https://doi.org/10.51885/3134-8025_JICS_2026_1_7
SRSTI 28.23.01

MULTIMODAL CLASSIFICATION OF SKIN NEOPLASMS BY USING DEEP LEARNING

ТЕРЕҢ ОҚЫТУДЫ ҚОЛДАНА ОТЫРЫП, ТЕРІ ӨСІНДІЛЕРІНІҢ МУЛЬТИМОДАЛЬДЫ ЖІКТЕЛУІ

МУЛЬТИМОДАЛЬНАЯ КЛАССИФИКАЦИЯ КОЖНЫХ НОВООБРАЗОВАНИЙ С ИСПОЛЬЗОВАНИЕМ ГЛУБОКОГО ОБУЧЕНИЯ

T. Sadvakassov [1]*, **G. Soltan** [1], **T. Karibekov** [1], **S. Abdrakhmanova** [1]
[1]Astana IT University, Astana, Kazakhstan
*Corresponding author: Temirlan Sadvakassov, e-mail: imtims.17@gmail.com

Keywords:
machine learning, classification of skin diseases, convolutional neural networks, transfer learning, deep learning, medical diagnostics.

ABSTRACT

The article discusses modern machine learning methods which are applicable to the task of automated classification of skin diseases based on dermoscopic images. Special attention is paid to convolutional neural network architectures and the use of transfer learning with pre-trained models. The publicly available HAM10000 dataset, which includes images of seven types of skin lesions as well as clinical metadata on patients (gender, age, and location), was used as the experimental basis. Several models were implemented and tested during the study. A basic convolutional neural network (CNN) trained from scratch showed limited results, achieving an accuracy of around 39%. The improved CNN model with class balancing provided higher accuracy (73%), but still had limitations when classifying rare and visually similar categories. The MobileNetV2 model, which uses transfer learning and clinical metadata integration, demonstrated the best performance. A test accuracy of 81% was achieved, while recall for melanoma increased from 0.38 in the baseline CNN to 0.60, significantly reducing the likelihood of missing the most dangerous disease.

The Grad-CAM method was used to interpret the decisions, allowing us to visualize the model's areas of attention and identify the causes of errors. The results obtained confirm the promise of deep learning in decision support tasks in dermatology and highlight the need for further clinical validation and expansion of the database.

Түйінді сөздер:
машиналық оқыту, тері ауруларының жіктелуі, конволюциялық нейрондық желілер,

ТҮЙІНДЕМЕ

Мақалада дермоскопиялық кескіндер бойынша тері ауруларын автоматтандырылған жіктеу міндетіне қолданылатын машиналық оқытудың заманауи әдістері қарастырылады. Конвульсиялық нейрондық желілердің архитектурасына және алдын ала дайындалған

трансферлік оқыту, терең оқыту, медициналық диагностика.

модельдермен трансферлік оқытуды қолдануға ерекше назар аударылады. Эксперименттік база ретінде HAM10000 жалпыға қол жетімді деректер жиынтығы пайдаланылды, оның ішінде терінің жеті түрінің суреттері, сондай-ақ пациенттердің клиникалық метадеректері (жынысы, жасы және локализациясы) бар.
Зерттеу барысында бірнеше модельдер енгізіліп, сыналды. Нөлден үйретілген негізгі конволюциялық нейрондық желі (CNN) шектеулі нәтижелерге қол жеткізіп, шамамен 39 % дәлдікке қол жеткізді. Жақсартылған CNN сыныпты теңдестіру моделі жоғары дәлдікті қамтамасыз етті (73 %), бірақ сирек кездесетін және визуалды ұқсас санаттарды жіктеуде әлі де шектеулер болды. Трансферлік оқыту мен клиникалық метадеректерді біріктіруді қолданатын MobileNetV2 моделі ең жақсы көрсеткіштерді көрсетті. 81 % сынақ дәлдігіне қол жеткізілді, меланома үшін толықтығы (recall) негізгі CNN-де 0.38-ден 0.60-қа дейін өсті, бұл ең қауіпті ауруды өткізіп жіберу мүмкіндігін айтарлықтай төмендетеді.
Шешімдерді түсіндіру үшін Grad-Cam әдісі қолданылды, бұл модельдің назар аудару аймақтарын визуализациялауға және қателіктердің себептерін анықтауға мүмкіндік берді. Нәтижелер дерматологиядағы шешім қабылдауды қолдау міндеттерінде терең оқытуды қолдану перспективасын растайды және одан әрі клиникалық валидация мен мәліметтер базасын кеңейту қажеттілігін көрсетеді.

Ключевые слова:
машинное обучение, классификация кожных заболеваний, сверточные нейронные сети, трансферное обучение, глубокое обучение, медицинская диагностика.

АННОТАЦИЯ

В статье рассматриваются современные методы машинного обучения, применимые к задаче автоматизированной классификации кожных заболеваний по дермоскопическим изображениям. Особое внимание уделено архитектурам свёрточных нейронных сетей и использованию трансферного обучения с предобученными моделями. В качестве экспериментальной базы использован общедоступный набор данных HAM10000, включающий изображения семи типов кожных новообразований, а также клинические метаданные пациентов (пол, возраст и локализация).
В ходе исследования реализованы и протестированы несколько моделей. Базовая свёрточная нейронная сеть (CNN), обученная с нуля, показала ограниченные результаты, достигая точности около 39 %. Улучшенная модель CNN с балансировкой классов обеспечила более высокую точность (73%), но по-прежнему имела ограничения при классификации редких и визуально схожих категорий. Наилучшие показатели продемонстрировала модель MobileNetV2, использующая трансферное обучение и интеграцию клинических метаданных. Достигнута тестовая точность 81 %, при этом полнота (recall) для меланомы выросла с 0,38 в базовой CNN до 0.60, что существенно снижает вероятность пропуска наиболее опасного заболевания.
Для интерпретации решений применялась методика Grad-CAM, позволившая визуализировать области внимания модели и выявить причины ошибок. Полученные результаты подтверждают перспективность применения глубокого обучения в задачах поддержки принятия решений в дерматологии и подчёркивают необходимость дальнейшей клинической валидации и расширения базы данных.

INTRODUCTION

Skin and subcutaneous diseases remain a significant medical and social problem worldwide. According to data from the Global Burden of Disease (GBD-2019), approximately 4.86 billion new cases of skin disease are registered annually, placing them in 7th place among global causes of temporary disability. Melanoma poses a particular danger — it is one of the most aggressive forms of skin cancer, accounting for only a small percentage of all dermatological cases, but characterized by a high mortality rate. According to the World Health Organization, more than 325 000 new cases of melanoma are diagnosed worldwide each year, with more than 57 000 deaths related to this disease. Early diagnosis significantly increases survival rates: in stages I–II, the five-year survival rate exceeds 90%, whereas late detection leads to a sharp decline.

In Kazakhstan, diseases of the skin and subcutaneous tissue traditionally occupy a high position in the structure of morbidity among the population, ranking fifth among all classes of diseases. The urgency of the problem is exacerbated by a shortage of dermatologists, especially in rural areas, as well as an increase in the number of consultations via telemedicine services. With limited access to experts and a heavy burden on the healthcare system, there is a need for tools that can provide rapid, standardized, and reproducible assessment of skin lesion images (Aitkazinova et al., 2024).

Traditional diagnostic methods include visual examination and dermatoscopy. Despite their effectiveness when performed by a highly skilled physician, such methods are subject to subjective errors and depend on the specialist's experience, the quality of the equipment, and lighting conditions. This leads to the risk of early stages of melanoma not being detected, especially at the primary health care level. In these conditions, the digitization of healthcare and the development of machine learning technologies create the conditions for the introduction of decision support systems. Modern mobile device cameras and accumulated databases of dermatoscopic images open up the possibility of developing algorithms that can act as a "second reader" and help doctors identify suspicious cases for further examination.

In recent years, machine learning, and especially convolutional neural networks (CNNs), has become the de facto standard in medical image analysis. In radiology, ophthalmology and pathology, CNN-based systems have already demonstrated expert-level performance in classification, segmentation and detection tasks (Haggenmüller et al., 2021; Wu et al., 2022; Esteva et al., 2019). According to Esteva et al. (2019), CNNs have become a standard tool in medical imaging because they can automatically extract complex spatial features without hand-crafted descriptors. In dermatology, multiple studies report that CNNs can reach or even match the diagnostic accuracy of expert dermatologists under controlled conditions, making them a natural candidate for automated triage and decision support (Cerminara et al., 2023; Rahkonen et al., 2019). At the same time, the translation of these results into routine clinical practice is still limited by data imbalance, heterogeneity of imaging protocols and the need for interpretable, well-calibrated predictions.

One of the key tasks in developing such systems is ensuring the reliability and reproducibility of results when working with images of varying quality and origin.

One of the most popular and accessible benchmark datasets for research is HAM10000, which includes 10,015 images of seven types of skin lesions. The dataset is widely used for testing architectures and learning protocols and has become the de facto standard for multi-class classification tasks. However, it has some limitations, including a significant class imbalance (more than 67% of images are related to nevus), the presence of duplicate images, and partial histological verification of diagnoses (Cassidy et al., 2022; Aydin et al., 2023). These factors require the use of additional techniques—augmentation, class balancing, specialized loss functions, and careful validation at the patient level. In addition, testing models on clinical images obtained in real-world conditions, rather than just in standardized imaging conditions, is becoming an important area of focus.

Recent studies show that the best results are achieved when using pre-trained models with transfer learning. Architectures such as ResNet, EfficientNet, and MobileNet provide higher accuracy and stability compared to training from scratch (Manole et al., 2024; Venugopal et al., 2023).In addition to architecture, the choice of model fine-tuning strategy, including the use of cyclic learning rate changes, regularization, and overfitting control methods, is of particular importance. At the same time, increasing attention is being paid to the integration of clinical metadata (gender, age, location), which allows the model to be brought closer to real-world conditions. This multimodal approach enhances diagnostic accuracy and model robustness, especially in the presence of rare or atypical lesions. All the same, explainable AI is also developing: attention visualization methods (Grad-CAM) help identify which features are taken into account by the model, thereby increasing doctors' confidence (van der Velden et al., 2022; Borys et al., 2023). Methods such as Integrated Gradients and Score-CAM are also used to provide a deeper understanding of the model's logic when making decisions, especially in situations where an explanation is needed as to why a particular case is classified as suspicious. (Hauser et al., 2022). Approaches to combining heterogeneous data sources can be identified. Multimodal architectures that use cross-attention mechanisms allow for the coordinated consideration of images and accompanying clinical information, improving the model's generalizability (Le et al., 2025).

Thus, despite significant progress in the field of computer dermatology, a research gap remains: most studies focus solely on images and rarely include clinical context or interpretability assessments. This study aims to fill this gap by developing and comparing CNN and MobileNetV2 models with the integration of clinical data and the use of explainable AI methods, making its results more relevant to real-world clinical practice.

One of the key technological strategies is transfer learning, the use of models pre-trained on large general collections (e.g., ImageNet) followed by further training on specialized dermatological data. This approach accelerates convergence, reduces the requirements for the size of the marked sample, and increases the generalizability. Modern architectures (ResNet, Inception, EfficientNet) provide a flexible balance between accuracy and computational cost, while regularization and augmentation techniques help combat overfitting. In recent years, particular attention has been paid to the EfficientNet architecture, which demonstrates high accuracy with fewer parameters, making it particularly suitable for telemedicine solutions (Manole et al., 2024). It is also worth noting MobileNetV2, which is actively used in applications for screening skin diseases, thanks to its compactness and high performance (Nirupama et al., 2024). The calibration of probabilistic model inferences plays a particularly important role. These methods help to improve the reliability of predictions and avoid underestimating or overestimating risk, which is particularly important in the early detection of melanoma.

The advantages of ML in this context are high speed, reproducibility of results, and technical suitability for integration into clinical systems (web interfaces, PACS/EMR modules, mobile applications) provided that probabilities are correctly calibrated and risk management is well thought out. Modern solutions are increasingly being developed with the possibility of integration into telemedicine platforms in mind, providing remote access to diagnostics and second opinions from doctors.

A separate area is the development of explainable AI (XAI), which is particularly relevant in dermatology. Grad-CAM, Integrated Gradients, and Score-CAM methods allow visualizing which areas of the image the model should pay attention to, thereby increasing doctors' confidence and facilitating the interpretation of results (Hauser et al., 2022). The integration of such tools has become particularly important in situations where AI is used as an element of clinical decision-making rather than simply as an auxiliary analysis tool.

Recent systematic reviews highlight a gradual shift from experimental “reader” studies towards validated clinical deployments of deep learning in dermatology. While convolutional neural networks show promising sensitivity in detecting malignant lesions, their real-world performance strongly depends on image acquisition protocols, patient pathways and decision thresholds, and therefore requires prospective evaluation and robustness assessment across diverse clinical sites [1–4]. Work performed in real-world settings (total body photography, in-clinic dermatoscopy) underlines both the potential and the limitations of augmented intelligence in dermatology: CNNs can improve early detection, but specificity and overall reliability remain highly context-dependent.

Data and metrics. The open dataset HAM10000 was used as the experimental basis, containing 10,015 dermatoscopic im ages of seven types of skin neoplasms: actinic keratosis and intradermal carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), nevus (nv), and vascular lesions (vasc). The dataset is widely used in research from 2019 to 2025 and is effectively the standard for multi-class classification of dermatoscopic images.

When using HAM10000, there is one key threat to validity. This is a pronounced imbalance between classes: more than 67% of images relate to nevus, while categories such as dermatofibroma or actinic keratosis are represented much less frequently. This leads to a shift in the model towards the dominant class and reduces recall for rare classes. Recent studies recommend using stratified or patient/lesion-level divisions, as well as techniques to combat imbalance: augmentation, class weighting, and specialized loss functions. In our study, stratification and class weight adjustment were applied, which helped to mitigate imbalances and increase the stability of the results.

Standard metrics were used to evaluate the quality of classification: accuracy, precision, recall, and F1 score. Additionally, a confusion matrix and classification report were constructed, which made it possible to identify the most problematic classes and analyze the nature of the model's errors.

Transfer learning as the de facto standard. Comparative studies confirm the consistent superiority of fine-tuning pre-trained ImageNet models (ResNet, DenseNet, Inception, EfficientNet) over training from scratch on moderately sized medical datasets. This provides better generalization ability and faster convergence; for HAM10000, the best "quality/resources" trade-offs are often achieved by the EfficientNet (B0–B5/B7) [5,9–11]. At the same time, transformer and hybrid (CNN+Transformer) variants are being investigated, which improve global context extraction and, with correct fine-tuning and regularization, provide an increase in ROC-AUC/macro-F1 [12–14]. Transfer learning has effectively become the de facto standard for HAM10000. Comparative studies consistently report overall accuracies in the mid-80% to mid-90% range when ImageNet-pretrained backbones are fine-tuned on dermoscopic images. For example, Aydin et al. achieved an accuracy of 96.5% and F1-score of 0.96 on HAM10000 using color histogram-based local descriptors combined with an XGBoost classifier (Aydin, 2023). Alam et al. reported that their RegNetY-320 model, trained with extensive augmentation, reached 91% accuracy, F1-score of 0.88 and ROC-AUC of 0.95 on the same dataset (Alam, 2022). Tajerian et al. fine-tuned an EfficientNet-B1 architecture and obtained 84.3% overall accuracy on HAM10000, with class-wise F1-scores ranging from 0.54 to 0.93 and melanoma F1-score of 0.58, highlighting the persistent difficulty of minority classes (Tajerian, 2023). A recent survey further summarizes that EfficientNet-B4 trained on HAM10000 can reach 87.9% accuracy (Ali et al., 2022), while transformer-based models such as SkinTrans/ViT achieve up to 94.3% accuracy on the same benchmark (Manole, 2024).

Taken together, these results indicate that modern transfer-learning pipelines with EfficientNet-like or transformer backbones define a strong state-of-the-art baseline for

HAM10000, but still leave room for improvement in the detection of under-represented lesions such as melanoma.

Working with imbalance and overall stability. The best results are demonstrated by combinations of targeted augmentation (random crop/flip, color jitter, CutMix/MixUp), class weights, or Focal-loss, as well as ensemble architectures. Individual studies show gains from combining EfficientNet with additional attention blocks and/or cascade ensembling; a correct validation procedure (k-fold with stratification at the lesion level) remains an important factor for stability [6–8,10,11].

Interpretability and clinical integration. Decision support systems require explainability and verification of the "plausibility" of explanations from a dermatologist's point of view. A review of XAI in medical imaging emphasizes that heat maps (Grad-CAM, etc.) must be supplemented with validation of the correctness of explanations and control of "provoked artifacts," otherwise the increase in model accuracy will not be accompanied by an increase in clinical confidence [15–17]. Publications in dermatology also draw attention to interfaces that enable the sharing of explanations between doctors and AI (human-AI teaming) [2,4].

Issues and challenges. Despite significant progress in the application of the deep learning approach to dermatological diagnosis tasks, there remain a number of unresolved issues that limit the practical implementation of such systems. One of the key obstacles is the limited and inconsistent labeling in open datasets. For example, in HAM10000, some diagnoses are histologically confirmed, while others are based on clinical judgment or expert consensus, which reduces the homogeneity of reference labels and may lead to bias in training results.

Another significant factor is the pronounced imbalance between classes: the predominance of images of nevuses and the lack of rare and clinically significant categories (such as melanoma or dermatofibroma) leads to a shift in the model toward "common" diagnoses and reduces the completeness of dangerous classes. To overcome this limitation, targeted augmentations, the use of specialized loss functions (Focal Loss), or synthetic data augmentation are required.

Additional problems remain due to differences in imaging devices, image acquisition protocols, patient populations, and the presence of artifacts (hair, glare, marks). These factors reduce the transferability of models trained on a single dataset to real-world clinical practice. The solution may be external and prospective validation using independent data sources, as well as verification of the calibration of probabilistic predictions.

Special attention should be paid to the interpretability of algorithms. Explainable AI methods, such as Grad-CAM, allow visualization of the model's areas of focus, but require critical review by experts to eliminate reliance on irrelevant features and increase trust in the system.

Finally, risk management and ethical compliance play an important role. Protecting personal data, ensuring fairness between different patient groups, and clearly documenting limitations and safe usage scenarios remain necessary conditions for integrating such technologies into medical practice.

Scientific novelty of the project. The scientific novelty of the project lies in the development of a multimodal approach to the automated classification of skin lesions, which combines the analysis of dermatoscopic images and clinical metadata (gender, age, location). This approach takes into account the real-world context of a dermatologist's work and improves the accuracy and completeness of classification compared to models that use only visual information.

Unlike most existing studies, this work directly compares a basic CNN trained from scratch with a pre-trained MobileNetV2 architecture using transfer learning. The multimodal version of MobileNetV2 demonstrated a significant increase in completeness for the melanoma class (from 0.38 to 0.60), which is of high clinical significance as it reduces the likelihood of missing the most dangerous type of tumors.

Additionally, scientific novelty is determined by the use of interpretable AI methods (Grad-CAM) to analyze model decisions. Visualization of areas of attention has made it possible not only to identify key features that influence classification, but also to show potential sources of error (e.g., focus on artifacts). This ensures a higher level of trust on the part of physicians and opens up opportunities for joint use of the system by specialists.

Finally, the paper critically examines the limitations of the HAM10000 dataset, including the heterogeneity of diagnosis verification methods, which is rarely taken into account in applied research. Thus, the project contributes to the development of medical decision support systems by offering a practice-oriented, interpretable, and clinically relevant approach to the diagnosis of skin diseases.

MATERIALS AND METHODS

As mentioned earlier, the HAM10000 (Human Against Machine with 10,000 training images) dataset, which is open and widely recognized in the scientific community, was used as the experimental basis for this work. This dataset is a collection of 10,015 dermatoscopic images covering seven different types of skin lesions and serves as an important resource for the automatic classification and diagnosis of dermatological diseases using machine learning and computer vision methods.

Each image in HAM10000 is accompanied by a label indicating the type of neoplasm, which allows it to be used as a training sample for building and evaluating classification models. The types of skin lesions represented in the dataset include both malignant and benign formations, providing the diversity and balance necessary for reliable algorithm training: actinic keratosis and intradermal carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), nevus (nv), and vascular lesions (vasc).

Below is a summary table showing the distribution of images by class. It shows the name of the class, its designation, the number of images belonging to that type, and the percentage of the total number of examples:

Table 1. Dataset distribution

Class	Designation	Number of images	Percentage of total
Actinic keratosis / carcinoma	akiec	327	3.3%
Basal cell carcinoma	bcc	514	5.1%
Benign keratosis-like lesions	bkl	1 099	11.0%
Dermatofibroma	df	115	1.1%
Melanoma	mel	1 113	11.1%
Nevus	nv	6 705	67.0%
Vascular lesions	vasc	142	1.4%
Total	-	10 015	100%
Note – compiled by the authors			

Clinical metadata is available for some images: gender, age, and lesion location. This data was integrated into the model as an additional source of information.

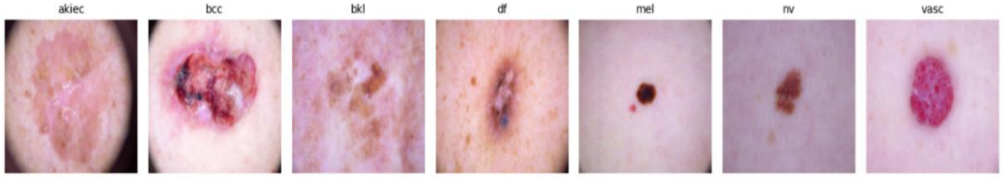


Figure 1. Examples of dermatoscopic images for each lesion class.
Note – compiled by the authors

During the preprocessing stage, images were scaled to 128×128 pixels and normalized to the range [0,1]. To reduce the risk of overfitting, standard augmentation techniques were used: horizontal and vertical reflections, random rotations, and brightness changes. To combat class imbalance, class weighting was applied during training, which compensated for the predominance of nevuses and increased completeness for rare categories.

Two groups of deep learning models were implemented. The baseline convolutional neural network (CNN) trained from scratch consisted of several convolutional blocks with 3×3 kernels (32 and 64 filters), each followed by BatchNormalization and MaxPooling2D layers, a Flatten operation and two fully connected layers with 64 units and ReLU activation. The final classification layer was a dense layer with 7 units and softmax activation, corresponding to the seven lesion classes.

In the multimodal variant, the image branch was complemented with a metadata branch that processed patient sex, age and lesion location. Categorical variables were one-hot encoded, concatenated with normalized age and passed through a dense layer with 64 units and ReLU activation to obtain a metadata embedding. The outputs of the image branch (64-dimensional feature vector) and the metadata branch (64-dimensional embedding) were concatenated into a 128-dimensional vector, followed by a dense layer with 64 units and ReLU activation and the final 7-way softmax classifier. For the MobileNetV2-based models, the convolutional backbone pre-trained on ImageNet was used as a fixed feature extractor, followed by GlobalAveragePooling2D and the same multimodal fusion block as described above.

Training was performed using the Adam optimizer with an initial learning rate of 0.001 and the sparse categorical cross-entropy loss function, while accuracy was used as the primary monitoring metric. The dataset was stratified by class and split into training and test subsets in an 80/20 ratio; within the training subset, 10–20% of the data were further reserved as a validation set by using the validation_split option in Keras. EarlyStopping (monitoring validation loss with a patience of 6 epochs) and ReduceLROnPlateau (factor 0.5, patience 4) callbacks were applied to prevent overfitting and stabilize convergence. Training was conducted for up to 30 epochs with a batch size of 32, using class weights inversely proportional to class frequencies to mitigate the strong class imbalance. Standard metrics (accuracy, precision, recall, and F1-score), as well as a confusion matrix and a per-class classification report, were used to evaluate model performance.

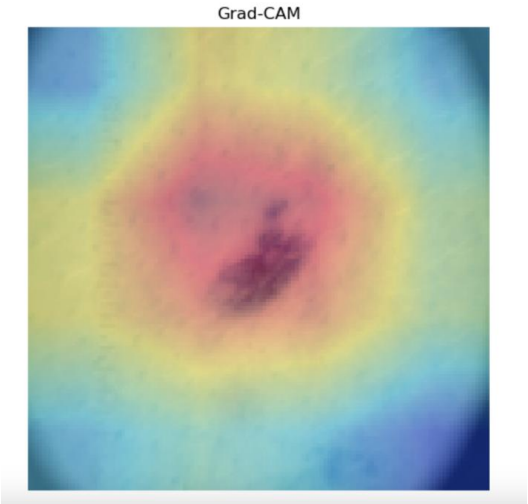


Figure 2. Grad-CAM visualization for a dermatoscopic image. The highlighted red region shows the areas most influential for the model's prediction

Note – compiled by the authors

RESULTS AND DISCUSSION

During the experiments, we evaluated several configurations belonging to two deep learning architecture families: a convolutional neural network (CNN) trained from scratch and a MobileNetV2 model with transfer learning. Within these two families, four main configurations were analyzed: (1) a simple baseline CNN trained on a small balanced subset of the HAM10000 dataset, (2) an improved CNN trained on the full dataset with class weighting and data augmentation, (3) a MobileNetV2-based model using only dermoscopic images, and (4) the final multimodal MobileNetV2 model that combines image features with clinical metadata (sex, age and lesion location).

The key characteristics and performance metrics of these four configurations are summarized in Table 2. The table reports test accuracy for each model and recall for the melanoma class, which is the most safety-critical outcome in this setting.

Table 2. Summary of evaluated models and their performance on the HAM10000 test set.

Model	Input data	Additional techniques	Test accuracy
Baseline CNN	Dermoscopic images, small balanced subset	No metadata, no class weighting	0.39
Improved CNN	Dermoscopic images, full HAM10000	Class weighting, data augmentation	0.73
MobileNetV2	Dermoscopic images, full HAM10000	Transfer learning, class weighting	0.53
Multimodal MobileNetV2	Images + metadata	Transfer learning, class weighting, metadata fusion	0.81
Note – compiled by the authors			

First, we analyzed the CNN-based configurations. The simplest baseline CNN, trained from scratch on a small balanced subset of the HAM10000 dataset, achieved only about 0.39 test accuracy. This result confirms that training a shallow architecture on limited data leads to underfitting and unstable generalization.

When the same CNN architecture was retrained on the full HAM10000 dataset with class weighting and data augmentation, the test accuracy increased to approximately 0.73. However, the confusion matrix still revealed systematic errors for rare categories such as dermatofibroma (df) and actinic keratoses (akiec), as well as persistent confusion between melanoma and other pigmented lesions. Thus, even the improved CNN trained from scratch remained limited when dealing with the heterogeneous and imbalanced real-world data.

Figure 3 shows the confusion matrix for the final multimodal MobileNetV2 model on the HAM10000 test set. Most errors arise between visually similar pigmented lesions such as nevus (nv), benign keratosis-like lesions (bkl) and melanoma (mel), as well as between bkl and actinic keratoses (akiec), which is consistent with previous work on this dataset.

For the MobileNetV2-based models, using transfer learning already provided a noticeable improvement compared to the CNN family. A MobileNetV2 model that used only dermoscopic images and class weighting reached a test accuracy of about 0.53, outperforming the baseline CNN trained on a small subset. However, the most substantial gain was obtained when clinical metadata were added as a second input branch. The multimodal MobileNetV2 model that combined image features with sex, age and lesion location achieved the best overall performance, with a test accuracy of 0.81. This improvement in sensitivity was obtained without a substantial drop in overall specificity, highlighting the benefits of using pre-trained weights together with additional clinical context.

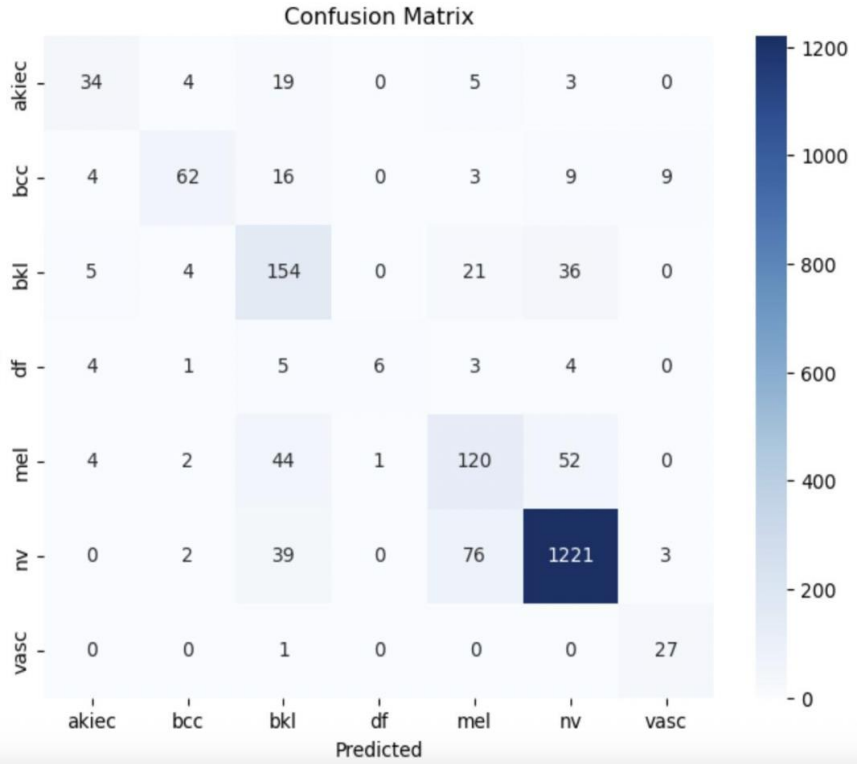


Figure 3. Confusion matrix for the multimodal MobileNetV2 model on the HAM10000 test set.

Note – compiled by the authors

	precision	recall	f1-score	support
akiec	0.70	0.68	0.69	65
bcc	0.69	0.87	0.77	103
bkl	0.70	0.68	0.69	220
df	0.89	0.70	0.78	23
mel	0.66	0.53	0.59	223
nv	0.92	0.94	0.93	1341
vasc	0.92	0.86	0.89	28
accuracy			0.85	2003
macro avg	0.78	0.75	0.76	2003
weighted avg	0.85	0.85	0.85	2003

Figure 4. Classification report for MobileNetV2 with integrated metadata, showing precision, recall, and F1-score across all seven lesion categories

Note – compiled by the authors

Overall, the experiments demonstrate a consistent improvement in performance as the models become more expressive and make better use of the available data (Table 2). Moving from a shallow CNN trained on a small balanced subset to an improved CNN on the full dataset yields a clear gain in accuracy, but the most clinically relevant improvements are achieved by combining transfer learning with multimodal fusion of clinical metadata in the MobileNetV2 family.

A comparison with previously published work on the HAM10000 dataset and related dermatoscopic benchmarks shows that the obtained results are competitive with other deep

learning approaches. Many earlier studies focused exclusively on image-based models, whereas our findings support the growing evidence that integrating visual and clinical information can improve both accuracy and sensitivity in skin cancer screening. At the same time, the remaining limitations highlighted in the confusion matrices and per-class metrics indicate the need for further work on imbalance-handling strategies and external validation before deploying such systems in routine clinical practice.

The use of Grad-CAM made it possible not only to interpret predictions, but also to identify potential sources of model errors. Analysis of heat maps showed that in cases of correct classification, the network focused on morphologically significant structures (e.g., melanoma pigment patterns), while in cases of error, attention shifted to artifacts (hair coverings, heterogeneous textures, glare). This coincides with the findings of studies on explainable AI (van der Velden et al., 2022; Borys et al., 2023), which emphasize the need to verify the "plausibility" of explanations in order to increase the trust of physicians.

However, the study has a number of limitations. First, the HAM10000 dataset is characterized by a pronounced class imbalance: more than 67% of images are non-nevi, while rare categories are extremely limited. Second, some of the diagnoses in the dataset are not histologically confirmed, which reduces the reliability of the reference labels (Cassidy et al., 2022). Third, the experiments did not include external validation on independent datasets (e.g., ISIC 2020, 2023), which is necessary to verify the model's transferability to new patient populations. In this work we did not investigate specialized loss functions such as Focal Loss or more aggressive synthetic augmentation strategies; these techniques are expected to further improve recall for rare classes and are therefore planned for future experiments.

Thus, the results of the study confirm the potential of deep learning for decision support tasks in dermatology, especially in the format of multimodal architectures. However, for practical implementation, it is necessary to solve the problems of class imbalance, clinical validation on external data, and the development of interfaces that allow doctors to jointly use model explanations.

CONCLUSION

This paper investigates deep learning methods for automated classification of skin lesions based on dermatoscopic images using the open-source HAM10000 dataset. A comparison of two approaches is implemented: a basic convolutional neural network (CNN) trained from scratch, and the MobileNetV2 architecture using transfer learning and integration of clinical metadata (gender, age, location).

The results showed that MobileNetV2 significantly outperforms the best CNN configuration in terms of overall accuracy (0.81 vs. 0.73) and recall for the most dangerous class, melanoma (0.60 vs. 0.38).

The use of Grad-CAM made it possible to analyze the model's decision-making process and demonstrated that the areas of focus coincided with clinically significant areas of the image, which increases doctors' confidence in the system. However, challenges remain related to class imbalance in the dataset, limited histological verification, and the need for external validation on independent datasets.

The practical value of this study lies in the fact that the proposed approach can be used to create decision support tools for dermatological practice. In particular, such systems can be integrated into telemedicine services and mobile applications for primary screening, which will reduce the burden on specialists and increase the availability of early diagnosis.

In the future, we plan to expand the experimental base by using more diverse datasets (e.g., ISIC 2020, 2023), investigating specialized loss functions such as Focal Loss to combat class imbalance, and introducing next-generation multimodal architectures, including transformer

models. This will create the conditions for the development of clinically reliable, interpretable, and scalable decision support systems in dermatology.

CONFLICT OF INTEREST: The authors declare no conflict of interest.

FUNDING: No funding was received for the preparation of this article.

ACKNOWLEDGEMENTS: The authors express their gratitude to colleagues for methodological support and helpful discussions, as well as to anonymous reviewers for valuable comments that helped improve the quality of the article.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES: No artificial intelligence (AI) technologies were used at any stage in the preparation of this article.

REFERENCES

Haggenmüller S., Maron R.C., Hekler A., et al. (2021). Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. European Journal of Cancer. 156, 202–216. https://doi.org/10.1016/j.ejca.2021.06.049

Cerminara S.E., Cheng P., Kostner L., et al. (2023). Diagnostic performance of augmented intelligence with 2D and 3D total body photography and convolutional neural networks in a high-risk population for melanoma under real-world conditions. European Journal of Cancer. 190, 112954. https://doi.org/10.1016/j.ejca.2023.112954

Wu Y., Xu H., Zhang J., et al. (2022). Skin Cancer Classification With Deep Learning: A Systematic Review. Frontiers in Oncology. 12, 893972. https://doi.org/10.3389/fonc.2022.893972

Jain A., et al. (2021). AI-Based Tool for Skin Condition Diagnosis in Primary Care. JAMA Network Open. 4, 217249. https://doi.org/10.1001/jamanetworkopen.2021.7249

Aydin Y., et al. (2023). A Comparative Analysis of Skin Cancer Detection Methods Using HAM10000. PLOS ONE. 18 (10), 0289870. https://doi.org/10.1371/journal.pone.0289870

Tajerian A., et al. (2023). Design and validation of a new machine-learning-based approach using HAM10000 dermoscopy images. PLOS ONE. 18 (5), 0284437. https://doi.org/10.1371/journal.pone.0284437

Alam T.M., et al. (2022). An Efficient Deep Learning-Based Skin Cancer Classifier (imbalanced HAM10000). Computational Intelligence and Neuroscience. 7801342. https://doi.org/10.1155/2022/7801342

Cassidy B., Kendrick C., Soyer H.P., et al. (2022). Analysis of the ISIC image datasets: usage, benchmarks and recommendations. Medical Image Analysis. 75, 102305. https://doi.org/10.1016/j.media.2021.102305

Manole I., et al. (2024). Enhancing Dermatological Diagnostics with EfficientNet. Bioengineering. 11 (8), 810. https://doi.org/10.3390/bioengineering11080810

Venugopal V., et al. (2023). A deep neural network using modified EfficientNet for skin lesion classification. Intelligent Medicine. 3(3), 100105. https://doi.org/10.1016/j.imed.2023.100105

Shetty B., et al. (2022). Skin lesion classification of dermoscopic images using deep learning. Scientific Reports. 12, 19224. https://doi.org/10.1038/s41598-022-22644-9

Xin C., et al. (2022). An improved transformer network for skin cancer classification. Computers in Biology and Medicine. 149, 105939. https://doi.org/10.1016/j.compbiomed.2022.105939

Yang G., Luo S., Greer P. (2023). A novel vision transformer model for skin cancer classification. Neural Processing Letters. https://doi.org/10.1007/s11063-023-11204-5

Himel G.M.S., et al. (2024) Skin Cancer Segmentation and Classification Using Vision Transformer. Healthcare. 12 (2), 139. https://doi.org/10.3390/healthcare12020139

van der Velden B.H.M., Kuijf H.J., Gilhuijs K.G.A., Viergever M.A. (2022). Explainable AI in deep learning-based medical image analysis. Medical Image Analysis. 79, 102470. https://doi.org/10.1016/j.media.2022.102470

Borys K., et al. (2023). Explainable AI in medical imaging: Beyond saliency-based approaches. European Journal of Radiology. 164, 110876. https://doi.org/10.1016/j.ejrad.2023.110876

Miller I., et al. (2024). Performance of commercial dermatoscopic systems with AI. Cancers. 16 (7), 1443. https://doi.org/10.3390/cancers16071443

Aitkazinova A., Saimova A., Urazayeva A. (2024). Epidemiological situation of skin cancer and melanoma in the Republic of Kazakhstan in 2012–2022. ResearchGate. https://doi.org/10.13140/RG.2.2.22583.70564

Zhou H., Wang Y., Li X., et al. (2024). Accurate Skin Lesion Classification Using Multimodal Learning on the HAM10000 Dataset. medRxiv. https://doi.org/10.1101/2024.05.30.24308213

Khan M., Ahmed S., Javed H., et al. (2023). A Novel Transfer Learning Framework for Multimodal Skin Lesion Analysis. Applied Sciences. 13 (21), 11974. https://doi.org/10.3390/app132111974

Rai R., Chawla A., Gupta P., et al. (2025). Comparative analysis of multimodal architectures for effective skin lesion classification. Frontiers in Artificial Intelligence. 8, 1608837. https://doi.org/10.3389/frai.2025.1608837

Hauser K., Kurz A., Meier F., et al. (2022). Explainable artificial intelligence in skin cancer recognition: A systematic review. Medical Image Analysis. 78, 102433. https://doi.org/10.1016/j.ejca.2022.02.025

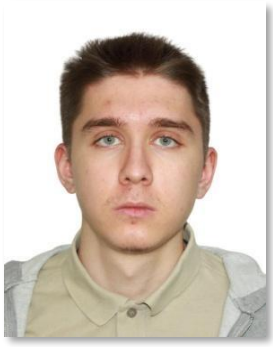
Tran-Van N.-Y. and Le K.-H. (2025). A multimodal skin lesion classification through cross-attention fusion and collaborative edge computing. Comput. Med. Imaging Graph. 124, 102588. https://doi.org/10.1016/j.compmedimag.2025.102588

Esteva A., Robicquet A., Ramsundar B., et al. (2019). A guide to deep learning in healthcare. Nat. Med. 25 (1), 24. https://doi.org/10.1038/s41591-018-0316-z

Rahkonen S., Neittaanmäki N., et al. (2019). Convolutional neural networks in skin cancer detection using spatial and spectral domain. Proc. Photonics in Dermatology and Plastic Surgery. https://doi.org/10.1117/12.2509871

Nirupama and Virupakshappa (2024). MobileNet-V2: An enhanced skin disease classification by attention and multi-scale features. J. Imaging Inform. Med. 38 (7). https://doi.org/10.1007/s10278-024-01271-y

Авторлар туралы мәліметтер
Информация об авторах
Information about authors



Садвакасов Темирлан – Компьютерлік ғылымдар және инженерия бағытындағы студент – магистрі, Astana IT University, Астана, Қазақстан

Садвакасов Темирлан – Студент магистратуры по направлению компьютерные науки и инженерия, Astana IT University, г. Астана, Казахстан

Sadvakassov Temirlan – Master's student of Computer Science and Engineering, Astana IT University, Astana, Kazakhstan,
e-mail: imtims.17@gmail.com,
ORCID: https://orcid.org/0009-0001-2423-1706,

Солтан Гульжан – техника ғылымдарының кандидаты, Компьютерлік инженерия мектебінің Associate Professor-і, Astana IT University, Астана қ., Қазақстан
Солтан Гульжан – кандидат технических наук, Associate Professor школы компьютерной инженерии, Astana IT University, г. Астана, Казахстан
Soltan Gulzhan – Candidate of Technical Sciences, Associate Professor, Department of Computer Engineering, Astana IT University, Astana, Kazakhstan,
e-mail: gulzhan.soltan@astanait.edu.kz,
ORCID: https://orcid.org/0000-0002-1603-7524,

Карибеков Темирлан – медицина ғылымының докторы, S&IC «MedTech» директоры, Astana IT University, Астана, Қазақстан
Карибеков Темирлан – доктор медицинских наук, директор S&IC «MedTech», Astana IT University, г. Астана, Казахстан
Karibekov Temirlan – Doctor of Medical Sciences, Director S&IC «MedTech», Astana IT University, Astana, Kazakhstan,
e-mail: t.karibekov@astanait.edu.kz,
ORCID: https://orcid.org/0009-0008-9801-1774,

Абдрахманова Саида – Компьютерлік ғылымдар және инженерия бағытындағы студент – магистрі, Astana IT University, Астана, Қазақстан
Абдрахманова Саида – Студент магистратуры по направлению компьютерные науки и инженерия, Astana IT University, г. Астана, Казахстан
Abdrakhmanova Saida – Master's student of Computer Science and Engineering, Astana IT University, Astana, Kazakhstan,
e-mail: abdrahmanovasaida48@gmail.com,
ORCID: https://orcid.org/0009-0008-2800-4704