AI system for melanoma recognition

1st Jakub Grelowski
Wrocław University of Science and Technology
Wroclaw, Poland
262754@student.pwr.edu.pl

2nd Kamil Kochan
Wrocław University of Science and Technology
Wroclaw, Poland
259141@student.pwr.edu.pl

3rd Maksym Malicki
Wrocław University of Science and Technology
Wroclaw, Poland
259216@student.pwr.edu.pl

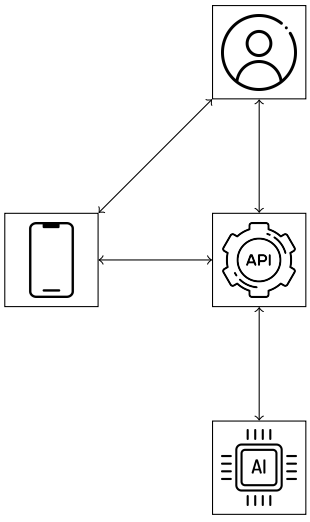
***Index Terms*—melanoma, artificial intelligence, computer vision, machine learning**

I. Introduction

The goal of this project is to create a tool that is able to recognize melanoma from pictures by using machine learning algorithms. The main scope is to prepare and train a computer vision model that is able to classify melanoma and distinguish it from benign birthmarks. As part of the project, we have investigated how different algorithms affect the accuracy and efficiency of the system.

A. Tools

To implement the project, we used Python with scikit-learn, XGBoost and PyTorch libraries. Moreover, created a mobile application for the detection input using Flutter. The communication between the client app and the model is carried out using a dedicated API with Python Fast API framework.



B. Data

To train our model, we gathered various datasets with labelled, professional videodermatoscopy images of malignant and benign melanoma cancer. The size of our datasets ranges from 3000 to 10000 instances with the equal classes sizes.

II. Review of the existing literature

A. The role of technology in melanoma screening and diagnosis

In the field of dermatologic oncology, recent research highlights a significant shift toward the use of improved technological methods for melanoma screening and diagnosis. This includes the integration of artificial intelligence and multispectral imaging techniques, which together aim to improve the accuracy and efficiency of early melanoma detection. The context of these advances is part of a drive to improve patient outcomes through more accurate and earlier diagnosis. Comprehensive reviews compare traditional diagnostic methods, such as visual inspections and biopsies performed by physicians, with these newer technologies, noting significant differences in scalability, objectivity and potential for early detection. However, existing approaches still face issues such as high false-positive rates and the availability of advanced diagnostic tools in less-resourced settings. A systematic review including nine studies that evaluated six mobile apps for assessing skin cancer risk found that evidence was poor and did not support clinical use, despite two apps having been awarded the CE [1]. The identification of these gaps not only underscores current technological limitations, but also the critical need for continued development in the field, as highlighted by research leading to more accessible and reliable diagnostic technologies.

B. Cutaneous Malignant Melanoma: A Review of Early Diagnosis and Management

One of the most aggressive cancers seen in humans is cutaneous melanoma (CM), a tumor formed from melanocytes [2]. Visual examination tools such as the dermatoscope play a key role in the diagnosis of cutaneous melanoma (CM). These diagnostic methods aim to increase the visual distinction between malignant and benign skin lesions, thereby facilitating early detection. In the context of the increasing incidence of melanoma reported worldwide, the ability to diagnose melanoma at an early stage using non-invasive technologies is becoming increasingly important. Prominent among these

is dermoscopic examination, especially when combined with sequential digital dermoscopy and whole-body photography, providing detailed visualisation of the skin surface to help identify suspicious lesions with greater accuracy than the naked eye. However, the publication reveals that despite advances, challenges such as underestimation of lesion thickness or misdiagnosis remain common, highlighting gaps in current methodologies. This highlights the continued need to improve visual diagnostic tools, pushing forward innovations to increase diagnostic accuracy and reduce melanoma mortality through earlier detection.

C. Human–computer collaboration for skin cancer recognition

The authors of this article focus on terms such as telemedicine, AI-based support and CBIR (Content-Based Image Retrieval), which are currently key to understanding the research methods used and their resulates. Due to the growing interest in the integration of AI systems into medical practice (not only in the field of dermatology), article positioned the literature review covering AI systems working autonomously as well as those cooperating with physicians. Authors also analyzed the impact of AI-based support in clinically relevant scenario [3]. They found that the collaborative approach between AI and physicians significantly improved diagnostic accuracy compared to either working alone. This highlights the potential of AI as a valuable tool in enhancing clinical decision-making processes. The analysis also showed that a large proportion are not ready to work in a real-world environment, so diagnostic errors can arise. There is also a risk factor caused by the relationship to AI-based systems, so in this case this and clinical experience may also influence the final decision.

D. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning

This paper presents a study that investigates the efficiency of traditional machine learning and deep learning approaches in image classification. In particular, it contrasts the performance of support vector machines (SVM), classical machine learning paradigms with convolutional neural networks (CNN), one of the foundations of modern deep learning techniques [4]. It provides a detailed breakdown of how SVMs work by applying kernel functions to process linearly separable and classified data. On the other hand, CNNs are distinguished by their ability to automatically learn hierarchical features directly from raw input data. This includes multiple layers of convolution filters to process complex patterns and structures in images - making them highly effective for tasks involving image data. It should be noted that SVM, although versatile and powerful in some cases, may not scale as well as CNNs with increasing data size and complexity and hence, CNN, regardless to be more proficient with large data sets requires greater computational power.

E. Data augmentation for improving deep learning in image classification problem

The frequent issue in machine learning often revolves around the lack of data or highly unbalanced datasets. This dilemma is particularly prominent in melanoma detection, where the distribution of samples may heavily favor benign moles. Researchers commonly address this challenge through the utilization of data augmentation techniques. In this paper, the authors provide a concise overview of the current state-of-the-art methodologies for data augmentation. They classify these methods into two primary categories: white-box methods, such as traditional transformations (shear, zooming, reflection, rotation, and histogram operations), and black-box methods, which are based on deep neural networks. While traditional transformations are easily replicable, reliable, and swift, they do not introduce significant visual alterations that could benefit model learning [5]. Although the application of white noise is conceivable, it typically fails to generate samples that are perceptibly distinct to the human eye. Conversely, proposed black-box methods, such as Generative Adversarial Networks or style transfer, offer the potential to create synthetic yet realistic image samples that introduce novel visual features. Nonetheless, their drawback lies in their high computational demands.

F. Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach

The paper ”Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach” by Daghrir et al. introduces a hybrid method for melanoma skin cancer detection, leveraging convolutional neural networks - CNN, and classical machine learning classifiers (kNN, SVM) trained with features describing skin lesions. In the related work section, the authors discuss existing methods for melanoma detection, emphasizing the use of visual clues like the ABCDE (asymetry, border, color, diameter, evolving) signs and dermoscopy in diagnosis. They note the importance of image acquisition, pre-processing, lesion segmentation, and classification methodologies in melanoma detection systems [6]. For the pre-processing DOG and thresholding methods were proposed for hairline removal. Segmentation methods like thersholding and hierarchical clustering are mentioned and the inpainting algorithm for repairing the hairline pixels. Features can be excracted using various algorithms, like colorSIFT for colors or HOG for texture. The model with the highest accuracy was designed, by majority voting among the models mentioned in the begining of the section. The most significant challenge highlighted in the paper is the segmentation of skin lesions from healthy skin, which is crucial for accurate diagnosis. Authors also discuss the difficulties in the extraction of features like texture descriptors, color features, and border characteristics for lesion classification. The experiments and results section presents the performance evaluation of the proposed method using a public dataset from the International Skin Imaging Collaboration.

G. Developing a Recognition System for Diagnosing Melanoma Skin Lesions Using Artificial Intelligence Algorithms

This is a more in depth analysis, where authors explore biomedical imaging, particularly within the scope of melanoma detection employing computer-aided diagnosis (CAD) systems enhanced by artificial intelligence. The research lies between dermatology and modern imaging technologies, targeting the early detection of melanoma, and therefore increasing survival rates through an early intervention [7]. This review also builds on ongoing research using modern datasets such as PH2 and ISIC 2018, highlighting the real-world effectiveness of the existing implementations. The authors point out the limitations of existing methods, especially in the areas of accurate separation, extraction, and flexibility of all previously trained models on complex cases of skin cancer. Experimental outcomes showed that the artificial neural network (ANN) model reached high accuracy rates of 97.50% for the PH2 datasets and 98.35% for the ISIC 2018 datasets, outperforming the convolutional neural network (CNN) models in these datasets.

III. Plan of the Experiment

Our research aims to prepare an accurate machine learning model that recognizes melanoma from pictures using traditional and deep learning techniques. The goal is to identify which models perform best in diagnosing skin cancer from dermatoscopic images. The results will help identify the most promising approaches.

As a result of our experiment, we want to find answers for following questions:

1) What machine learning models (Extreme Gradient Boosting, Convolutional Neural Network) are most effective for classifying skin lesions as melanoma or benign from dermatoscopic images?
2) How does the resolution and quality of input images affects the performance of melanoma detection models?
3) How data augmentation techniques can impact the performance of melanoma detection models?

A. Datasets

1) Melanoma Skin Cancer Dataset of 10000 Images [8]: This dataset is split into two groups: test and train. Each group consists of two sets of images, presenting benign and malignant melanoma. Training set is mildly imbalanced, with benign class containing 4605 samples and malignant class 5000 samples. Test set is balanced with both classes containing 500 images. The entire test set has 103.28 MB and contains 10605 samples. Images are in JPEG format. The sizes of images are in range 5 - 30KB. Figures 1 and 2 presents examples from each class of the dataset.

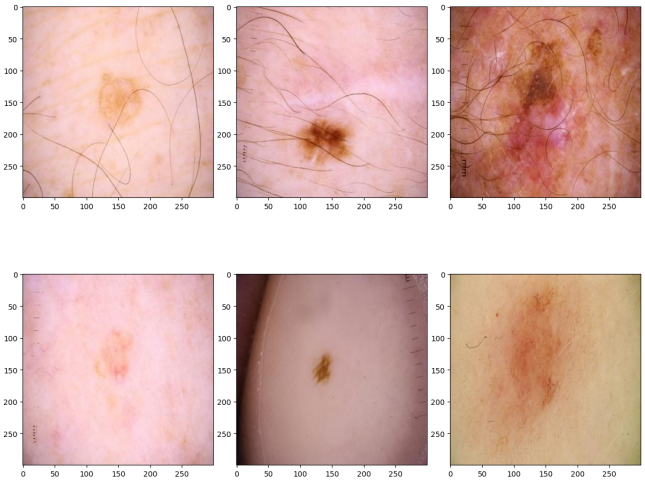


Fig. 1. Example images labeled as benign

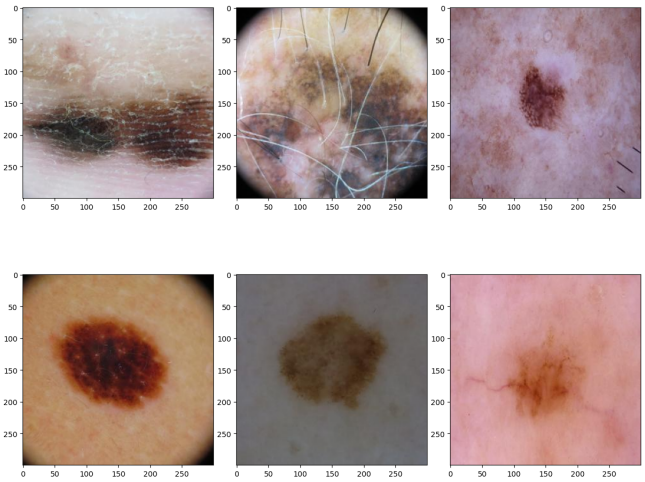


Fig. 2. Example images labeled as malignant

IV. Research Protocol

Our research involves several key steps, including data collection, pre-processing, network design, training, and evaluation.

A. Classification

Since the task involves classification only two categories - malignant and benign - a binary classification approach was used. Thanks to that system was able to correctly classify most cases of images where suspicious skin mark is visible. The techniques described below can be used to carry out preprocessing. Our goal is to compare following classification methods:

- Extreme Gradient Boosting
- Convolutional Neural Networks

B. Pre-processing

We increased the learning stability and convergence rate by normalizing the pixel values of the images and standardizing

the range of input features. To achieve this, each pixel value is divided by 255 and shifted from 0 to 1, as typical image pixels range from 0 to 255 pixels.

To increase the diversity of the dataset and generalise models, image enhancement techniques such as rotation (up to 20 degrees), random resize, color jittering, horizontal and vertical flipping are used.

C. Evaluation

Evaluation of the quality of the classification was carried out using the following metrics:

Precision: Calculates the ratio at which the predicted positives were actually positives in binary classification.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall: Calculates the ratio at which the actual positives were correctly classified.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Accuracy: Measures the overall correctness of the model by calculating the ratio of correct predictions to the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

F1-Score: This is the harmonic mean of Precision and Recall, presenting a balanced mean of these parameters. F1-Score indicates if the classifier has good performance on detecting positives.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

Each of those metrics is measured 5 times during **Cross Validation**, with the usage of **KFold** method. On the dataset to be trained, a division into five segments is performed. This means that in each of the five learning cycles, 80% of the data was used for training and 20% for validation.

Once the results from cross validation are gathered for each of the models, the mean values and standard deviation can be calculated for each instance of the model.

To ensure that results of cross validation evaluations are a part a normally distributed population, we run Shapiro-Wilk tests.

Paired t-tests will also be carried out to determine the relevance of average accuracy values between models and between different pre-processing methods.

In addition, methods such as Grad-CAM (for the CNN model) and LIME (for the XGBoost model) were used to explain and understand the decisions that the models make. These techniques mark areas of relevance to the model in the image.

- Grad-CAM (Gradient-weighted Class Activation Mapping) is a method for visualising the activation of individual neural network layers.
- LIME (Local Interpretable Model-agnostic Explanations) is a method that allows model interpretation with linear models.

V. Experimental environment

A. XGBoostClassifier

The XGBoost model is an efficient and optimised machine learning algorithm based on decision trees. It works well with both large and smaller learning datasets.

During pre-processing in this model, images were loaded from folders containing categorised files (*benign* and *malignant*). Based on the folders in which the images are located, the data is labelled accordingly. Images are flattened, that is transforming a two-dimensional array of pixels into a one-dimensional feature vector. This is followed by a normalization of the image pixels by dividing their values by 255.0 to obtain values between 0 and 1.

The model has the following parameters:

- **n_estimators=100** - number of trees
- **max_depth=5** - maximum depth of the tree
- **learning_rate=0.1** - learning rate
- **tree_method='hist'** - efficient tree construction method
- **device='cuda'** - use of GPU (if available) for accelerated computations

Once the cross-validation stage was completed and the model was satisfied that it was working correctly, the final training of the model on the entire training dataset (*x_train* and *y_train*) and its evaluation on the test dataset proceeded with the same hyperparameters as during cross-validation. The final training aims to use the full availability of the training data to make the model as optimal as possible.

B. Convolutional Neural Network

Convolutional Neural Networks are a class of deep neural networks that are particularly well suited for analyzing visual data. They are widely used for image recognition, classification, and various computer vision tasks. CNNs utilize convolutional layers, which apply filters to the input data to extract relevant features like edges, textures, and shapes.

Convolutional Neural Networks key components are:

- **Convolutional layers** - These layers apply convolution operations to the input image, extracting features using multiple filters. Each filter detects specific features such as edges or patterns.
- **Activation function** - A Rectified Linear Unit (ReLU) is applied after each convolutional operation to introduce non-linearity.
- **Pooling Layers** - Reduce the spatial dimensions of the data (downsampling), which helps reduce computational requirements and extracts dominant features.
- **Fully connected layers** - Act similar to traditional neural networks and are used for classification at the final stages.
- **Flattening** - Converts the 2D matrix data output from convolutional/pooling layers into a 1D vector to feed into the fully connected layers.

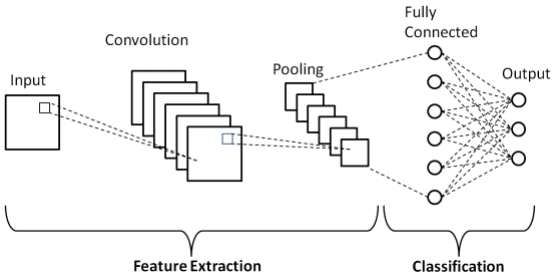


Fig. 3. CNN architecture for image classification [9]

Our architecture consists of 3 convolutional layers. Each layer contains batch normalization (32, 128, and 256 features for each layer) for faster and more stable training, ReLU activation function and max pooling of 2x2 kernel size. 3 fully connected layers are applied. First layer flattens the output of the last convolutional layer. Next fully connected layers contain 1000, 500 units and ReLU activation functions. Final fully connected layer contains 2 output units for binary classification.

Main training loop consists of a KFold cross validation split for a train and validation subsets for each fold. Each fold consists of 40 learning epochs. For each epoch, images are loaded to GPU using CUDA (if it is available). Gradients for each tensor are cleared using Adam optimizer with learning rate set to 0.001. Model then learns on batches of 32 loaded images and loss is calculated with cross entropy criterion. Automatic mixed precision *(AMP)* is used to improve performance and reduce memory usage on GPUs. Loss is then scaled and backpropagation is performed to compute new gradients. After that, model is validated on the training data. Each epoch ends with logging out precision and loss. Each fold is validated to acquire final metrics.

VI. Research results

To find answers for our research questions, we performed tests including XGBoost and Convolutional Neural Network model. We have run tests for following scenarios:

- Without any data augmentation.
- With data augmentation techniques.
- With scaled down and noised data (Gaussian noise).

A. Model evaluation

We evaluated the performance of both XGBoost and Convolutional neural network models using accuracy, precision, recall and F1-score metrics. Models were trained on data with no augmentation techniques.

Results of the evaluation are presented in Tables I, II and Figure 4.

TABLE I
EVALUATION RESULTS FOR THE CNN MODEL

fold	**Accuracy**	**Precision**	**F1 Score**	**Recall**
1	0.8543	0.8577	0.8542	0.8543
2	0.8558	0.8569	0.8559	0.8554
3	0.8329	0.8330	0.8329	0.8327
4	0.8626	0.8626	0.8626	0.8626
5	0.8287	0.8300	0.8287	0.8288
average	0.8469	0.8480	0.8469	0.8468
std deviation	0.0151	0.0153	0.0151	0.0150

TABLE II
EVALUATION RESULTS FOR THE XGBOOST MODEL

fold	**Accuracy**	**Precision**	**F1 Score**	**Recall**
1	0.9179	0.9404	0.9118	0.8848
2	0.9241	0.9363	0.9223	0.9087
3	0.9250	0.9277	0.9197	0.9119
4	0.9274	0.9276	0.9229	0.9183
5	0.9222	0.9264	0.9196	0.9130
average	0.9233	0.9317	0.9193	0.9074
std deviation	0.0035	0.0063	0.0045	0.0130

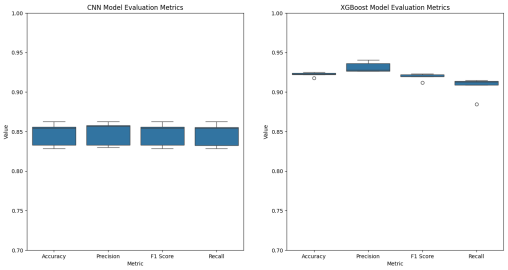


Fig. 4. Comparison of XGBoost and CNN metrics without data augmentation.

To check the normality of the distribution of the results, we run Shapiro-Wilk tests. Results can be seen in Tables III and IV.

TABLE III
RESULTS OF THE SHAPIRO-WILK TEST FOR CNN MODEL METRICS

Metric	**W-statistic**	**p-value**
Accuracy	0.8673	0.2558
Precision	0.8160	0.1088
F1 Score	0.8685	0.2605
Recall	0.8651	0.2470

TABLE IV
RESULTS OF THE SHAPIRO-WILK TEST FOR XGBOOST MODEL METRICS

Metric	W-statistic	p-value
Accuracy	0.9635	0.8322
Precision	0.8245	0.1265
F1 Score	0.8217	0.1204
Recall	0.7952	0.0741

Because p-values for all metrics are above 0.05, we can run paired t-tests to check if there is a significant difference between models accuracies.

The null hypothesis assumes there is no significant difference between the average accuracies of the XGBoost and CNN models. Table V provides details - the null hypothesis can be rejected.

TABLE V
RESULTS OF PAIRED T-TESTS FOR CNN AND XGBOOST MODELS WITHOUT AUGMENTATION

Metric	Value
T-statistic	11.0516
P-value	4.0047e-06

These results indicate that the XGBoost model performs better on raw, unmodified images. This is confirmed by all measured metrics. However, this may mean that the model is not learning the patterns correctly.

B. Impact of data augmentation

In order to check the impact of data augmentation, we trained our models on data with following transformations:

- Horizontal and vertical flips.
- Rotation up to 20 degrees.
- Color jitter.

Example of a transformation can be seen on Figure 5.

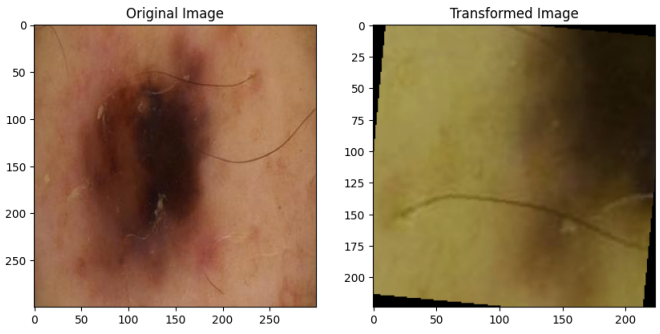


Fig. 5. Example of a transformation applied on an image from the dataset.

Tables VI, VII and Figure 6 shows metrics for models that were trained on data with augmentation techniques applied.

TABLE VI
EVALUATION RESULTS FOR CNN MODEL WITH DATA AUGMENTATION TECHNIQUES

fold	Accuracy	Precision	F1 Score	Recall
1	0.8943	0.8948	0.8943	0.8941
2	0.8777	0.8790	0.8777	0.8773
3	0.8985	0.8987	0.8985	0.8985
4	0.8996	0.8999	0.8995	0.8995
5	0.9095	0.9102	0.9100	0.9100
average	0.8959	0.8965	0.8960	0.8959
std deviation	0.0116	0.0113	0.0118	0.0119

TABLE VII
EVALUATION RESULTS FOR XGBOOST MODEL WITH DATA AUGMENTATION TECHNIQUES

fold	Accuracy	Precision	F1 Score	Recall
1	0.8439	0.7875	0.8480	0.9184
2	0.8378	0.8019	0.8441	0.8909
3	0.8732	0.8464	0.8697	0.8944
4	0.8769	0.8401	0.8788	0.9211
5	0.8553	0.8131	0.8585	0.9092
average	0.8574	0.8178	0.8598	0.9068
std deviation	0.0173	0.0250	0.0147	0.0137

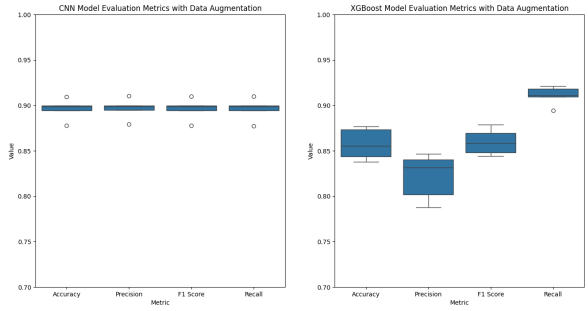


Fig. 6. Comparison of XGBoost and CNN metrics with data augmentation.

Results of the Shapiro-Wilk tests can be seen on Tables VIII and IX.

TABLE VIII
RESULTS OF THE SHAPIRO-WILK TEST FOR CNN MODEL METRICS TRAINED ON AUGMENTED DATA

Metric	W-statistic	p-value
Accuracy	0.9317	0.6085
Precision	0.9402	0.6679
F1 Score	0.9340	0.6302
Recall	0.9347	0.6293

TABLE IX
RESULTS OF THE SHAPIRO-WILK TEST FOR XGBOOST MODEL METRICS TRAINED ON AUGMENTED DATA

Metric	W-statistic	p-value
Accuracy	0.9102	0.4689
Precision	0.9326	0.6147
F1 Score	0.9462	0.7102
Recall	0.8903	0.3588

All p-values are above 0.05, meaning that data comes from a normally distributed population.

To check if the average accuracies of the models with data augmentation are equal to the average accuracy of the model without data augmentation, we carried out a paired t-test comparing the average accuracies from each of the folds. Results can be seen on Tables X, XI and XII.

The null hypothesis assumes there is no significant difference between the average accuracies of the XGBoost and CNN models with augmentation. Table X provides details - the null hypothesis can be rejected.

TABLE X
RESULTS OF PAIRED T-TESTS FOR CNN AND XGBOOST MODELS WITHOUT AUGMENTATION

Metric	Value
T-statistic	-4.1277
P-value	0.0033

The null hypothesis assumes there is no significant difference between the average accuracies of the XGBoost model with augmentation and without augmentation. Table X provides details - the null hypothesis can be rejected.

TABLE XI
RESULTS OF PAIRED T-TESTS FOR XGBOOST ACCURACIES WITHOUT AND WITH AUGMENTATION

Metric	Value
T-statistic	8.3365
P-value	3.2422e-05

The null hypothesis assumes there is no significant difference between the average accuracies of the CNN model with augmentation and without augmentation. Table XII provides details - the null hypothesis can be rejected.

TABLE XII
RESULTS OF PAIRED T-TESTS FOR CNN ACCURACIES WITHOUT AND WITH AUGMENTATION

Metric	Value
T-statistic	-5.7680
P-value	0.0004

To better understand the decisions made by the XGBoost model, the LIME algorithm was used to mark the relevant area for classification. By looking at Figures 7 and 8, it is easy to see that the model, after augmentation, marks random areas that are irrelevant for classification.

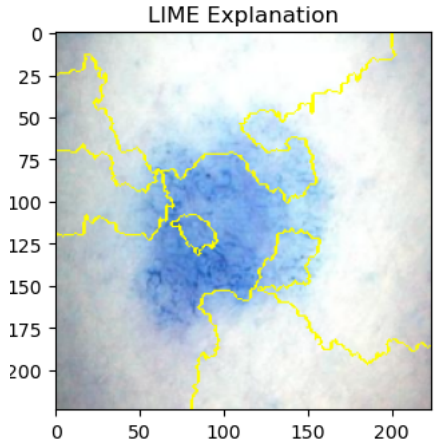


Fig. 7. Visual Lime Explanation for XGBoost without data augmentation.

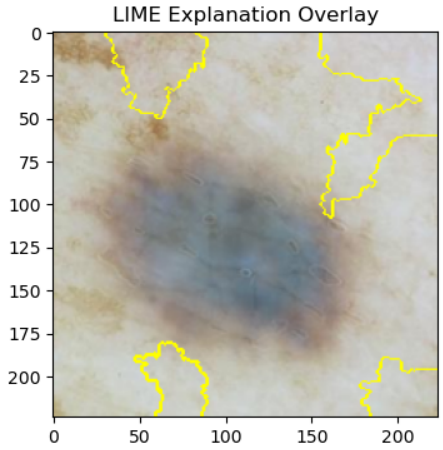


Fig. 8. Visual Lime Explanation for XGBoost with data augmentation.

The CNN model has increased in all metrics by a significant amount. An important fact is that the recall value has increased, which can be particularly significant in skin cancer diagnosis (minimise the number of missed melanoma cases).

To check how data augmentation impacts CNN layers, we created heatmaps for each activation layer for both CNN models using Gradient-weighted Class Activation Mapping (Grad-CAM) [10].

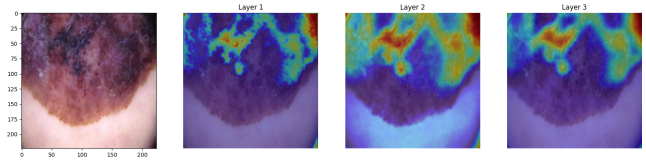


Fig. 9. Visualization of each layer activation for CNN without data augmentation.

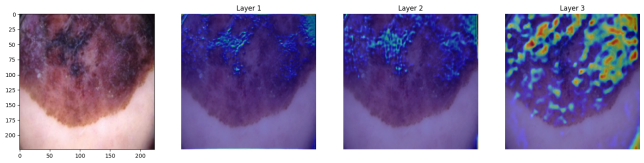


Fig. 10. Visualization of each layer activation for CNN with data augmentation.

Results of the visualization reveal that the CNN model trained on augmented data is able to recognize more complex and diverse features, indicating a better generalization of the network. This shows that it is less prone to over-fitting, leading to better classification quality on unknown data.

C. Impact of image quality and resolution

To check how image quality and resolution impacts performance of the models, we trained both models on dataset which had scaled down images to 90x90 resolution and Gaussian noise applied. All of data augmentation techniques were performed as before. Example of a transformation can be seen on Figure 11.

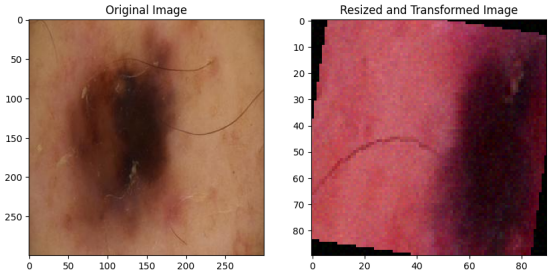


Fig. 11. Example of a resize and transformation applied on an image from the dataset.

Tables XIII, XIV and Figure 12 present validation results for each fold for CNN and XGBoost models.

TABLE XIII
EVALUATION RESULTS FOR CNN MODEL WITH SIZED DOWN IMAGES.

fold	**Accuracy**	**Precision**	**F1 Score**	**Recall**
1	0.8537	0.8542	0.8538	0.8534
2	0.8652	0.8667	0.8652	0.8648
3	0.8657	0.8709	0.8657	0.8647
4	0.8423	0.8578	0.8422	0.8402
5	0.8652	0.8652	0.8652	0.8651
average	0.8584	0.8630	0.8584	0.8576
std deviation	0.0103	0.0068	0.0104	0.0109

TABLE XIV
EVALUATION RESULTS FOR XGBOOST MODEL WITH SIZED DOWN IMAGES.

fold	**Accuracy**	**Precision**	**F1 Score**	**Recall**
1	0.7742	0.7002	0.7934	0.9154
2	0.7902	0.7443	0.8042	0.8746
3	0.8062	0.7461	0.8139	0.8954
4	0.8213	0.7872	0.8241	0.8647
5	0.8274	0.8058	0.8257	0.8467
average	0.8039	0.7567	0.8123	0.8794
std deviation	0.0220	0.0412	0.0136	0.2676

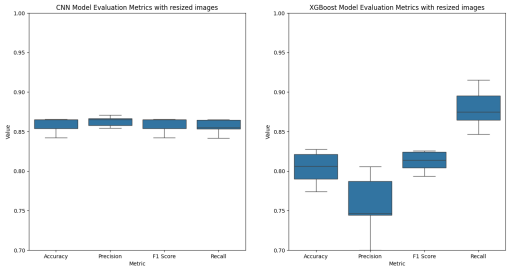


Fig. 12. Comparison of XGBoost and CNN metrics with scaled down images.

Results for the Shapiro-Wilk tests can be seen on Tables XV and XVI.

TABLE XV
RESULTS OF THE SHAPIRO-WILK TEST FOR CNN MODEL METRICS WITH SCALED DOWN IMAGES

Metric	**W-statistic**	**p-value**
Accuracy	0.7865	0.0627
Precision	0.9465	0.7124
F1 Score	0.7859	0.0619
Recall	0.8651	0.7778

TABLE XVI
RESULTS OF THE SHAPIRO-WILK TEST FOR XGBOOST MODEL METRICS WITH SCALED DOWN IMAGES

Metric	**W-statistic**	**p-value**
Accuracy	0.9519	0.7514
Precision	0.9540	0.7657
F1 Score	0.9274	0.5788
Recall	0.9846	0.9578

All p-values are above 0.05, meaning that data comes from a normally distributed population.

Results of the paired t-test comparing the average accuracies from each of the folds can be seen on Tables XVII, XVIII and XVIII.

The null hypothesis assumes there is no significant difference between the average accuracies of the XGBoost and CNN models with scaled data. The Table XVII provides details - the null hypothesis can be rejected.

TABLE XVII
RESULTS OF PAIRED T-TESTS FOR XGBOOST AND CNN ACCURACIES WITH SIZED DOWN IMAGES

Metric	Value
T-statistic	-5.0201
P-value	0.0010

The null hypothesis assumes there is no significant difference between the average accuracies of the XGBoost model with and without scaled data. Table XVIII provides details - the null hypothesis can be rejected.

TABLE XVIII
RESULTS OF PAIRED T-TESTS FOR XGBOOST ACCURACIES WITHOUT AND WITH SCALED DATA.

Metric	Value
T-statistic	11.9927
P-value	2.1539e-06

The null hypothesis assumes there is no significant difference between the average accuracies of the CNN model with and without scaled data. Table XIX provides details - the null hypothesis can't be rejected.

TABLE XIX
RESULTS OF PAIRED T-TESTS FOR CNN ACCURACIES WITHOUT AND WITH SCALED DATA.

Metric	Value
T-statistic	-1.4150
P-value	0.1948

Because p-value of the test between CNN model with and without scaling data is higher than 0.05, we can assume that image quality does not cause any statistically significant difference for CNN model. However, visualization of activation layers (Figure 13) shows, that even though the performance of the CNN model has improved (comparing to CNN model trained without any data augmentation), it also focuses on less significant features. This could be caused by artifacts that lower quality images contain.

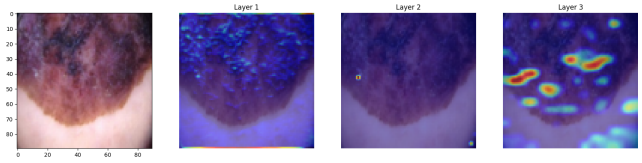


Fig. 13. Visualization of each layer activation for CNN with scaled down images.

D. Conclusions

The XGBoost model performed better than the Convolutional Neural Network (CNN) across all metrics when data augmentation was not applied. This indicates that the model might have been over-fitted to the training data.

Data augmentation notably enhanced the performance of the Convolutional Neural Network model. The CNN metrics significantly improved, highlighting why this model is widely used in image recognition. CNN can handle more diverse and complex features introduced through our augmentation techniques. On the other hand, the performance of the XGBoost model decreased when augmentation was applied. This could be due to the model's reliance on specific features that became less identifiable after augmentation, leading to a decrease in relevant area identification as shown in the LIME visual explanations.

Grad-CAM visualisations showed that CNNs trained with augmented data captured complex features and focus on the distinctive stain that prompted the medical check-up. This indicates better generalization capabilities, making the model less prone to over-fitting.

Training models with reduced image resolution and added Gaussian noise did not degrade the CNN's performance substantially and even slightly improved assessment metrics. This suggests that CNNs are versatile and maintain good performance even with lower-resolution images and added noise. However, the XGBoost model's performance was less impacted by these changes, but also showing decreasing of quality.

In summary, although XGBoost shows good performance in simple image classification tasks, CNNs outperform it in more demanding contexts where data augmentation techniques and noise immunity are necessary. This is particularly important in applications requiring high accuracy and reliability in diverse and complex datasets, such as recognizing melanomas in images.

REFERENCES

[1] Albert T Young, Niki B Vora, Jose Cortez, Andrew Tam, Yildiray Yeniay, Ladi Afifi, Di Yan, Adi Nosrati, Andrew Wong, Arjun Johal, et al. The role of technology in melanoma screening and diagnosis. *Pigment Cell & Melanoma Research*, 34(2):288–300, 2021.

[2] Piyu Parth Naik. Cutaneous malignant melanoma: a review of early diagnosis and management. *World journal of oncology*, 12(1):7, 2021.

[3] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[4] Pin Wang, En Fan, and Peng Wang. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 141:61–67, 2021.

[5] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. pages 117–122, 2018.

[6] Jinen Daghrir, Lotfi Tlig, Moez Bouchouicha, and Mounir Sayadi. Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach. pages 1–5, 2020.

[7] Fawaz Waselallah Alsaade, Theyazn HH Aldhyani, and Mosleh Hmoud Al-Adhaileh. Developing a recognition system for diagnosing melanoma skin lesions using artificial intelligence algorithms. *Computational and mathematical methods in medicine*, 2021:1–20, 2021.

[8] Muhammad Hasnain Javid. Melanoma Skin Cancer Dataset of 10000 Images. https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images, 2022. [last access 22/5/2024].

[9] MK Gurucharan. Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network. https://www.upgrad.com/blog/basic-cnn-architecture/, 2022. [last access 22/5/2024].

[10] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.