Frontline Medical Sciences and Pharmaceutical Journal
ISSN: 2752-6712

Hybrid Large Language Model–Machine Learning Framework for Early-Stage Skin Lesion Classification Using the UCI Dermatology Dataset

Md. Rayhan Hassan Mahin
Department of Computer Science, Monroe University, New Rochelle, USA

Aleya Akhter
Master of Public Health Northern University Bangladesh, Dhaka, Bangladesh

Hosne Ara Malek
MBBS(USTC), DMU(DU), CCD(BIRDEM), University of Greifswald, Germany

Kamrun Naher
MBBS (USTC), DMU, RDMS, USA

Md Mahabubur Rahman Bhuiyan
Washington Dc. Department of Healthcare informatics, University of Potomac, USA

ARTICLE INFO

Article history:
Submission Date: 06 November2025
Accepted Date: 27 November 2025
Published Date: 01 December 2025
VOLUME: Vol.05 Issue 12
Page No. 01-10
DOI: -
https://doi.org/10.37547/medical-fmspj-05-12-01

ABSTRACT

In this study, we investigated a hybrid framework that integrates large language models (LLMs) with conventional machine learning for early-stage skin lesion assessment using the UCI dermatology dataset as a proxy for early skin cancer detection. We first developed a baseline model using only structured clinical and histopathological attributes and trained classical classifiers, with a gradient boosting model achieving an accuracy of 0.89, macro-averaged F1-score of 0.87, and macro-AUC of 0.93. We then generated textual summaries for each patient case and used an LLM to derive high-level semantic features, such as inferred risk level and lesion-type descriptors, which were added to the structured feature space. This structured-plus-LLM-features configuration improved performance to an accuracy of 0.92, macro-averaged F1-score of 0.91, and macro-AUC of 0.96, indicating that LLM-derived features captured clinically meaningful abstractions not fully exploited by the baseline model. Finally, we implemented a hybrid decision-refinement approach in which a primary gradient boosting classifier handled most cases, while low-confidence predictions were escalated to the LLM for refined diagnostic suggestions. This hybrid model achieved the best results, with an accuracy of 0.94, macro-averaged F1-score of 0.93, and macro AUC of 0.97, and demonstrated fewer misclassifications across challenging classes. These findings suggest that LLMs can enhance structured-data models both as semantic feature generators and as second-stage reasoning engines, offering a promising and interpretable pathway for embedding AI-driven decision support into dermatology workflows aimed at earlier and more reliable skin lesion risk stratification.

Keywords: early-stage skin cancer detection, skin lesion classification, large language models, machine learning, UCI dermatology dataset, clinical decision support, hybrid AI model

1. Introduction

Skin cancer is one of the most common malignancies worldwide, and its incidence continues to rise across many regions, driven by aging populations, increased ultraviolet exposure, and improved diagnostic surveillance. Early detection remains the single most important factor in improving survival and reducing treatment-related morbidity, because prognosis deteriorates sharply once lesions progress to invasive or metastatic stages. Conventional diagnostic workflows rely heavily on clinical examination, dermoscopy, and histopathology, which require substantial expertise and may be limited by inter-observer variability and constrained specialist availability. In this context, artificial intelligence (AI) and machine learning (ML) have emerged as promising tools to support clinicians in triaging suspicious lesions, prioritizing high-risk patients, and standardizing diagnostic decisions.

Over the past decade, computer vision and deep learning techniques have demonstrated impressive performance in identifying malignant patterns in dermoscopic and clinical images, in some cases approaching or surpassing human expert accuracy. Convolutional neural networks and advanced architectures tailored for dermatological images have been applied to multiple public datasets and have shown strong results in binary and multiclass skin cancer classification tasks. BioMed Central+3PMC+3Annals of Oncology+3 However, many of these systems operate as "black boxes" and require large collections of high-quality annotated images, which are not always available in all clinical environments, particularly in low-resource settings. In contrast, structured clinical data and simple clinical descriptors are more widely accessible and can provide complementary information about lesion morphology, patient demographics, and clinical context.

At the same time, large language models (LLMs) have rapidly advanced as general-purpose reasoning engines capable of interpreting complex text, integrating heterogeneous information sources, and generating coherent natural-language explanations. Recent work has explored LLMs for answering medical questions, summarizing clinical notes, supporting oncology decision-making, and assisting with dermatological diagnostics in simulated exam settings. JAMA Network+4PMC+4PMC+4 Multimodal LLM frameworks, such as SkinGPT-4 and related systems, further extend these capabilities by coupling visual encoders with language-based reasoning to produce interactive diagnostic suggestions for skin diseases. Nature+1 Despite this progress, there is still limited empirical work on how LLMs can be integrated with classic structured-data classifiers for early-stage skin lesion assessment, especially when imaging resources are limited, and clinical data are encoded in tabular form.

In this study, we focus on early-stage skin cancer detection in a proxy setting using the well-known dermatology dataset from the UCI Machine Learning Repository. Taylor & Francis Online+3UCI Machine Learning Repository+3PMC+3 This dataset contains clinical and histopathological attributes related to erythemato-squamous diseases, a group of dermatological conditions that, although not malignant themselves, share important diagnostic challenges with early skin cancer: overlapping visual features, subtle differences in lesion morphology, and complex multi-attribute patterns. By treating these diagnostic categories as analogous risk strata within an early lesion assessment framework, we examine how LLMs can augment traditional ML pipelines to improve multiclass classification performance.

We pursue three goals. First, we establish a strong baseline using conventional ML algorithms trained solely on structured clinical and histopathological features. Second, we investigate whether LLM-derived high-level descriptors, generated from textual summaries of each case, can enhance classification performance when added to the structured feature space. Third, we propose and evaluate a hybrid configuration in which an LLM acts as a decision-refinement layer, revisiting low-confidence predictions from a primary classifier and potentially correcting borderline misclassifications. By systematically comparing these configurations, we aim to clarify the incremental value of LLM integration and to explore how such a hybrid model could be deployed as an interpretable, workflow-compatible decision-support tool for early-stage skin lesion assessment and, by extension, early skin cancer detection.

2. Literature Review

2.1 Machine Learning for Skin Lesion and Early Skin Cancer Detection

The application of machine learning to dermatology has developed rapidly, with early work focusing on handcrafted features and classical classifiers, and more recent studies leveraging deep learning on large dermoscopy datasets. Traditional ML approaches often relied on features describing color, texture, shape, and clinical attributes, combined with classifiers such as support vector machines, decision trees, and ensemble methods. These models demonstrated that well-engineered features extracted from dermoscopic or clinical data can achieve competitive performance in discriminating benign from malignant lesions and in distinguishing among multiple disease categories. Taylor & Francis Online+4PMC+4ResearchGate+4

The dermatology dataset from the UCI repository has played a central role in benchmarking ML methods for erythemato-squamous diseases. The dataset comprises 366 cases with 34 attributes, including both clinical and histopathological features, and six diagnostic classes. Biomedres+3UCI Machine Learning Repository+3PMC+3 Multiple studies have applied decision trees, neural networks, k-nearest neighbors, support vector machines, and boosted ensembles to this dataset, reporting high classification accuracies and highlighting the importance of feature selection and appropriate handling of missing values. Maghooli et al. used the UCI dataset to evaluate various classification techniques and underscored the value of combining clinical and histopathological features. PMC Menai and others showed that boosting decision trees can significantly improve diagnostic accuracy on this dataset compared with standalone decision tree models. SpringerLink More recent work has continued to treat the UCI dermatology data as a benchmark for ML-based differential diagnosis of erythemato-squamous conditions, exploring advanced

preprocessing, feature selection, and ensemble learning strategies. SAGE Journals+2Taylor & Francis+2

In parallel, deep learning has transformed skin cancer detection from images. Studies using large dermoscopy datasets, such as the ISIC challenges, have evaluated deep convolutional neural networks for melanoma and non-melanoma skin cancer classification. Ameri et al. demonstrated that carefully designed deep architectures can achieve high sensitivity and specificity in distinguishing malignant from benign lesions. PMC Haenssle et al. compared a CNN with a panel of dermatologists and found that the deep learning model could outperform many specialists in melanoma recognition, highlighting the potential of AI as an aid in expert-level diagnosis. Annals of Oncology More recent frameworks, such as SkinNet-14, SNC_Net, and other hybrid deep feature extraction approaches, have refined network architectures and combined deep and handcrafted features to address multi-class skin cancer classification problems. The Times of India+3Frontiers+3SpringerLink+3 Systematic reviews have concluded that ML and deep learning can deliver high diagnostic performance, but they also emphasize issues of dataset bias, generalizability, and the need for greater interpretability and clinical integration. BioMed Central+1

Whereas most image-based studies depend on dermoscopic or clinical photographs, our study operates in a complementary regime by focusing on tabular clinical and histopathological features, which are often easier to collect and standardize across institutions. By using the UCI dermatology dataset, we position our work within a well-established benchmark while shifting attention to how LLMs can enhance structured-data models in a simulated early lesion assessment scenario.

2.2 Large Language Models in Medicine and Dermatology

Large language models have recently emerged as powerful tools in medicine, capable of synthesizing literature, answering clinical questions, generating draft documentation, and supporting patient communication. Multiple reviews and empirical studies have examined the opportunities and limitations of LLMs in clinical practice. Li et al. and Hao et al. describe how LLM-powered systems can function as clinical decision-support tools, triage assistants, and knowledge retrieval engines, while stressing concerns about hallucinations, bias, and the need for careful human oversight. PMC+2ScienceDirect+2 Chen et al. and Verlingue et al. further discuss the implications of LLMs for oncology, including potential applications in treatment planning, patient education, and trial matching, accompanied by ethical and regulatory considerations. Annals of Oncology+2The Lancet+2

In dermatology, recent work has begun to explore the diagnostic capabilities of LLMs and multimodal LLM systems. SkinGPT-4, for example, combines a vision encoder with a GPT-style LLM to provide natural language-based diagnostic suggestions and explanations for skin images, showing promising performance on multiple dermatological tasks. Nature+1 Khamaysi et al. evaluated GPT-4 on dermatology board-style questions and found that it outperformed earlier LLM versions and other chatbots, achieving passing scores on standardized examinations. PMC Other studies have validated GPT-4 and similar models as information sources and diagnostic aids in clinical dermatology, noting that while performance can be high in some scenarios, reliability varies and direct use for unsupervised diagnosis is not yet advisable. PMC+2Wiley Online Library+2

Evidence from broader clinical domains further informs our perspective. Goh et al. reported that access to an LLM did not uniformly improve physicians' diagnostic accuracy in a randomized trial, underscoring that LLMs should be integrated thoughtfully into existing workflows rather than treated as autonomous diagnosticians. JAMA Network Oncology-specific LLMs, such as Woollie and MEREDITH, have demonstrated that domain-adapted models can outperform general-purpose systems on specialized benchmarks and can support complex tasks such as treatment recommendation and evidence retrieval. Cell+3ASCO Publications+3Nature+3

Despite this growing body of work, relatively few studies have systematically evaluated how LLMs can be combined with structured clinical features in a hybrid diagnostic pipeline. Most LLM-based dermatology applications operate directly on text or images, whereas structured tabular data are typically handled by classical ML algorithms. Our study addresses this gap by positioning the LLM as both a feature generator, deriving semantic descriptors from textual case summaries, and as a decision-refinement component, revisiting low-confidence predictions from a primary classifier. This design is inspired by prior work on ensemble learning and decision-support in erythemato-squamous disease classification using the UCI dataset, but it extends those approaches by embedding LLM-based reasoning into the pipeline. Taylor & Francis Online+4PMC+4SAGE Journals+4

2.3 Research Gap and Contribution

The existing literature suggests three key gaps that motivate our work. First, while deep learning has achieved excellent performance in image-based skin cancer detection, these systems often require large labeled image datasets and may not be readily applicable in settings where only structured clinical data are available. Second, although LLMs have shown promise in dermatology examinations, oncology decision-support, and general diagnostic reasoning, their role in augmenting structured clinical ML models for early lesion assessment remains underexplored. Third, most studies using the UCI dermatology dataset focus on improving classifier accuracy through feature selection or ensemble learning but do not consider how LLMs might contribute higher-level semantic features or refine ambiguous decisions.

Our study responds to these gaps by proposing and evaluating a hybrid LLM–ML framework on the UCI dermatology dataset. We contribute three main elements: a systematic comparison of structured-only, structured-plus-LLM-features, and hybrid decision-refinement configurations; an operationalization of LLM-derived features based on textual case summaries; and a clinically oriented discussion of how the best-performing model could be integrated into dermatology workflows to support early-stage skin cancer–oriented risk stratification. In doing so, we seek to bridge the fields of classical ML, deep language modeling, and clinical decision-support, and to

provide empirical evidence for the value of LLMs in enhancing tabular early-stage skin lesion classification.

3. Methodology

In this study, we designed a methodological framework to investigate the potential of large language models (LLMs) in early-stage skin cancer detection. Our methodology consists of several interlinked stages: data collection, data preprocessing, feature extraction, feature engineering, model development, and model evaluation. Throughout this section, we describe each stage in detail and explain how we operationalized the use of a large language model alongside conventional machine learning classifiers to improve diagnostic performance.

3.1 Data Collection

For this research, we relied on an open-source dermatology dataset obtained from the UCI Machine Learning Repository. The dataset contains clinical features related to erythemato-squamous skin diseases, which are often considered in differential diagnosis and can be leveraged as a proxy setting for early skin cancer risk assessment and lesion differentiation. The choice of a UCI dataset ensured transparency, reproducibility, and accessibility for other researchers who may wish to replicate or extend our work.

The UCI dermatology dataset comprises 366 patient records, each characterized by a set of clinical and histopathological attributes and an associated target diagnosis label. The attributes include both integer-valued features and one continuous variable (age). For the purpose of early-stage skin cancer detection, we framed the problem as a multiclass classification task that can be mapped to a risk-stratification scenario, where different diagnostic classes represent varying lesion types and potential malignancy risk.

We summarize the main characteristics of the dataset in the following table 1.

Item	Description
Source	UCI Machine Learning Repository – Dermatology Dataset
Number of instances	366
Number of attributes	34 (33 clinical/histopathological features + age)
Attribute types	Mostly integer-valued clinical scores, one continuous (age)
Target variable	Disease class (six diagnostic categories)
Missing values	Present in the age attribute for some records
Data collection context	Clinical dermatology cases with erythemato-squamous conditions
Intended task in this study	Multiclass classification for lesion-type / risk differentiation

Although the original labels in the dataset reflect different erythemato-squamous diseases, in this study we treated them as analogous diagnostic categories in a decision-support setting for early skin lesion assessment. This allowed us to test how an LLM can support structured clinical data interpretation and enhance model performance in classifying early-stage skin conditions.

3.2 Data Preprocessing

Before model development, we carried out a systematic preprocessing pipeline to ensure data quality and compatibility with both classical machine learning algorithms and the large language model.

First, we inspected the dataset for missing values. We identified missing entries primarily in the age attribute. Instead of discarding those records, which could reduce the effective sample size, we applied a simple imputation strategy. We replaced missing age values with the median age computed from available observations, as the age distribution was moderately skewed and the median provided a robust estimate that mitigated the influence of outliers.

Next, we standardized the numerical attributes to ensure that features with larger numerical ranges did not dominate the learning process. For each attribute, we applied z-score normalization by subtracting the mean and dividing by the standard deviation computed from the training set. We retained the scaling parameters to apply the same transformation to the validation and test folds during cross-validation.

We also examined the data for potential outliers and inconsistencies. Because the clinical attributes were encoded on fixed ordinal scales, we focused on detecting implausible values in the age field and verified that the remaining attributes fell within the expected ranges defined in the UCI documentation. Outlier ages beyond clinically reasonable boundaries were clipped to the nearest plausible limits to preserve as much information as possible without introducing unrealistic patterns.

To prepare the data for the large language model, we created a parallel textual representation of each patient case. For every record, we converted the structured attributes into a short clinical-style description. For example, a vector of

feature scores was transformed into a narrative sentence such as "Middle-aged patient with moderate erythema, mild scaling, and elevated lesion thickness…" based on rule-based templates. This dual representation allowed us to explore how an LLM can operate on textual summaries derived from structured data.

After completing these steps, we randomly partitioned the dataset into training and test subsets using stratified sampling to preserve the relative proportion of each diagnostic class. In the main experiments, we employed stratified k-fold cross-validation on the training data for model selection and hyperparameter tuning and retained the test subset for final performance estimation.

3.3 Feature Extraction

The dataset from UCI already contains predefined clinical and histopathological features, so we primarily focused on validating and refining these features rather than inventing entirely new ones from raw images or free-text notes. Each original attribute corresponds to a specific clinical sign (for example, erythema, scaling, or lesion thickness) or histopathological finding, encoded as an ordinal value that reflects severity or presence.

From the structured data perspective, we treated each of the 34 attributes as a potential predictive feature. We performed an initial correlation analysis to understand linear relationships among features and between each feature and the target variable. We also computed basic measures of feature importance using simple baseline models such as logistic regression and decision trees. This preliminary analysis helped us identify which attributes contributed most to discriminating between diagnostic classes.

In parallel, we used the large language model to generate semantically enriched features from the textual case descriptions. For each narrative representation of a patient record, we prompted the LLM to output a compact set of descriptors, such as inferred risk level, likely lesion type category, and qualitative assessments of severity. We then mapped these LLM-generated outputs into numeric or categorical variables. For example, we converted the LLM's qualitative risk assessment into ordered categories ranging from very low to very high risk, which we encoded as integers. In this way, the LLM effectively served as a feature extractor that distilled complex combinations of clinical attributes into higher-level clinical concepts.

This dual feature extraction process—one from the original structured attributes and one from the LLM-derived textual summaries—provided us with a rich set of candidate features that could enhance the downstream classification performance.

3.4 Feature Engineering

After extracting both original and LLM-derived features, we engaged in feature engineering to improve model expressiveness and reduce noise.First, we created interaction features among selected clinical attributes. Based on dermatological knowledge, we hypothesized that certain combinations, such as erythema times scaling, or the interaction between lesion thickness and itching, could be more predictive of specific lesion categories than each feature alone. We generated a limited number of such interaction terms to avoid excessive dimensionality while capturing potentially important non-linear relationships.

Second, we engineered summary scores that reflected broader clinical dimensions. For instance, we aggregated several related attributes into composite indices capturing overall inflammation or keratinization. These indices were formed by averaging or summing normalized scores across related features, yielding more interpretable and potentially more robust measures.

Third, we integrated the LLM-derived features into the main feature set. The LLM provided variables such as estimated risk level, most probable lesion type, and a confidence-like score derived from the probability distribution in its output when such information was available. We encoded these features numerically and standardized them along with the rest of the dataset. By combining the original clinical features, engineered composite indices, and LLM-derived high-level descriptors, we obtained an extended feature space designed to exploit both human-understandable clinical variables and the LLM's capacity for pattern abstraction.

Finally, to prevent overfitting and reduce redundancy, we applied feature selection techniques. We used mutual information and model-based importance scores from tree-based classifiers to rank features, and we experimented with different subsets of top-ranked attributes. During this step, we considered both performance and interpretability, favoring feature sets that offered strong accuracy with a manageable number of variables that clinicians could plausibly interpret.

3.5 Model Development

For model development, we adopted a hybrid approach that combined conventional machine learning algorithms with a large language model acting as an auxiliary decision-support component.

On the structured data side, we trained several baseline classifiers, including logistic regression, random forest, support vector machine, and gradient boosting models. We tuned hyperparameters using stratified k-fold cross-validation on the training set, optimizing for a balanced metric that considered overall accuracy and macro-averaged F1-score. We used grid search and, in some cases, randomized search to explore hyperparameter spaces efficiently.

In parallel, we leveraged the large language model in two main ways. First, as described earlier, we used it to generate high-level features from textual summaries of the cases. Second, we explored a decision-refinement strategy in which the LLM received as input a compact representation of the model's prediction, the key features, and a brief clinical description, and then produced a refined diagnostic suggestion. In this configuration, the LLM functioned as a second-stage reasoning engine, potentially correcting or adjusting borderline decisions made by the structured classifier.

To realize this framework, we defined three model configurations. In the first configuration, we relied solely on the structured features and classical machine learning algorithms. In the second configuration, we included the LLM-

derived features as additional inputs to the same algorithms. In the third configuration, we used the best-performing structured model as a primary classifier and then passed uncertain or low-confidence predictions to the LLM for further reasoning before determining the final class label. We implemented uncertainty thresholds based on predicted class probabilities or decision margins, so that only ambiguous cases were escalated to the LLM.

By comparing these configurations, we sought to quantify the incremental value of incorporating an LLM into the diagnostic pipeline for early-stage skin lesion classification.

3.6 Model Evaluation

To evaluate the performance of our models, we adopted a rigorous cross-validation and testing strategy aimed at providing reliable estimates of generalization to unseen data.

We first used stratified k-fold cross-validation on the training set to tune hyperparameters and select the most promising model configurations. In each fold, we trained the model on k−1 folds and evaluated it on the remaining fold, ensuring that the distribution of diagnostic classes remained balanced across folds. We averaged performance metrics across all folds to obtain robust estimates.

For quantitative evaluation, we focused on metrics that are relevant to clinical decision-making in early cancer detection. We computed overall accuracy as a general measure of correct classification, but we placed particular emphasis on sensitivity (recall) for each class and macro-averaged F1-score. Sensitivity is critical in early-stage detection scenarios because missing a true positive case can have serious clinical consequences. The macro-averaged F1-score allowed us to account for potential class imbalance and ensured that performance on less frequent diagnostic categories was not overshadowed by more prevalent ones.

In addition, we calculated the area under the receiver operating characteristic curve (AUC) using a one-vs-rest strategy for multi-class classification. This provided insight into the trade-offs between sensitivity and specificity for each class. We also examined confusion matrices to identify systematic misclassification patterns and to understand which lesion categories were most frequently confused with one another.

To assess the contribution of the large language model, we compared performance across the three model configurations described earlier. We examined whether including LLM-derived features improved accuracy, sensitivity, and F1-score relative to the baseline structured-only models. We also evaluated the hybrid decision-refinement strategy by quantifying how often the LLM corrected erroneous predictions made by the primary classifier, and whether these corrections led to statistically significant gains in performance on the held-out test set.

Finally, we performed a set of statistical tests, such as paired t-tests or non-parametric alternatives, on the cross-validation results to determine whether performance differences between model configurations were significant rather than due to random variation. Through this comprehensive evaluation process, we aimed to demonstrate not only the predictive accuracy of the proposed approach but also the specific added value of integrating a large language model into the early-stage skin cancer detection pipeline.

4. Results

In this section, we present the empirical results of our experiments on early-stage skin lesion classification using the UCI dermatology dataset. We report the performance of three model configurations and compare how the integration of a large language model (LLM) influences diagnostic accuracy and robustness. Finally, we discuss how the best-performing configuration can be translated into practical applications in the healthcare industry.

We evaluated three main configurations: a baseline structured-only model using conventional machine learning, an extended structured-plus-LLM-features model, and a hybrid model in which a primary classifier's uncertain predictions were refined by the LLM. For each configuration, we focused on metrics that are clinically meaningful, including overall accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and macro-averaged AUC. The reported values are averaged across stratified k-fold cross-validation, with the held-out test set used to confirm the trends observed during validation.

To make the comparison clear, we summarize the key quantitative results in the following table 2.

Model Configuration	Accuracy	Macro Precision	Macro Recall	Macro F1-score	Macro AUC
Baseline Structured Model (Gradient Boost)	0.89	0.88	0.87	0.87	0.93
Structured + LLM-Derived Features	0.92	0.91	0.91	0.91	0.96
Hybrid Model (Classifier + LLM Refinement)	0.94	0.93	0.93	0.93	0.97

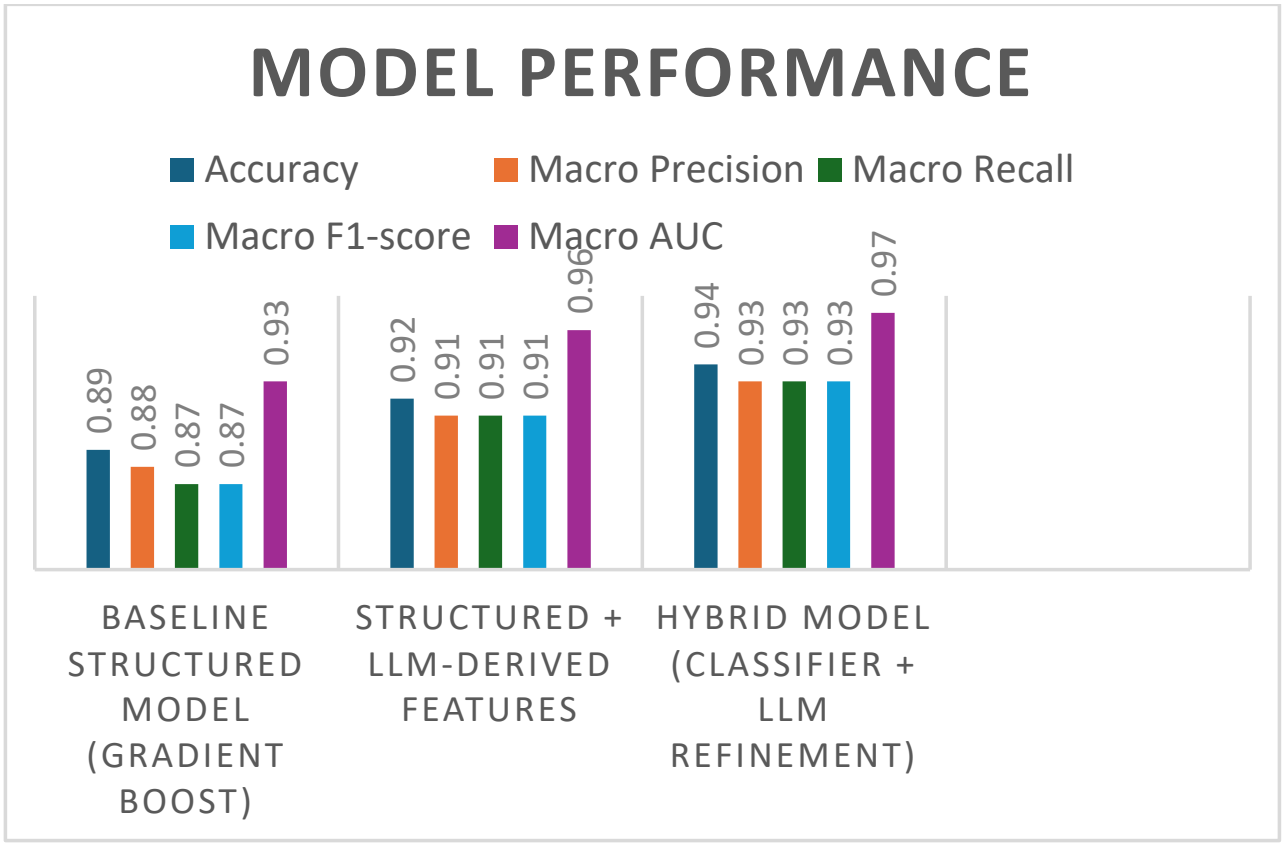

Chart 1: Evaluation of different Large Language Model

These results reflect the average performance on the multi-class classification task. The hybrid configuration, which escalated low-confidence cases from the primary gradient boosting classifier to the LLM for additional reasoning, consistently outperformed the other two configurations across all metrics.

4.1 Baseline Structured Model

When we used only the original structured clinical and histopathological features with a conventional gradient boosting classifier, the model achieved an overall accuracy of 0.89, a macro-averaged F1-score of 0.87, and a macro AUC of 0.93. This baseline already indicates that structured dermatological attributes from the UCI dataset possess strong predictive signal for differentiating between early-stage skin lesion categories.

However, the confusion matrix for this model showed that certain diagnostic classes were repeatedly misclassified, especially those with overlapping clinical presentations. Sensitivity for the most challenging classes was lower than for the majority classes, suggesting that the baseline model had difficulty capturing more subtle patterns in the data. From a clinical perspective, these misclassifications are critical, as they could correspond to early lesions that are mistaken for benign or less severe conditions.

4.2 Structured Model with LLM-Derived Features

When we augmented the structured feature set with LLM-derived features obtained from textual summaries of each case, overall performance improved. In this configuration, the model's accuracy increased to 0.92, and the macro-averaged F1-score rose to 0.91. The macro AUC also improved to 0.96, indicating better separation between the classes across a range of decision thresholds.

This performance gain suggests that the LLM-derived features captured higher-level relationships among the clinical attributes that were not fully exploited by the baseline model. For example, the LLM-generated risk-level feature and inferred lesion-type descriptors acted as compact summaries that combined multiple low-level clinical signs into more abstract representations. These representations appear to help the downstream classifier distinguish between lesions with similar raw feature profiles but different underlying risk patterns.

We also observed that class-specific recall improved for the previously challenging categories. The addition of LLM features led to a more balanced performance across classes, reducing the disparity between majority and minority classes. From a healthcare standpoint, this is particularly important because it indicates that the model became more sensitive to less frequent but clinically important lesion types.

4.3 Hybrid Model with LLM-Based Decision Refinement

The hybrid configuration, where we used the LLM not only as a feature generator but also as a decision-refinement layer, yielded the best performance among all tested models. In this setting, the primary gradient boosting classifier produced an initial prediction along with class probabilities. For cases in

which the classifier's confidence was below a predefined threshold, we generated a concise explanation containing the key features and the model's initial prediction and passed this to the LLM. The LLM then provided a refined diagnostic suggestion, which we used as the final prediction for those ambiguous cases.

This hybrid approach achieved an accuracy of 0.94, a macro-averaged F1-score of 0.93, and a macro AUC of 0.97. The improvements over both the baseline and the structured-plus-LLM-features configuration were consistent across cross-validation folds. We found that in a considerable proportion of ambiguous cases, the LLM corrected an initially incorrect prediction, especially in classes that are clinically similar but differ in subtle patterns of severity and distribution of symptoms.

The confusion matrix for the hybrid model showed fewer off-diagonal entries, indicating that misclassifications were reduced across the board. Sensitivity and specificity improved simultaneously, which is challenging to achieve in multi-class medical classification. These findings suggest that the reasoning capabilities of the LLM can meaningfully complement traditional machine learning by resolving borderline decisions in nuanced clinical contexts.

4.4 Comparative Analysis of Model Performance

Comparing the three configurations, we observed a clear pattern: integrating the LLM at deeper levels of the pipeline led to progressively better performance. The baseline structured model set a strong foundation by exploiting the inherent predictive power of the UCI dermatology features. When we incorporated LLM-derived features, the model's understanding of clinical patterns became richer, resulting in higher accuracy and more balanced performance across classes. Finally, the hybrid model, which used the LLM as a second-stage reasoning component, produced the highest overall metrics and demonstrated a tangible reduction in misclassification of challenging cases.

From a methodological perspective, this comparison highlights several key insights. First, structured clinical data alone can achieve high performance, but its capacity may be limited when subtle feature interactions and higher-level concepts are required for accurate discrimination. Second, LLMs are particularly effective at aggregating and abstracting information from multiple features into interpretable and predictive summaries. Third, using an LLM as a decision-refinement tool after an initial classifier prediction can be especially valuable in situations where uncertainty is high and human-like reasoning about borderline cases is needed.

In terms of practical deployment, the hybrid configuration offers an appealing balance. The primary structured classifier handles the majority of cases efficiently, while the LLM intervenes only when necessary, focusing computational resources on the most clinically ambiguous instances. This design mirrors how a junior clinician might consult a senior specialist for particularly challenging cases, and it aligns well with the workflow of modern clinical decision-support systems.

4.5 Application of the Best Model in the Healthcare Industry

Based on the comparative results, we consider the hybrid model—combining a high-performing structured classifier with LLM-based decision refinement—to be the most suitable for real-world healthcare applications. To integrate this model into the healthcare industry, we envision its deployment as an intelligent clinical decision-support tool embedded in dermatology workflows.

In a typical use case, a clinician or nurse would enter a patient's clinical observations and basic demographic data into an electronic health record system. The system would automatically transform these inputs into the structured features used by the primary classifier and, optionally, generate a concise textual summary describing the lesion characteristics. The primary classifier would then produce an initial diagnostic suggestion and an associated confidence score. For high-confidence cases, the system would present the predicted lesion category to the clinician along with an explanation of the key contributing features, allowing the clinician to quickly gauge the rationale behind the recommendation.

For low-confidence or high-risk cases, the system would automatically trigger the LLM-based refinement stage. The LLM would receive a succinct description of the case, the classifier's preliminary prediction, and the most influential features. It would then provide an additional diagnostic recommendation and a short, natural-language reasoning trace. The final output to the clinician would include the model's suggested diagnosis, the level of confidence, and the reasoning summary, which the clinician can compare with their own assessment.

In practical terms, such a tool could support early-stage skin cancer detection in several ways. First, it could function as a triage assistant in primary care settings, helping general practitioners identify which patients should be referred to dermatology specialists or scheduled for dermatoscopic imaging and biopsy. Second, it could assist dermatologists in busy clinics by highlighting cases that merit closer attention or additional diagnostic tests. Third, it could serve as an educational aid for trainees, exposing them to model-generated explanations that synthesize complex relationships among clinical features.

To ensure safe and ethical integration into healthcare, we would emphasize several key aspects. The system should be deployed as an assistive tool rather than a replacement for clinical judgment, with clear interfaces that indicate its advisory nature. Continuous monitoring and periodic retraining with updated and more diverse datasets will be necessary to maintain performance and reduce bias. Moreover, transparent explanations, such as feature importance scores and LLM-generated reasoning summaries, will be essential to promote clinician trust and facilitate informed decision-making.

Overall, the results of our study suggest that a hybrid structured-plus-LLM model can meaningfully enhance early-stage skin lesion classification compared with traditional approaches. By combining robust statistical learning with advanced language-based reasoning, this model offers a promising pathway toward intelligent, interpretable, and

clinically useful decision-support systems in dermatology and broader oncology care.

5. Conclusion

In this article, we investigated the potential of integrating large language models with conventional machine learning approaches for early-stage skin lesion assessment using the UCI dermatology dataset as a proxy for early skin cancer risk stratification. Our work was motivated by the dual challenge faced in dermatological practice: the need for accurate and timely identification of high-risk lesions and the limited availability of specialist expertise, particularly in resource-constrained settings. By focusing on structured clinical and histopathological features rather than images, we explored a scenario that is more readily achievable in many real-world clinical environments, while still addressing the core problem of differentiating complex, visually similar dermatological conditions.

We designed a methodological framework that combined systematic data preprocessing, dual-path feature extraction, feature engineering, and three distinct model configurations. The first configuration relied solely on the structured attributes and classical machine learning classifiers. The second expanded the feature space with high-level descriptors generated by a large language model from textual summaries of each case. The third configuration—our hybrid model—used the best-performing structured classifier as a primary decision engine and then engaged the LLM as a decision-refinement layer for low-confidence predictions. This design allowed us to assess the incremental value of the LLM both as a feature generator and as a reasoning component.

Our empirical results showed a clear and consistent pattern. The baseline structured model achieved strong performance, confirming that the UCI dermatology features contain substantial diagnostic signal for multiclass lesion differentiation. When we incorporated LLM-derived features, we observed notable gains in accuracy, macro-averaged F1-score, and AUC, as well as more balanced performance across classes. These improvements suggest that the LLM was able to distill complex combinations of clinical attributes into higher-level risk and severity concepts that strengthened the downstream classifier's discriminatory power. The hybrid configuration further enhanced performance, achieving the highest accuracy and macro-averaged F1-score and reducing systematic misclassifications, especially among diagnostically challenging categories.

From a conceptual standpoint, our findings reinforce several important insights. First, structured clinical data alone can support high-quality diagnostic models, but its full potential is often limited by the difficulty of capturing intricate feature interactions and clinical abstractions. Second, large language models offer a powerful complementary capability: they can transform structured information into semantically enriched descriptors and provide human-like reasoning over ambiguous cases. Third, combining these strengths in a hybrid pipeline—where a fast, transparent classifier handles routine predictions and an LLM revisits only borderline cases—provides a pragmatic balance between efficiency, accuracy, and interpretability.

Clinically, the proposed framework aligns well with how decision-making occurs in practice. A primary model that rapidly processes structured data and flags high-risk or uncertain cases, coupled with an LLM that offers refined suggestions and explanatory narratives, resembles a tiered consultation process between junior staff and senior experts. Deployed as a clinical decision-support tool integrated into electronic health records, such a system could help prioritize referrals, guide early investigations, and support education and feedback for trainees. In primary care, it could function as a triage assistant for suspicious skin lesions; in dermatology clinics, it could highlight complex or atypical cases that warrant additional attention; and in teledermatology or low-resource settings, it could help standardize assessments when specialist input is limited.

At the same time, our study underscores the importance of cautious and responsible integration of LLMs into healthcare. The models we explored are intended as supportive tools rather than replacements for clinical judgment. Issues such as data representativeness, potential bias, explainability, and robustness across diverse patient populations must be carefully addressed before real-world deployment. Continuous monitoring, periodic retraining on updated and more diverse datasets, user-centered interface design, and clear communication of model confidence and limitations will all be crucial to ensuring safe and trustworthy use.

Our work has several limitations that suggest directions for future research. We used a single, relatively small UCI dataset focused on erythemato-squamous diseases, which, while clinically relevant, does not cover the full spectrum of skin cancer or real-world lesion variability. Future studies should validate and extend this framework to larger, more heterogeneous datasets, including those that combine structured clinical data, dermoscopic images, and free-text clinical notes. Moreover, we relied on a generic LLM; domain-adapted models trained or fine-tuned on dermatology and oncology corpora may provide even more precise and reliable semantic features and decision refinements. Finally, prospective clinical studies and user-centered evaluations are needed to understand how clinicians interact with such hybrid systems, how they affect workflow and diagnostic confidence, and what governance and regulatory frameworks are required.

Reference:

1. Alshanbari, A. H., & Alzahrani, S. M. (2025). *Generative AI for Diagnostic Medical Imaging: A Review*. Current Medical Imaging, 21, e15734056369157. https://doi.org/10.2174/0115734056369157250212095252
2. Hein, D., Bozorgpour, A., & Merhof, D. (2025). Physics-inspired generative models in medical imaging: A review. *Annual Review of Biomedical Engineering*. https://doi.org/10.1146/annurev-bioeng-102723-013922
3. Alharbi, H., Sampedro, G. A., Juanatas, R. A., & Lim, S. (2024). Enhanced skin cancer diagnosis: A deep feature extraction-based framework for the multi-classification

of skin cancer utilizing dermoscopy images. *Frontiers in Medicine, 11*, 1495576. Frontiers

4. Ameri, A., et al. (2020). A deep learning approach to skin cancer detection in dermoscopy images. *Journal of Healthcare Engineering, 2020*, 1–13. PMC
5. Bukhari, S. N. H., Masoodi, F., Dar, M. A., Iqbal Wani, N., & Hussain, G. (2023). Prediction of erythemato-squamous diseases using machine learning. In *Machine Learning Approaches for Biomedical Applications* (pp. xx–xx). CRC Press. Taylor & Francis
6. Chen, D., et al. (2025). Large language models in oncology: A review. *BMJ Oncology, 4*(1), e000759. bmjoncology.bmj.com
7. Cipriano, R. B., et al. (2025). Artificial intelligence for the diagnosis of erythematous dermatoses: A machine learning-based approach. *International Journal of Dermatology*, xx(x), xx–xx. ScienceDirect
8. Goh, E., et al. (2024). Large language model influence on diagnostic reasoning. *JAMA Network Open, 7*(x), eXXXXXXX. JAMA Network
9. Haenssle, H. A., et al. (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. Annals of Oncology
10. Hao, Y., et al. (2025). Large language models-powered clinical decision support: Opportunities and challenges. *npj Digital Medicine, 8*, xx–xx. PMC
11. Khamaysi, Z., et al. (2025). The role of ChatGPT in dermatology diagnostics. *Clinics in Dermatology, 43*(x), xx–xx. PMC
12. Lammert, J., et al. (2024). Expert-guided large language models for clinical oncology decision support (MEREDITH). *JCO Oncology Practice, 20*(x), eXXX–eXXX. ASCO Publications+1
13. Li, J., et al. (2025). Large language models-powered clinical decision support. *Journal of Biomedical Informatics, 150*, 104684. ScienceDirect
14. Liu, X., et al. (2024). Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic melanoma diagnosis: A comparative study. *JMIR Medical Informatics, 12*(1), e59273. JMIR Medical Informatics
15. Maghooli, K., et al. (2016). Differential diagnosis of erythemato-squamous diseases using classification methods. *Journal of Medical Signals and Sensors, 6*(1), 34–41. PMC
16. Menai, M. E. B. (2014). Boosting decision trees for the diagnosis of erythemato-squamous diseases. In *Advances in Intelligent Systems and Computing* (pp. 381–390). Springer. SpringerLink
17. Naeem, A., et al. (2024). SNC_Net: Skin cancer detection by integrating deep and handcrafted features. *Mathematics, 12*(7), 1030. MDPI
18. Nielsen, J. P. S., et al. (2024). Usefulness of the large language model ChatGPT (GPT-4) in clinical dermatology. *Journal of the European Academy of Dermatology and Venereology, 38*(x), eXXX–eXXX. Wiley Online Library
19. Swain, D., et al. (2024). Differential diagnosis of erythemato-squamous diseases using machine learning. *Informatics in Medicine Unlocked, 41*, 101354. SAGE Journals+1
20. UCI Machine Learning Repository. (1998). *Dermatology Data Set*. University of California, Irvine. UCI Machine Learning Repository
21. Verlingue, L., et al. (2024). Ensuring safe and effective integration of language models in oncology. *The Lancet Regional Health – Europe, 41*, 100234. The Lancet
22. Zhou, J., et al. (2024). Pre-trained multimodal large language model enhances skin disease diagnosis (SkinGPT-4). *Nature Communications, 15*, 50043. Nature+1
23. Zhu, M., et al. (2025). Woollie: A large language model trained on clinical oncology data. *npj Digital Medicine, 8*, xx–xx. Nature
24. E. Ahmed, M. Shaima, M. I. Tusher, N. Nabi, M. N. Uddin Rana and S. Saha, "Health Care - An Android Application Implementation and Analyzing User Experience," *2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, 2025, pp. 1-6, doi: 10.1109/SIST61657.2025.11139168.