A Fair Multimodal Deep Learning System for Skin Cancer Detection in Patients with Diverse Skin Types

Author: Nurbol Agybetov, Yermek Khaknazar, Baglan Yessenkeldi, Tokhtar Nuralin

Instructor: Seitenov Altynbek

Programme: Research Methods and Tools

Department: Software Engineering

Institution: Astana IT university

Date: 10.10.2025

Table of content

1. Introduction

1.1. Background and Research Context

One of the most prevalent types of cancer globally is skin cancer, and, especially in the case of melanoma, early diagnosis is key for effective treatment and survival of the patient. Success of melanoma treatment is timing.1 Subjectivity and the experience level of the clinician involved are some of the factors in traditional diagnosis through a visual inspection. Over the past few years, revolutionary advancements within Deep Learning technologies and, in particular, Convolutional Neural Networks (CNNs) in the analysis of medical images is a significant breakthrough.

Comprehensive studies demonstrate that current DL models achieve the same level of diagnostic accuracy as many trained dermatologists or even surpass them [1, 1]. In Haenssle et al. (2018), e.g., an Inception v4 convolutional neural network surpassed 58 dermatologists representing 17 countries with better specificity matched to the same level of sensitivity. Similarly, Jinnai et al. (2020) reported their model outperforming 20 experts when judging routine clinical photos. These results demonstrate the tremendous potential DL possesses for achieving diagnostic accuracy enhancement, speed, as well as access. However, despite such exemplary diagnostic success under lab circumstances, their implementation under clinical circumstances has been negligible. The discrepancy between potentiality versus implementation reality reflects presence of underlying flaws running deep even beyond classification accuracy.

1.2. Problem Statement

The key issue is that the current best solutions are extremely precise on test data sets but do not guarantee reasonable, reliable, and trustworthy utilization under actual clinical practice. Emphasis on precision led to neglecting key concerns such as fairness issues in algorithms, comprehension from the clinical environment, and clear decision-making. Present models have three associated issues:

1. Systemic Algorithmic Bias. The majority of the common data sets utilized to train models, such as the ISIC Archive and HAM10000, predominantly consist of images of fair-skinned individuals [1, 1]. Models trained under this imbalanced data have significantly and perilously declining performance when experimented on images of individuals with varied skin tones, thereby aggravating healthcare inequalities. The utilization of this type of tool in a heterogeneous clinical environment is not only from a technical perspective but also poses danger to individuals from certain groups by producing incorrect outcomes. Thus, the problem is not only from the "research gap" but also encompasses medical ethics as well as patient safety.

2. Reductionist Unimodal Method. The majority of the systems consider only the dermoscopic images but neglect vital clinical information such as patient information, medical histories, and doctors' comments that are utilized by the human experts to arrive at decisions [1, 1]. This pigeonholes the work the systems are capable of doing and makes them tools for pattern recognition rather than complete clinical decision support systems. The "Black Box" Issue. The black box nature of deep learning models makes them uninterpretable, therefore doctors are reluctant towards them. They are unable to monitor the process by which an AI provided the diagnosis [1, 1]. It is hard to utilize the tools, since the doctor won't be able to trust the recommendation without the justification for the recommendation.

1.3. Research Question

The principal research question is this: "How do we design and verify a mixed deep learning model that integrates dermoscopy images with both structured and unstructured clinical data to reduce demographic bias and enhance readability in the diagnosis of different patient groups for skin cancer?"

1.4. Relevance and Significance of the Research

The article is informative and useful for the following aspects:

- Reducing Health Disparities. The endeavor is centered on the serious issue of health-based inequality in medical AI. Using the method of "Fairness by Design," the research is centered on creating an instrument that is equally good to all the patient groups.

- Adding Clinical Value. By evolving from pattern recognition to the construction of an actual clinical decision support system that resembles the overall practice of an expert clinician, this work bridges the gap from pattern recognition to the implementation of an effective clinical support tool.

- Developing Trust for Medical AI. Above all, an Explainable AI (XAI) method is required such that medical AI is no longer only research conceptualization but is something trustworthy for doctors, according to Layode et al. (2019) [1, 1].

2. Literature Review

Thematic analysis of scientific journals illustrates the main trends of development, outcomes, and controversial problems in diagnostics of skin cancer using deep learning. The thematic review ensures

better understanding of conducted research, summarizes amassed data, and establishes rational grounds for further research.

2.1. Evolution of Architectural Solutions: From Transfer Learning to Ensembles

Transfer learning has become a baseline approach in this field. Previously published and landmark studies have irrefutably shown that the application of pre-trained convolutional neural networks on large-scale datasets (e.g., ImageNet), i.e., VGG, ResNet, and Inception, significantly outperforms from-scratch trained models and conventional machine learning methods by a great margin. The works of Kalouche (2016) and Rahi et al. (2019) are good examples demonstrating the success of this approach and rendering it a standard baseline methodology.

Creating model ensembles was the next approach for improving reliability and accuracy. Instead of building a single, extremely powerful network, researchers began combining predictions from various models, initially as a way to compensate for the weaknesses of each and later to achieve superior performance. For instance, Imran et al. (2022) demonstrated that an ensemble containing three models was able to achieve an accuracy of 93.5%, a performance that surpassed the best individual model in the set by 14%. Evidence from Ghosh et al. (2024) and Bajwa et al. (2020) also provide support for this. Moreover, the importance of task-specific design is illustrated by the segmentation first, then classification architecture by Yu et al. (2017), first-prize winner of the ISBI 2016 challenge, as well as other specially dedicated designs.

2.2. Data Strategies: From Preprocessing to Taxonomy Utilization

Scientific sources agree that data quality and representativeness can be just as important as model architecture.1 Sources like those from Vijayalakshmi (2019) and Mahmud et al. (2025) pay special attention to image preprocessing as a required step to remove artifacts such as hair, glare, and air bubbles.

A common issue is class imbalance, where benign lesions are far more frequent than malignant lesions. The papers rectify this, and highlight three primary methods:

1. Resampling uses methods like undersampling or oversampling such that the classes will be even, as illustrated by Gururaj et al. (2023) and Mijwil (2021).
2. Class Weights: Clarifying the loss functions for the rare but very informative classes, as illustrated by Ghosh et al. (2024).
3. Data Augmentation: Adding more data to the dataset by modifying the images, in one form or another. Data augmentation is performed to create well-balanced classes as well as to help prevent overfitting.

2.3. New Frontiers: Enhancing Clinical Value Beyond the Image

The field is maturing and the research focus is starting to shift from purely visual analysis to incorporating additional information to increase the clinical applicability of the models. This is the beginning of a more holistic approach to developing a system for diagnosis.

- Integration of Patient Metadata. The research by Ahmadi Mehr & Ameri (2022) was major step in showing that adding simple patient metadata (age, sex, anatomical location of the lesion) to the image improved classification accuracy by over 5% [1, 1]. This provided measurable evidence that an entirely image-based approach is inadequate and negates important knowledge that a physician would always apply.
- Use of Domain Knowledge via Taxonomy. A novel research study by Bajwa et al. (2020) proposed using a disease taxonomy to aid in classification across a vast amount of classes (n=23). By providing the model with the training of the hierarchical relationships between diagnoses, the context of the medical knowledge was embedded into the model which was reflected in model accuracy [1, 1].
- Building Trust through Explainable AI (XAI). For the "black box" dilemma researchers have started implementing XAI. Mahmud et al. (2025) utilized Grad-CAM and Saliency Maps methods to evaluate the areas of the image attended to by the model in decision making. Layode et al. (2019) developed a model that, in addition to the diagnosis, yields similar case examples from a database to substantiate its conclusion [1, 1]. These examples reveal an apparent trend toward transparent and interpretable models.

2.4. Synthesis and Identification of Research Gaps

Critical review of the literature emerges several key gaps preventing the equitable and robust deployment of DL systems in clinical situations. These gaps result from the aforementioned trends and limitations.

1. Fairness and data diversity gap. Even as newer and more sophisticated architectures have been deployed, their credibility is eroded by the known lack of diversity of the benchmark datasets (Dildar et al., 2021; Naqvi et al., 2023) [1, 1]. This is the single most important and unsolved issue that limits the worldwide applicability of models.
2. Multimodal integration gap. While prior works (Ahmadi Mehr & Ameri, 2022; Bajwa et al., 2020) demonstrate the advantages of utilizing non-visual data, they have not yet been adopted routinely. 1 There is no systematic method to integrate images, structured metadata and unstructured text in a single architecture.
3. Explainability and trust gap. There are XAI technologies (Mahmud et al., 2025) but use is sporadic, and almost no research exists on how physicians use such explanations, or if these explanations really lead to increased levels of diagnostic trust [1, 1].
4. Clinical validation gap. Almost all studies use retrospective designs conducted on carefully curated datasets. There is a "glaring shortage" of prospective studies that would provide evaluations of model performance in practice and real clinical environments, and how it can impact physician workflows [1, 1].

The following table clearly links key works with the identified research gaps, forming a logical basis for the proposed project.

Author(s) and Year	Main Contribution	Connection to Identified Gap

Haenssle et al. (2018)	Proved that a CNN can outperform a large group of dermatologists.	Establishes a performance benchmark, but on a dataset lacking sufficient diversity.
Ahmadi Mehr & Ameri (2022)	Showed an accuracy increase of >5% by adding patient metadata.	**Multimodality Gap:** Provides quantitative proof of the inadequacy of image-only models.
Bajwa et al. (2020)	Used disease taxonomy for classification across 23 classes.	**Multimodality Gap:** Demonstrates the power of integrating structured medical knowledge.
Naqvi et al. (2023) / Dildar et al. (2021)	Review articles explicitly pointing to data bias towards light skin as the main problem in the field.	**Fairness Gap:** Provides authoritative, meta-analytic evidence of the core problem.
Mahmud et al. (2025)	Implemented XAI (Grad-CAM) to visualize model decisions.	**Explainability Gap:** Shows the availability of XAI tools but highlights their non-standardized application.

3. Research Design and Methods

This section illustrates an explicit plan to provide an answer to the research question. It describes the planned design of the system, the data approach with an emphasis on fairness, and an extremely rigorous evaluation plan.

3.1. Aim and Objectives of the Research

Aim: The research aim is to develop multimodal diagnostic system (FMDS) that takes into account structured clinical data and different skin types that improve objectivity of early skin cancer detection.

Objectives:

1. The goal is to design and deploy an original hybrid neural network architecture (FMDS) capable of processing and integrating simultaneously the dermoscopic images, structured patient metadata.
2. The goal is to design and integrate a "Fairness by Design" protocol under the model training process utilizing stratified sampling and fairness-aware loss functions to identify and minimize bias related to skin phototype.
3. To integrate a multimodal XAI module that provides holistic explanations for the prediction made by the system, illustrating the role of visual as well as non-visual features
4. The goal is to carry out a rigorous evaluation of the FMDS by comparing its performance to unimodal baselines models, with both traditional accuracy metrics and established fairness metrics.

3.2. Research Design

This study employs a quantitative, experimental research design based on secondary data sources.

- **Quantitative:** The study is considered quantitative, as it includes the assessment of model performance using objective numerically based metrics, e.g., area under the ROC curve (AUC-ROC), F1-score, as well as established fairness indexes. e.g., equal opportunity difference and statistical parity difference.
- **Experimental:** The study is considered experimental because it is developing a new system (FMDS) and enacting a series of controlled experiments (ablation studies) to systematically assess the generated prototypes against a range of baseline models. This approach allows for validation of the contribution of each proposed architecture component.
- **Secondary sources:** The study will utilize existing publicly available datasets (secondary sources, e.g., HAM10000, ISIC Archive), and no primary data will be collected from individuals.

3.3. Methods and Sources

This section describes the tools, data sources, and procedures for collecting, selecting, and analyzing data.

3.3.1. Proposed FMDS Architecture (Tools and Procedures)

A three-parallel-branch architecture is proposed. Each branch is made to process one type of data, namely:

- **Image Processing Branch:** The feature extraction will be done by utilizing an extremely efficient and recently effective type of convolutional neural network, namely EfficientNetV2. The reason behind utilizing this network is the very high accuracy-to-computing-cost ratio, rendering it suitable for potential utilization in clinics.

- **Structured Data Processing Branch:** A basic Multi-Layer Perceptron (MLP) will process structured data, for example, age, sex, where the lesion is located, and most importantly, the Fitzpatrick skin type.

- **Text Processing Branch:** A special type of language model, called BioBERT, will be used to find features in unstructured clinical notes (for example, patient history). BioBERT was chosen because it has been trained on a lot of biomedical literature, making it better for this job than general language models.

Fusion Module: The feature vectors from the three branches will be combined and fed into the final set of fully connected layers for making the final classification.

3.3.2. Data Sources and "Fairness by Design" Protocol (Data Selection and Processing)

- **Datasets:** The only datasets are HAM10000 and the ISIC Archive. One important goal is to include datasets where the images come from subjects with different skin tones (e.g., the Diverse Dermatology Images dataset), if they are available, to directly answer the fairness question.

- **"Fairness by Design" Protocol:** This is the foundation of the planned research.

    - **Clear Stratification:** All patients will be grouped by their Fitzpatrick skin type.

    - **Balanced Batch Sampling:** In the process of training, each data batch will have all the skin types. This prevents the majority group from dominating the training process.

    - **Fairness-Aware Loss Function:** A re-weighting/debiasing method will be utilized in the loss function to directly punish the model for generating predictions that are related to the sensitive feature (skin phototype).

3.3.3. Multi-faceted Evaluation Protocol (Data Analysis)

- **Performance Metrics:** Standard metrics will be used: AUC-ROC, F1-score, Precision, and Recall.

- **Fairness Metrics:** Existing fairness metrics will be used to validate the primary hypothesis of the project: Equal Opportunity Difference and Statistical Parity Difference.

- **Ablation Studies:** There will be a sequence of experiments to evaluate the contribution of each system component. The full FMDS model will be compared to a unimodal model, a bimodal model, and the complete model without the fairness-aware loss function.

- **XAI Evaluation:** We will qualitatively evaluate the output of the XAI module for consistency and clinical relevance. Explanations will summarize the important regions of the image (using Grad-CAM) in addition to highlighting important metadata and text snippets (using SHAP values) that led to the diagnosis.

3.4. Practical Considerations, Limitations, and Ethical Issues

This section addresses potential obstacles, the limitations of the study, and relevant ethical issues.

- **Potential Obstacles:**

- **Data Availability:** The ambition to increase datasets through a more diverse image dataset hinges on the available datasets and their quality. If the datasets are not available, the model will arguably have limited merit in bias reduction.

- **Data Annotation:** Some datasets we use may be missing metadata such as Fitzpatrick skin type and thus will require a manual annotation that is both time-consuming and subjective in its labeling

- **Technical Challenges with Developing a Model:** Successfully integrating and training a three-branch multimodal architecture can have serious technical challenges associated with convergence related issues and high computational costs.

- **Limitations of the Research:**

  - **Retrospective Data Set:** Because the study is on datasets that exist, it is retrospective. While the FMDS may have good performances, the performance of the FMDS in a real-time prospective clinical setting cannot be guaranteed, and a separate validation must occur.

  - **Generalizability Issues:** While this study is intended to improve fairness, the model performances will be limited by the diversity of the final combined dataset and many not generalize equally to every demographic or every recalcitrant skin condition.

  - **Explainability Issues:** XAI methods namely Grad-CAM or SHAP provide useful information, but those reasoning interpretations are approximations of the model reasoning and thus should not be considered as a learning classification.

- **Ethical Issues:**

  - **Data privacy:** Throughout this research, only openly available (anonymized) data will be used to ensure patient privacy. All activities and procedures will be compliant with data protection policies.

  - **Responsible AI:** The ethical impetus is to minimize diagnostic bias. Nonetheless, it is recognized that any AI model may perpetuate latent bias. Accordingly, the findings will be interpreted with appropriate caution, and any potential deployment in a real-world context would require clinical trials and regulatory procedures.

4. Implications and contributions to knowledge

This research seeks to help advance medical AI at the intersection of science theory and clinical practice.

4.1. Theoretical Contribution

- **A Blueprint for Fair AI in Medicine:** "Fairness by Design" will be a starting point for addressing demographic inequity in other medical imaging specialties.
- **Advancement of Multimodal Fusion Methods:** This research will contribute to the understanding of integrated empirical data created by fused imaging, along with structured and unstructured data in clinical environments.
- **Development of Multimodal XAI:** This initiative will pioneer the research and validation of explanation models that stretch across different data modalities, a critical and emerging XAI research area.

4.2. Practical Significance

- **Reducing Healthcare Disparities:** The FMDS, upon completion, has the potential to be a foundational diagnostic tool that equates the quality of care and mitigates misdiagnosis in people with dark skin.
- **Improving Diagnostic Accuracy and Confidence:** The system will provide a comprehensive "second opinion" that rationalizes the entire patient context, thus improving diagnostic accuracy and augmenting clinician trust in AI tools.
- **Foundation for Prospective Clinical Trials:** The groundwork for future prospective studies aimed at validating the FMDS in clinical practice will be built as a direct result of this research. This will help close the gap between its development in a research lab and its actual use in the clinical setting.

5.Reference

1. Abunadi, I., & Senan, E. M. (2021). Deep Learning and Machine Learning Techniques of Diagnosis Dermoscopy Images for Early Detection of Skin Diseases. Electronics, 10(24), 3158. https://doi.org/10.3390/electronics10243158
2. Ahmadi Mehr, R., & Ameri, A. (2022). Skin Cancer Detection Based on Deep Learning. Journal of Biomedical Physics & Engineering, 12(6), 559-568. https://doi.org/10.31661/jbpe.v0i0.2207-1517
3. Akinrinade, O., & Du, C. (2025). Skin Cancer Detection Using Deep Machine Learning Techniques. Intelligence-Based Medicine, 11, 100191. https://doi.org/10.1016/j.ibmed.2024.100191

4. Ameri, A. (2020). A Deep Learning Approach to Skin Cancer Detection in Dermoscopy Images. Journal of Biomedical Physics & Engineering, 10(6), 801-806. https://doi.org/10.31661/jbpe.v10i6.2004-1107
5. Bajwa, M. N., Muta, K., Malik, M. I., Siddiqui, S. A., Braun, S. A., Homey, B., Dengel, A., & Ahmed, S. (2020). Computer-Aided Diagnosis of Skin Diseases Using Deep Neural Networks. Applied Sciences, 10(7), 2488. https://doi.org/10.3390/app10072488
6. Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., Alsaiari, S. A., Saeed, A. H. M., Alraddadi, M. O., & Mahnashi, M. H. (2021). Skin Cancer Detection: A Review Using Deep Learning Techniques. International Journal of Environmental Research and Public Health, 18(10), 5479. https://doi.org/10.3390/ijerph18105479
7. Ghosh, H., Rahat, I. S., Mohanty, S. N., Ravindra, J. V. R., & Sobur, A. (2024). A Study on the Application of Machine Learning and Deep Learning Techniques for Skin Cancer Detection. International Journal of Computer and Systems Engineering, 18(1).
8. Gururaj, H. L., Manju, N., Nagarjun, A., Manjunath Aradhya, V. N., & Flammini, F. (2023). DeepSkin: A Deep Learning Approach for Skin Cancer Classification. IEEE Access, 11, 50205-50215. https://doi.org/10.1109/ACCESS.2023.3274848
9. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A., & Uhlmann, L. (2018). Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists. Annals of Oncology, 29(8), 1836–1842. https://doi.org/10.1093/annonc/mdy166
10. Imran, A., Nasir, A., Bilal, M., Sun, G., Alzahrani, A., & Almuhaimeed, A. (2022). Skin Cancer Detection Using Combined Decision of Deep Learners. IEEE Access, 10, 118198-118214. https://doi.org/10.1109/ACCESS.2022.3220329
11. Jinnai, S., Yamazaki, N., Hirano, Y., Sugawara, Y., Ohe, Y., & Hamamoto, R. (2020). The Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning. Biomolecules, 10(8), 1123. https://doi.org/10.3390/biom10081123
12. Kalouche, S. (2016). Vision-Based Classification of Skin Cancer Using Deep Learning. Stanford University.
13. Layode, O., Alam, T., & Rahman, M. M. (2019). Deep Learning Based Integrated Classification and Image Retrieval System for Early Skin Cancer Detection. 2019 IEEE International Conference on Big Data (Big Data), 5521-5526. https://doi.org/10.1109/BigData47090.2019.9006206
14. Mahmud, M. A. A., Afrin, S., Mridha, M. F., Alfarhood, S., Che, D., & Safran, M. (2025). Explainable Deep Learning Approaches for High Precision Early Melanoma Detection Using Dermoscopic Images. Scientific Reports, 15, 24533. https://doi.org/10.1038/s41598-025-09938-4
15. Mijwil, M. M. (2021). Skin Cancer Disease Images Classification Using Deep Learning Solutions. Multimedia Tools and Applications. https://www.google.com/search?q=https://doi.org/10.1007/s11042-021-10952-7
16. Naqvi, M., Gilani, S. Q., Syed, T., Marques, O., & Kim, H.-C. (2023). Skin Cancer Detection Using Deep Learning—A Review. Diagnostics, 13(11), 1911. https://doi.org/10.3390/diagnostics13111911
17. Rahi, M. M. I., Ullah, A. K. M. A., Khan, F. T., Alam, M. G. R., Mahtab, M. T., & Alam, M. A. (2019). Detection of Skin Cancer Using Deep Neural Networks. 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). https://doi.org/10.1109/ICASERT.2019.8934512

18. Vijayalakshmi, M. M. (2019). Melanoma Skin Cancer Detection Using Image Processing and Machine Learning. International Journal of Trend in Scientific Research and Development, 3(4), 780-784.
19. Wei, L., Ding, K., & Hu, H. (2020). Automatic Skin Cancer Detection in Dermoscopy Images Based on Ensemble Lightweight Deep Learning Network. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2997710
20. Yu, L., Chen, H., Dou, Q., Qin, J., & Heng, P. A. (2017). Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. IEEE Transactions on Medical Imaging, 36(4), 994-1004. https://doi.org/10.1109/TMI.2016.2642839

6. Research Schedule

To demonstrate the feasibility of the project within a standard PhD or Master's program timeline, the following detailed schedule is proposed [1, 1].

Phase	Key Tasks	Estimated Duration	Deliverables
Phase 1: Preparatory	Complete literature review; Collect and preprocess datasets; Develop protocol for skin phototype annotation.	Months 1-3	Annotated bibliography; Prepared dataset.
Phase 2: Model Development	Implement and train baseline unimodal models; Develop and train the three-component FMDS architecture.	Months 4-7	Working code for baseline models and FMDS.

Phase 3: Fairness and XAI Integration	Implement and tune the fairness-aware loss function; Develop and integrate the multimodal XAI module.	Months 8-10	Final version FMDS v1.0 with integrated fairness and XAI modules.
Phase 4: Evaluation and Analysis	Conduct comprehensive performance and fairness evaluation; Perform ablation studies; Analyze XAI results.	Months 11-14	Full set of experimental results and graphs.
Phase 5: Dissemination of Results	Write thesis/paper; Prepare manuscript for submission to a journal/conferenc e.	Months 15-18	Draft of thesis; Submission-ready manuscript.