22] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin cancer classification with deep learning: a systematic review," *Frontiers in Oncology*, vol. 12, p. 893972, 2022.

[23] E. R. Gordon, M. H. Trager, D. Kontos, C. Weng, L. J. Geskin, L. S. Dugdale, and F. H. Samie, "Ethical considerations for artificial intelligence in dermatology: a scoping review," *British Journal of Dermatology*, p. ljae040, 2024.

[24] X. Ma, J. Shan, F. Ning, W. Li, and H. Li, "Effnet: A skin cancer classification model based on feature fusion and random forests," *PLOS ONE*, vol. 18, no. 10, pp. 1–17, 10 2023.

[25] A. Alassaf, "Deep learning-based stacked auto-encoder with dynamic differential annealed optimization for skin lesion diagnosis," *Computer Systems Science and Engineering*, vol. 47, no. 3, pp. 2773–2789, 2023.

[26] T. H. Aldhyani, A. Verma, M. H. Al-Adhaileh, and D. Koundal, "Multi-class skin lesion classification using a lightweight dynamic kernel deep-learning-based convolutional neural network," *Diagnostics*, vol. 12, no. 9, p. 2048, 2022.

[27] Q. Su, H. N. A. Hamed, M. A. Isa, X. Hao, and X. Dai, "A gan-based data augmentation method for imbalanced multi-class skin lesion classification," *IEEE Access*, 2024.

[28] A. Roge, P. Ting, A. Chern, and W. Ting, "Deep ensemble learning using a demographic machine learning risk stratifier for binary classification of skin lesions using dermatoscopic images," *Journal of Medical Artificial Intelligence*, vol. 6, no. 0, 2023.

[29] M. Tahir, A. Naeem, H. Malik, J. Tanveer, R. A. Naqvi, and S.-W. Lee, "Dscc_net: Multiclassification deep learning models for diagnosing of skin cancer using dermoscopic images," *Cancers*, vol. 15, no. 7, 2023.

[30] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and S. Abawi, "A deep learning approach based on explainable ai for skin lesion classification," *IEEE Access*, 2022.

[31] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the isic image datasets: Usage, benchmarks and recommendations," *Medical image analysis*, vol. 75, p. 102305, 2022.

[32] 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, p. 180161, 2018.

[33] T. Mendonça, M. Celebi, T. Mendonca, and J. Marques, "Ph2: A public database for the analysis of dermoscopic images," *Dermoscopy image analysis*, vol. 2, 2015.

[34] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[35] A. G. Pacheco, G. R. Lima, A. S. Salomao, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro *et al.*, "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in brief*, vol. 32, p. 106221, 2020.

[36] B. Zhang, X. Zhou, Y. Luo, H. Zhang, H. Yang, J. Ma, and L. Ma, "Opportunities and challenges: Classification of skin disease based on deep learning," *Chinese Journal of Mechanical Engineering*, vol. 34, pp. 1–14, 2021.

[37] A. Cascarano, J. Mur-Petit, J. Hernandez-Gonzalez, M. Camacho, N. de Toro Eadie, P. Gkontra, M. Chadeau-Hyam, J. Vitria, and K. Lekadir, "Machine and deep learning for longitudinal biomedical data: a review of methods and applications," *Artificial Intelligence Review*, vol. 56, no. Suppl 2, pp. 1711–1771, 2023.

[38] K. Frasier, V. Li, M. Coleman, E. Rodriguez, T. Karatas *et al.*, "Digital intelligence for predicting skin disease progression and treatment outcomes," *Ameri J Clin Med Re: AJCMR-137*, 2024.

[39] I. S. A. Abdelhalim, M. F. Mohamed, and Y. B. Mahdy, "Data augmentation for skin lesion using self-attention based progressive generative adversarial network," *Expert Systems with Applications*, vol. 165, p. 113922, 2021.

[40] D. Popescu, M. El-Khatib, and L. Ichim, "Skin lesion classification using collective intelligence and deep learning," *Sensors*, 2022.

[41] K. Thurnhofer-Hemsi, E. López-Rubio, E. Dominguez, and D. A. Elizondo, "Skin lesion classification by ensembles of deep convolutional networks and regularly spaced shifting," *IEEE Access*, vol. 9, pp. 112 193–112 205, 2021.

[42] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Graph-based intercategory and intermodality network for multilabel classification and melanoma diagnosis of skin lesions in dermoscopy and clinical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3266–3277, 2022.

[43] Y. Zhang, F. Xie, and J. Chen, "Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis," *Computers in Biology and Medicine*, vol. 157, p. 106712, 2023.

[44] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *The Visual Computer*, vol. 39, no. 7, pp. 2781–2793, 2023.

[45] V. D. Nguyen, N. D. Bui, and H. K. Do, "Skin lesion classification on imbalanced data using deep learning with soft attention," *Sensors*, vol. 22, no. 19, p. 7530, 2022.

[46] W. Yin, J. Huang, J. Chen, and Y. Ji, "A study on skin tumor classification based on dense convolutional networks with fused metadata," *Frontiers in Oncology*, vol. 12, p. 989894, 2022.

[47] Y. Ding, Z. Yi, M. Li, J. long, S. Lei, Y. Guo, P. Fan, C. Zuo, and Y. Wang, "Hi-mvit: A lightweight model for explainable skin disease classification based on modified mobilevit," *Digital Health*, vol. 9, p. 20552076231207197, 2023.

[48] Q. Abbas, Y. Daadaa, U. Rashid, and M. E. Ibrahim, "Assist-dermo: A lightweight separable vision transformer model for multiclass skin lesion classification," *Diagnostics*, vol. 13, no. 15, p. 2531, 2023.

[49] G. Yang, S. Luo, and P. Greer, "A novel vision transformer model for skin cancer detection," *Neural Processing Letters*, 2023.

[50] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, and U. Tariq, "An effective skin cancer classification mechanism using hybrid cnn and svm models," *Sensors*, 2022.

[51] Y. Nie, P. Sommella, M. Carratù, M. O'Nils, and J. Lundgren, "A deep cnn transformer hybrid model for skin lesion classification of dermoscopic images using focal loss," *Diagnostics*, vol. 13, no. 1, 2023.

[52] Z. Li, Z. Chen, X. Che, Y. Wu, D. Huang, H. Ma, and Y. Dong, "A classification method for multi-class skin damage images combining quantum computing and inception-resnet-v1," *Frontiers in Physics*, vol. 10, p. 1046314, 2022.

[53] M. A. Khan, T. Akram, M. Sharif, S. Kadry, and Y. Nam, "Computer decision support system for skin cancer localization and classification." *Computers, Materials & Continua*, vol. 68, no. 1, 2021.

[54] B. Zhou and B. Arandian, "An improved cnn architecture to diagnose skin cancer in dermoscopic images based on wildebeest herd optimization algorithm," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 7567870, 2021.

[55] M. H. Annaby, A. M. Elwer, M. A. Rushdi, and M. E. Rasmy, "Melanoma detection using spatial and spectral analysis on superpixel graphs," *Journal of digital imaging*, vol. 34, no. 1, pp. 162–181, 2021.

[56] M. Saeed, A. Naseer, H. Masood, S. U. Rehman, and V. Gruhn, "The power of generative ai to augment for enhanced skin cancer classification: A deep learning approach," *IEEE Access*, vol. 11, pp. 130 330–130 344, 2023.

[57] F. Afza, M. Sharif, M. A. Khan, U. Tariq, H.-S. Yong, and J. Cha, "Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine," *Sensors*, vol. 22, no. 3, p. 799, 2022.

[58] S. Nivedha and S. Shankar, "Melanoma diagnosis using enhanced faster region convolutional neural networks optimized by artificial gorilla troops algorithm," *Information Technology and Control*, vol. 52, no. 4, pp. 819–832, 2023.

[59] A. Damarla and D. Sumathi, "An approach for optimization of features using gorilla troop optimizer for classification of melanoma," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022.

[60] M. Attique Khan, M. Sharif, T. Akram, S. Kadry, and C.-H. Hsu, "A two-stream deep neural network-based intelligent system for complex skin cancer types classification," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 10 621–10 649, 2022.

[61] O. O. Abayomi-Alli, R. Damasevicius, S. Misra, R. Maskeliunas, and A. Abayomi-Alli, "Malignant skin melanoma detection using image augmentation by oversamplingin nonlinear lower-dimensional embedding manifold," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 8, pp. 2600–2614, 2021.

[62] T. Akram, R. Junejo, A. Alsuhaibani, M. Rafiullah, A. Akram, and N. A. Almujally, "Precision in dermatology: Developing an optimal feature selection framework for skin lesion classification," *Diagnostics*, vol. 13, no. 17, 2023.

[63] C. Calderón, K. Sanchez, S. Castillo, and H. Arguello, "Bilsk: A bilinear convolutional neural network approach for skin lesion classification," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100036, 2021.

[64] M. A. Khan, T. Akram, Y.-D. Zhang, M. Alhaisoni, A. Al Hejaili, K. A. Shaban, U. Tariq, and M. H. Zayyan, "Skinnet-endo: Multiclass skin lesion recognition using deep neural network and entropy-normal distribution optimization algorithm with elm," *International Journal of Imaging Systems and Technology*, vol. 33, no. 4, pp. 1275–1292, 2023.

[65] M. Rasel, U. H. Obaidellah, and S. A. Kareem, "convolutional neural network-based skin lesion classification with variable nonlinear activation functions," *IEEE Access*, vol. 10, pp. 83 398–83 414, 2022.

[66] A. C. Foahom Gouabou, R. Iguernaissi, J.-L. Damoiseaux, A. Moudafi, and D. Merad, "End-to-end decoupled training: A robust deep learning method for long-tailed classification of dermoscopic images for skin lesion classification," *Electronics*, vol. 11, no. 20, p. 3275, 2022.

[67] S. Shen, M. Xu, F. Zhang, P. Shao, H. Liu, Y. Xu, and L. Ma, "A low-cost high-performance data augmentation approach for skin lesion classification using convolutional neural networks," *Journal of Healthcare Engineering*, 2022.

[68] R. Kaur, H. GholamHosseini, R. Sinha, and M. Lindén, "Melanoma classification using a novel deep convolutional neural network with dermoscopic images," *Sensors*, vol. 22, no. 3, p. 1134, 2022.

[69] V. Dillshad, M. A. Khan, M. Nazir, O. Saidani, N. Alturki, and S. Kadry, "D2lfs2net: Multi-class skin lesion diagnosis using deep learning and variance-controlled marine predator optimisation: An application for precision medicine," *CAAI Transactions on Intelligence Technology*, 2023.

[70] E. S. Nugroho, I. Ardiyanto, and H. A. Nugroho, "Boosting the performance of pretrained cnn architecture on dermoscopic pigmented skin lesion classification," *Skin Research and Technology*, vol. 29, no. 11, p. e13505, 2023.

[71] A. Mehmood, Y. Gulzar, Q. M. Ilyas, A. Jabbari, M. Ahmad, and S. Iqbal, "Sbxception: a shallower and broader xception architecture for efficient classification of skin lesions," *Cancers*, vol. 15, no. 14, p. 3604, 2023.

[72] S. B. Mukadam and H. Y. Patil, "Skin cancer classification framework using enhanced super resolution generative adversarial network and custom convolutional neural network," *Applied Sciences*, vol. 13, no. 2, 2023.

[73] A. Dahou, A. O. Aseeri, A. Mabrouk, R. A. Ibrahim, M. A. Al-Betar, and M. A. Elaziz, "Optimal skin cancer detection model using transfer learning and dynamic-opposite hunger games search," *Diagnostics*, vol. 13, no. 9, 2023.

[74] C. Supriyanto, A. Salam, J. Zeniarja, and A. Wijaya, "Two-stage input-space image augmentation and interpretable technique for accurate and explainable skin cancer diagnosis," *Computation*, vol. 11, no. 12, p. 246, 2023.

[75] T. Khater, S. Ansari, S. Mahmoud, A. Hussain, and H. Tawfik, "Skin cancer classification using explainable artificial intelligence on pre-extracted image features," *Intelligent Systems with Applications*, vol. 20, p. 200275, 2023.

[76] I. A. Ahmed, E. M. Senan, H. S. A. Shatnawi, Z. M. Alkhraisha, and M. M. A. Al-Azzam, "Multi-models of analyzing dermoscopy images for early detection of multi-class skin lesions based on fused features," *Processes*, vol. 11, no. 3, p. 910, 2023.

[77] S. İlkin, T. H. Gençtürk, F. K. Gülağız, H. Özcan, M. A. Altuncu, and S. Şahin, "hybsvm: Bacterial colony optimization algorithm based svm for malignant melanoma detection," *Engineering Science and Technology, an International Journal*, vol. 24, no. 5, pp. 1059–1071, 2021.

[78] M. S. Pitchiah and T. Rajamanickam, "Efficient feature based melanoma skin image classification using machine learning approaches." *Traitement du Signal*, vol. 39, no. 5, 2022.

[79] T. M. Alam, K. Shaukat, W. A. Khan, I. A. Hameed, M. U. Sarwar, and I. A. Hameed, "An efficient deep learning-based skin cancer classification model using ham10000 dataset," *Diagnostics*, 2022.

[80] M. Arshad, M. A. Khan, U. Tariq, A. Armghan, and S. Almotiri, "A computer-aided diagnosis system using deep learning for melanoma detection," *Journal of Healthcare Engineering*, 2021.

[81] I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and H. Mirza, "Automated multi-class classification of skin lesions using deep learning," *Computers in Medicine and Imaging*, 2021.

[82] I. Abunadi and E. M. Senan, "Deep learning and machine learning techniques of diagnosis dermoscopy images for early detection of skin diseases," *Electronics*, vol. 10, no. 24, p. 3158, 2021.

[83] S. Gopikha and M. Balamurugan, "Regularised layerwise weight norm based skin lesion classification using deep learning," *Computer Systems Science and Engineering*, 2023.

[84] Z. Lan, S. Cai, X. He, and X. Wen, "Fixcaps: An improved capsules network for diagnosis of skin cancer using the ham10000 dataset," *IEEE Access*, 2022.

[85] S. Moldovanu, F. A. D. Michis, K. C. Biswas, and C. Culea, "Skin lesion classification based on surface fractal features," *Cancers*, 2021.

[86] X. Deng, Q. Yin, and P. Guo, "Efficient structural pseudoinverse learning-based method for skin lesion classification," *Complex & Intelligent Systems*, 2022.

[87] D. N. Anggraini Ningrum, S.-P. Yuan, W.-M. Kung, C. H. Tang, and C.-H. Yu, "Deep learning classifier with patient's metadata for improving skin lesion diagnosis," *Journal of Multidisciplinary Healthcare*, 2021.