ENS 491-492 – Graduation Project
Final Report

Project Title: Skin Lesion Classification with Computer Vision

Group Members:

Ali Oğuz Doğru 27923

Eylül Öykü Şen 28297

Gizem Topsakal 27924

Supervisor(s): Erchan Aptoula

Date: 07.01.2024

1. EXECUTIVE SUMMARY

This project, Skin Lesion Classification with Computer Vision, addresses the critical issue of diagnosing cancerous skin lesions, a growing concern in dermatology. Each year, thousands of people are diagnosed with skin cancer, and early detection is pivotal for successful treatment. However, due to the complexity and variability of skin lesions, skin lesion classification poses significant challenges. Our project aimed to develop a computer vision based deep learning algorithm to accurately classify skin lesions. The system will be designed to support dermatologists in diagnosing skin cancer more efficiently, therefore improving patient outcomes and contributing to the public health care systems.

The project is developed by implementing state-of-the-art CNN architectures, such as ResNet, ResNext and EfficientNet. These models were selected based on their proven capabilities in image classification tasks. The project involves extensive preprocessing techniques of the ISIC2019 dataset, which includes various skin lesions. Utilized techniques were the removal of black corners, hair masking using DullRazor algorithm, and application of the Shade of Gray method to enhance image quality. Additionally, to balance the data various augmentation techniques were used.

Our findings indicate significant advancements in the accuracy of skin lesion classification. The project will demonstrate that ensembling various CNN models could lead to improved performance compared to individual models. The use of advanced preprocessing and augmentation techniques further enhanced the system's effectiveness, addressing the challenges posed by diverse and complex skin lesion images.

In conclusion, this project will try and contribute to the field of medical imaging and computer vision by providing a system for skin lesion classification. It holds promise for practical applications in healthcare, supporting the early detection of skin cancers and potentially saving lives.

2. PROBLEM STATEMENT

2.1 Background

The original problem our project handles is the complex and critical task of accurately classifying skin lesions for early detection of skin cancers such as melanoma. Skin cancer is

becoming increasingly prevalent worldwide and is one of the common forms of cancer. The American Academy of Dermatology reports that approximately 9,500 people in the U.S. are diagnosed with skin cancer daily. Detecting and treating skin lesions promptly greatly enhances outcomes. However visually assessing these lesions poses challenges due to their appearances and subtle variations, between malignant types. Incorrect classification can result in delayed treatment or unnecessary biopsies negatively impacting both being and healthcare resources.

We decided to take on this project because we believe that dermatology could benefit from cutting-edge technology. Thanks to advancements in computer vision and deep learning there is a chance to greatly enhance the precision and effectiveness of classifying skin lesions. Our aim was to leverage these technologies to create a system that can assist dermatologists in making more accurate and faster diagnoses, ultimately leading to better patient care and outcomes.

In terms of goals, our project aimed to develop a computer vision system using deep learning algorithms to classify skin lesions more accurately than current methods. This involved preprocessing a comprehensive dataset of skin lesion images, selecting and training advanced CNN architectures, and employing data augmentation techniques to balance the dataset.

2.2 Related Work

Prior research in this domain has made significant strides. Research like Nils Gessert et al. (2019) addressed the ISIC 2019 Skin Lesion Classification Challenge by developing a method that utilized dermoscopic images and patient data. Their unique approach to handling data imbalance and varying image resolutions, which included loss balancing and a range of EfficientNet models with different cropping techniques, was highly effective. This led to their top ranking in the challenge, achieving balanced accuracies of 63.6% and 63.4% in the two tasks, setting a new benchmark in this area.

Following this, Hassan et al. (2020) also addressed skin lesion classification using densely connected convolutional networks, achieving around 92% accuracy but faced challenges with data imbalance and quality. However, their work highlighted ongoing challenges with data imbalance and quality.

Another significant contribution in this area was made by Mohamed A. Kassem, Khalid M. Hosny, and Mohamed M. Fouad (2020). They developed a deep convolutional neural network model, leveraging transfer learning and the pre-trained GoogleNet model, to classify skin lesions into eight categories. Their model, initially based on GoogleNet parameters and further refined through training, demonstrated exceptional performance on the ISIC 2019 dataset. The model achieved impressive metrics: 94.92% accuracy, 79.8% sensitivity, 97% specificity, and 80.36% precision.

Taken as a whole, these studies offer significant developments in the field of skin lesion classification, demonstrating the effectiveness of advanced computational models in tackling complex medical imaging problems.

2.3 Objectives/Tasks

Objective 1: Development of a Preprocessing Pipeline

- Task 1.1: Implement an algorithm for removing black corners from images to focus on the lesion itself.
- Task 1.2: Apply the DullRazor algorithm for hair removal from skin lesion images.
- Task 1.3: Utilize the Shade of Gray method to eliminate illumination and noise in the images, enhancing image quality for better analysis.

Objective 2: Data Augmentation

- Task 2.1: Employ ImageDataGenerator for random augmentation of images (rotation, flipping, shearing, scaling) to increase dataset diversity.
- Task 2.2: Ensure balanced class distributions in the dataset by augmenting under-represented classes to match the size of the most populated class.

Objective 3: Implementation of Advanced CNN Architectures

- Task 3.1: Integrate and optimize state-of-the-art CNN models such as ResNet, ResNeXt, and EfficientNet for skin lesion classification.
- Task 3.2: Experiment with different model architectures and parameters to identify the most effective combination for our specific task.

Objective 4: Training and Optimization of Models

- Task 4.1: Conduct training sessions with a combination of different models and optimizers on both preprocessed and non-preprocessed datasets.

- Task 4.2: Implement techniques like early stopping to prevent overfitting and improve model performance.

Objective 5: Validation and Testing of Models

- Task 5.1: Use a validation set to fine-tune the models and select the best performing model based on accuracy.
- Task 5.2: Test the models on an unseen test set to evaluate real-world performance, employing test time augmentation for enhanced accuracy.

Objective 6: Ensembling and Final Model Selection

- Task 6.1: Explore various ensemble methods to combine different models for improved accuracy and reliability.
- Task 6.2: Finalize the model based on different metrics like accuracy, sensitivity, and specificity.

2.4 Realistic Constraints

One of the most significant challenges we faced in our project was the economic limitation in accessing high-performance GPUs which are crucial for the computational demands of our deep learning models. Advanced CNN architectures, such as ResNet and EfficientNet, require huge computational resources for efficient training and testing. Our initial approach included exploring paid cloud-based platforms like Google Colab, but the costs were too high for our project's budget.In response to this challenge, we turned to our university's resources. We utilized TOSUN HPC, our university's high-performance computing facility, which offered access to GPUs. This was a strategic move that significantly reduced our project costs while providing us with the necessary computational power. By leveraging TOSUN HPC, we were able to conduct our research without the financial burden of expensive cloud computing services.

3. METHODOLOGY

3.1 Dataset Properties

The dataset selected for the project, ISIC2019 Training Dataset, consists of images from three different datasets: BCN_20000 Dataset (Combalia et al., 2019), HAM10000 Dataset (Tschandl et al., 2018) and MSK Dataset (Codella et al., 2017). In total there are 25331 images in various sizes from 8 different skin lesion classes. Table 1 presents the name of the classes and the class distributions.

Table 1: ISIC2019 Training Dataset class distributions

Class	Number of Images
Melanoma (MEL)	4522
Melanocytic nevus (NV)	12875
Basal cell carcinoma (BCC)	3323
Actinic keratosis (AK)	867
Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis) (BKL)	2624
Dermatofibroma (DF)	239
Vascular lesion (VASC)	253
Squamous cell carcinoma (SCC)	628

All of the images are given in RGB color mode, JPG format but various different sizes. Considering the class imbalances and images coming from three different datasets, it can be concluded that the task requires extensive preprocessing and augmentation of the images.

3.2 Preliminary Design

With the conceptual groundwork in place, the preliminary design focused on the development and training of individual CNN models along with varying factors. The motivation was to explore the effect of different elements such as optimizers and data augmentations on the dataset in hand. The fact that skin lesion datasets differ from widely used datasets such as ImageNet in terms of number of instances, class imbalances and noise makes it more important to combine and test the state-of-art approaches to find out which factors are more important and useful. We hope that this approach paves the way for a better understanding of the characteristics of the problem in hand and a better focused future research in this area.

Leveraging the insights gained from our literature review, we implemented three distinct state-of-art CNN architectures: ResNet, ResNeXt, EfficientNet.

1. EfficientNets represent a series of convolutional neural network models designed to attain top-tier performance while prioritizing computational efficiency. Employing a compound scaling approach that carefully adjusts the network's depth, width, and resolution, EfficientNets aim to achieve optimal efficiency. Gessert et al. (2020) demonstrated the effectiveness of EfficientNets in skin lesion classification tasks, achieving an impressive mean accuracy of 0.926. This success underscores the suitability of EfficientNets for tasks like skin lesion identification.
2. ResNext: ResNext is an extension of the ResNet architecture, which introduces the concept of a "cardinality" parameter to increase the model's capacity (Xie et al., 2016). This parameter controls the number of parallel paths in each residual block, allowing ResNext to capture more diverse and fine-grained features. By leveraging a combination of depth and width, ResNext models excel at capturing both local and global features in images, making them well-suited for skin lesion identification.

The above mentioned models were trained six times each by changing one factor at a time. An overall architecture of the proposed method is presented in Figure 1. A more detailed description of each step in the architecture can be found in subsequent sections.

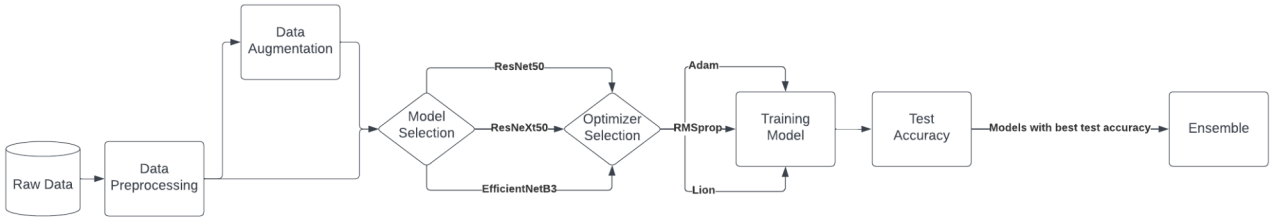
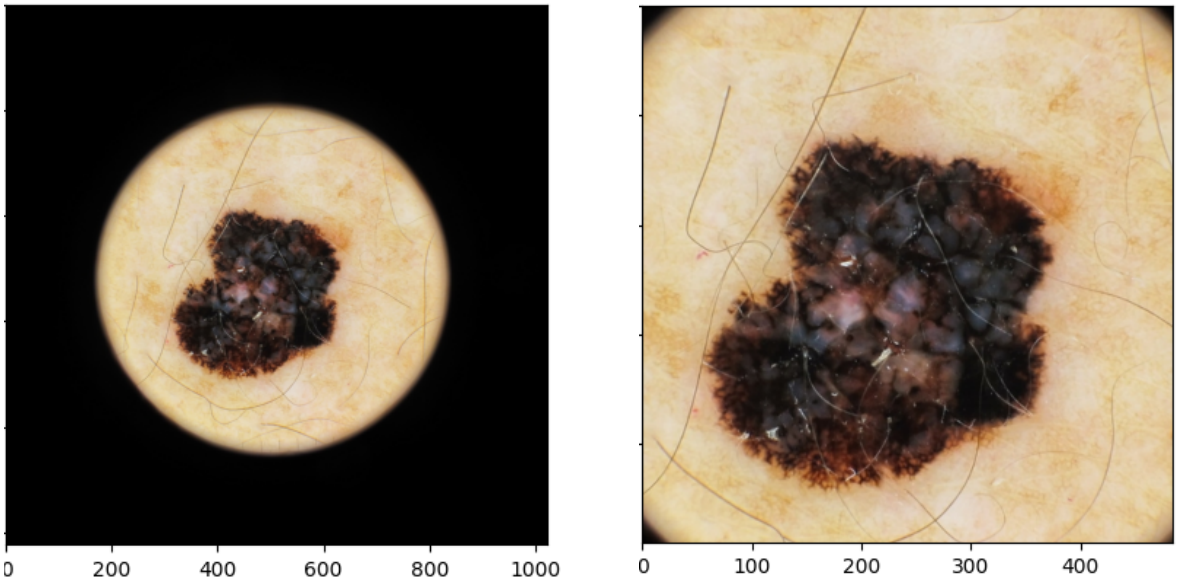


Figure 1: Overall architecture of the proposed methodology

3.3 Pre-processing

To enhance the robustness of our models in accordance with our preliminary design, we implemented preprocessing techniques to refine the input data. Like most of the skin image datasets, the ISIC2019 dataset consists of images that have unwanted artifacts, such as hairs and black corners other than the lesion itself. In order to address this problem we first removed black corners caused by microscopic images. By finding the biggest rectangle that does not contain black borders and cropping the image more by a safety margin we successfully removed the black corners. An example result of this algorithm is given in Figure 2.



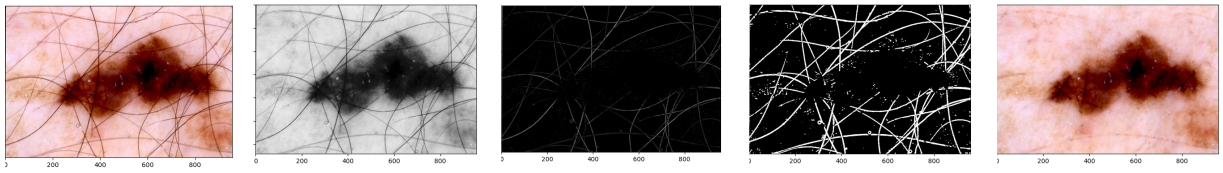
Original image Cropped image

Figure 2: Black corner removal

After cropping black corners, we implemented an algorithm to remove the hairs from the images using a modern implementation of DullRazor, which was first proposed by Lee et al. (1997). The implementation follows the below steps:

1. Converting the original image to grayscale.
2. Closing to the grayscale image, using a linear or cross-shaped kernel.
3. Calculating the difference between the resulting image and the original.
4. Applying binary thresholding to obtain a mask with the hairs in the image.
5. Replacing the pixels in common between the mask and the original image, with pixels from the latter.

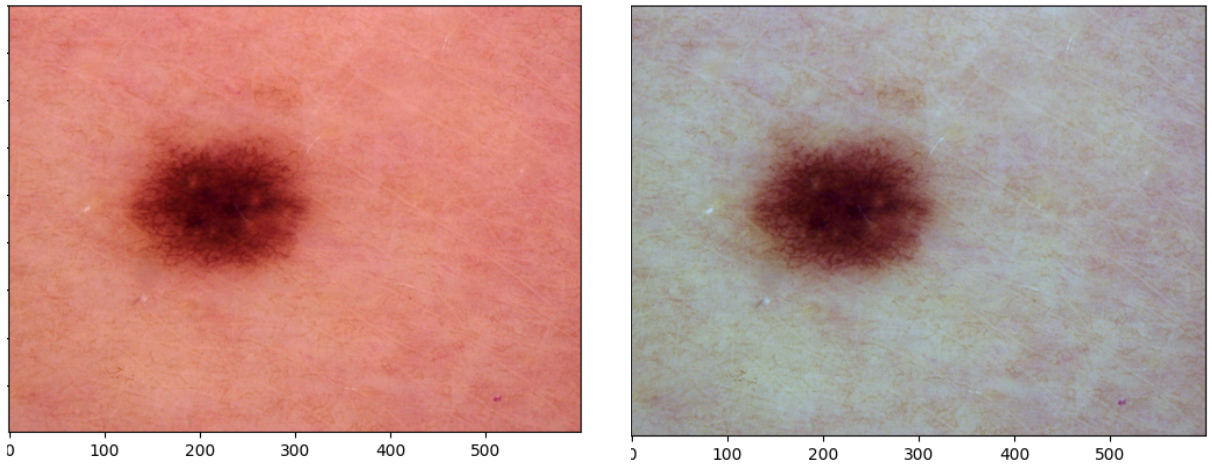
As can be seen from Figure 3, this approach ensured successful removal of the hairs from the image resulting in a better focus on skin lesion itself.



Original image Gray Scale Blackhat Mask Clean Image

Figure 3: DullRazor algorithm's result on an image

Additionally we applied the Shade of Gray method with Minkowski norm p = 6 following the winners of the ISIC2018 (Codella et al., 2019) and ISIC 2019 (Gessert et al., 2020) challenges. The method allowed us to get rid of illumination and noise present in the images as can be found in Figure 4.



Original image Shade of Gray applied image

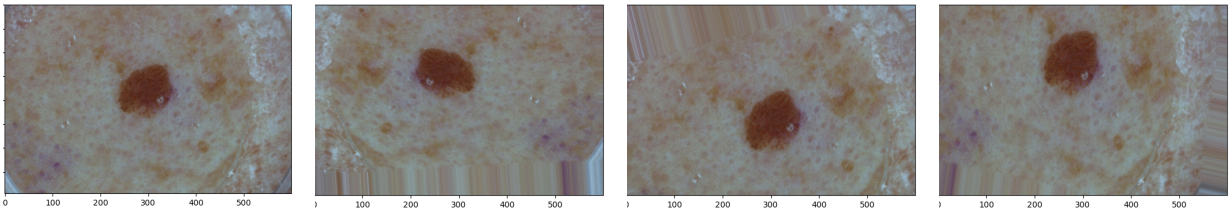
Figure 4: Shade of Gray method

3.4 Data Augmentation

Data augmentation played a crucial role in enhancing model robustness. We employed the ImageDataGenerator to randomly augment images, introducing variations such as rotation, flipping, shearing and scaling. This approach aimed to diversify the dataset, enabling our models to generalize well to unseen data. Besides, we used ImageDataGenerator in a way such that each image in a particular class was augmented as many times as it is required to make the class

distributions equal. Consequently, we increased the dataset size from 25330 to approximately 96000 where each of the 8 classes contain around 12000 images.

A sample of augmentations on the same image by ImageDataGenerator can be found in Figure 5. The image belongs to class “BKL” which has 2624 instances in total. As a result this image is augmented 3 times which makes 4 instances of it in total. This number is equal to approximately the division of the most crowded class size to “BKL” class size.



Original image Augmented 1 Augmented 2 Augmented 3

Figure 5: Data augmentation using ImageDataGenerator

3.5 Training and Testing

In the training phase different approaches were followed to determine the best combination of models, optimizers, hyperparameters and augmentation techniques. We employed the models in the preliminary design along with three different optimizers: Adam (Kingma & Ba, 2014), RMSprop (Tieleman et al., 2012) and Lion (Chen et al., 2023c). Using the above mentioned models as base, we added a few more layers for combatting overfitting, reducing the number of parameters and increasing efficiency. Figure 5 presents the added layers to the base models.

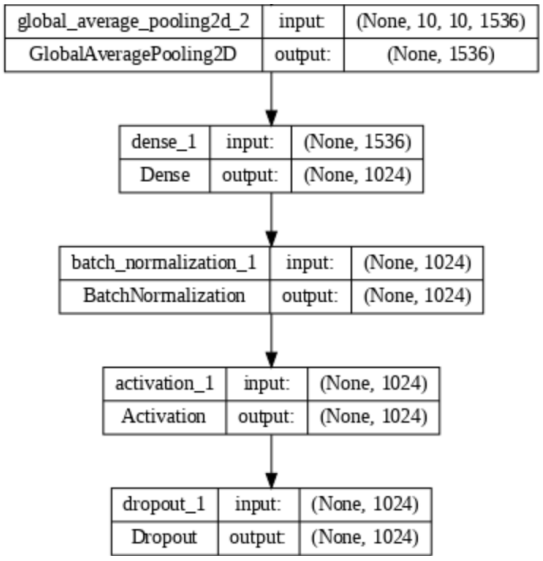


Figure 6: Model architecture after base model

Each model and optimizer combination was inputted a preprocessed but not augmented dataset and both preprocessed and augmented dataset. In all datasets 80%-20% split was used for training and testing. Besides, 10% of the train data was used as a validation set. Table 2 shows the train-validation-test split for both augmented and non-augmented data.

Table 2: Train-Validation-Test split on augmented and non-augmented data

Augmentation	Split	Number of Images
No	Train	17728
	Validation	2541
	Test	5062
Yes	Train	49027
	Validation	7008
	Test	14006

For the non-augmented dataset class-weights were adjusted according to the number of images in the corresponding class to combat class imbalance. Moreover, early stopping after 5 epochs of patience and learning rate scheduler with the parameters factor=0.2, patience=3, min_lr=0.00001 are applied to reduce overfitting resulting from having not enough data points.

Batch size was also fine tuned to 16 after experimenting with different batch sizes. Each model with the given combinations of optimizers and dataset were trained for at least 20 epochs with the training data and the best model according to the validation accuracy was saved during training. After the training phase the saved model was used again for at least 5 more epochs of training on combination of training and validation dataset. The class weights are adjusted again according to the distribution in the combined dataset. Table 3 presents the class weights used during training.

Table 3: Class weights during training with non-augmented data

Class	Class Weight
Melanoma (MEL)	0.70
Melanocytic nevus (NV)	0.25
Basal cell carcinoma (BCC)	0.95
Actinic keratosis (AK)	3.66
Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis) (BKL)	1.21
Dermatofibroma (DF)	13.27
Vascular lesion (VASC)	12.52
Squamous cell carcinoma (SCC)	5.05

After the training and validation phase, the model is tested with the unseen test set. In order to increase the testing performance, test time augmentation is also applied. For this purpose, an image is augmented using techniques like flipping, rotation, shifting and different combinations of augmentations were formed. Later, all of the instances coming from the combinations of augmentations of that image (including the original one but with a higher weight) was predicted using the model and the most probable class was chosen. Lastly, the best models from each CNN used were ensemble using four different techniques: soft voting, hard voting, weighted soft voting and weighted hard voting. The weights were also fine tuned using Optuna (Akiba et al., 2019). The results and comparisons of all methods are presented in the Results & Discussion section.

4. RESULTS & DISCUSSION

This section presents and discusses the results obtained from experimenting with various configurations of ResNet, ResNeXt50, and EfficientNetB3 models on our dataset. The performance of each model was evaluated based on their accuracy and loss metrics across training, validation. Table 4 presents the results for each configuration and test accuracy. A more detailed table can be found in the Appendix A.

Table 4: Model results

Model	Optimizer	Augmentation	Test Accuracy
ResNet	Adam	Yes	0.62
		No	0.62
	RMSprop	Yes	0.66
		No	**0.67**
	Lion	Yes	0.47
		No	0.34
ResNeXt50	Adam	Yes	0.62
		No	0.64
	RMSprop	Yes	0.67
		No	**0.68**
	Lion	Yes	0.46
		No	0.32
EfficientNetB3	Adam	Yes	0.50
		No	0.55
	RMSprop	Yes	0.59
		No	**0.66**
	Lion	Yes	0.44
		No	0.31

The results of the study shows that excessive data augmentation does not improve model performance in most of the configurations, in contrast it gives lower accuracy than non-augmented, class-weight based approaches. For instance the accuracy of EfficientNetB3 with RMSprop has decreased by 0.07 when the model is trained with augmented dataset. Only exception that data augmentation gives higher accuracy is using Lion as the optimizer. In fact, Lion performs worse than any other optimizer in any other settings but it starts to converge when the data is oversampled.

Another observation is that RMSprop performs the best in all settings. Using RMSprop increases the accuracy of ResNet and EfficientNetB3 by at least 0.04 and ResNext50 by at least 0.03. Even though official Keras documentation (Fu, 2020) suggests not using RMSprop as optimizer for EfficientNetB3, the study shows that it can still be the best choice for transfer learning in certain contexts.

In line with these observations, best performing settings from each model are given by no data augmentation (class-weight based approach) and RMSprop optimizer. In these settings, EfficientNetB3 has an accuracy of 66%, ResNet has 67%, ResNeXt50 has 68%. Although the difference between training and test accuracy is relatively high, the initial results still present an opportunity for improvement.

Thu study shows that one significant improvement comes from Test Time Augmentation. In Table 5, the result of Test Time Augmentation on best performing settings of each model is presented. Even though ResNet accuracy did not improve with Test Time Augmentation despite fine tuning parameters, other models showed 0.007 improvement on average.

Table 5: Test Time Augmentation results

Model	Accuracy (without TTA)	Accuracy (with TTA)
ResNet	**0.670**	0.540
ResNeXt	0.680	**0.690**
EfficientNetB3	0.659	**0.664**

While fine tuning and testing different settings and combining with Test Time Augmentation brought a lot of value, the final model was formed using ensemble learning which presented the biggest improvement. Among four different techniques (soft voting, hard voting,

weighted soft voting and weighted hard voting), weighted soft voting performed the best by increasing the accuracy of the overall to 74%. While Table 6 shows the results of the ensemble model using weighted soft voting, detailed tables of all techniques can be found in Appendix B.

Table 6: Weighted soft voting ensemble results

Weighted Soft Voting Accuracy: 0.738

Class	Precision	Recall	F1 Score	Support
Melanoma (MEL)	0.63	0.72	0.67	904
Melanocytic nevus (NV)	0.85	0.80	0.82	2575
Basal cell carcinoma (BCC)	0.64	0.89	0.75	664
Actinic keratosis (AK)	0.52	0.67	0.58	173
Benign keratosis	0.71	0.40	0.51	524
Dermatofibroma (DF)	0.65	0.47	0.54	47
Vascular lesion (VASC)	0.89	0.48	0.62	50
Squamous cell carcinoma (SCC)	0.76	0.46	0.57	125
Macro Average	0.70	0.61	0.63	5062
Weigthed Average	0.75	0.74	0.73	5062

Class 0 and Class 2 showed a strong balance between precision and recall, with f1-scores of 0.67 and 0.75, respectively. Class 2, in particular, exhibited high recall (0.89), indicating the model's effectiveness in identifying most true positives in this category. Class 1, being the class with the highest sample size, showed a high precision of 0.85 and a slightly lower recall of 0.80, resulting in an f1-score of 0.82. This suggests that the model is quite reliable in predicting this class accurately. Class 3, Class 4, and Class 5 experienced a trade-off between precision and recall. Class 3, for instance, had a lower precision but higher recall, indicating a tendency to

classify non-Class 3 samples as Class 3. Class 4 and Class 5, conversely, had higher precision but lower recall, suggesting cautiousness in predicting these classes, potentially leading to missed true positives. Class 6 and Class 7, despite having high precision, showed lower recall, leading to moderate f1-scores. This indicates challenges in the model's ability to consistently identify true positives in these categories.

Overall efficacy of the final ensemble model can be evaluated by weighted and average scores for each metric. The macro-average scores, which treat all classes equally regardless of their sample size, showed lower values (avg precision: 0.70, avg recall: 0.61, avg f1-score: 0.63) compared to the weighted averages. This suggests that the model's performance is more favorable in classes with larger sample sizes. The weighted-average scores, which consider the imbalance in class distribution, were higher (avg precision: 0.75, avg recall: 0.74, avg f1-score: 0.73), indicating that the model performs well in classes with more samples.

In summary, while several approaches improve model performance, careful consideration must be given to class imbalances and the unique characteristics of each class to achieve a more uniform performance across the board. This not only points to the need of research in deep learning for skin lesion classification but also the requirement of collaboration with external entities that can collect and provide diverse and more comprehensive datasets.

5. IMPACT

The "Skin Lesion Classification with Computer Vision" project holds significant potential for scientific, technological, and socio-economic impact in several areas:

Scientific and Technological Impact:

- Advancement in Medical Imaging and Dermatology: Our project uses modern deep learning algorithms to improve the accuracy of skin lesion classification, which advances the field of medical imaging.
- Contribution to Computer Vision and AI in Healthcare: The project shows how AI may be effectively applied in healthcare by combining advanced convolutional neural network (CNN) architectures, and preprocessing methods with data augmentation techniques.

Socio-economic Impact:

- Improving Healthcare Outcomes: The project's ability to make more accurate skin lesion classifications, can lead to earlier detection of skin cancers, significantly improving patient outcomes and potentially saving lives.
- Cost Efficiency in Healthcare: By reducing the need for unnecessary biopsies and enabling more accurate diagnoses, the system can contribute to cost savings in healthcare systems, making diagnostics more affordable and accessible.
- Public Health Awareness: The project's success can raise awareness about the importance of early skin cancer detection and the potential of technology in health monitoring and preventive care.

Innovative and Commercial Aspects:

- Commercial Viability: The technology developed could be integrated into diagnostic tools for dermatologists, offering a valuable commercial product. Its potential for integration into mobile applications for preliminary skin checks also presents an avenue for widespread public use.

Freedom-to-Use (FTU) Issues:

- Intellectual Property: The project utilized publicly available datasets and open-source technologies, mitigating major FTU concerns. However, as we move towards commercialization, a thorough IP review would be necessary to ensure compliance with any existing patents, especially concerning specific CNN architectures and algorithms used.
- Data Usage and Privacy: While the project adhered to data privacy standards in using the ISIC2019 Training Dataset, further development, especially in commercial applications, must rigorously maintain patient data privacy and adhere to regulations like HIPAA (Health Insurance Portability and Accountability Act).

In conclusion, our project not only advances the field of medical imaging and AI but also has the potential to make a tangible socio-economic impact by enhancing healthcare outcomes and efficiency. The innovative aspects offer commercial opportunities, although FTU considerations will be crucial in navigating the path to market.

6. ETHICAL ISSUES

Our project is primarily focused on improving the outcomes of healthcare through technological innovation. It also presents several ethical considerations:

- Data Privacy and Consent: The project uses open-source skin lesion images for now, but if it is ever to become an app, it will need to use patient-derived skin lesion images, which raises concerns about data privacy. It's essential to ensure that all data used in the project, especially in training AI models, is anonymized and used with proper consent.
- Bias and Fairness in AI Models: AI models, including those used in medical diagnostics, can inadvertently exhibit bias based on the data they are trained on. Ensuring our models are trained on diverse datasets representing various skin types and conditions is crucial to prevent any form of demographic or racial bias, thus promoting equitable healthcare.
- Reliability and Misdiagnosis: While our project aims to support dermatologists in diagnosis, over-reliance on AI tools could lead to misdiagnoses if the system fails or errs. The ethical responsibility is to ensure that the system is reliable, but also to clarify that it's a decision-support tool, not a replacement for professional medical advice.
- Transparency and Explainability: There's an ethical need for transparency in how AI models make decisions, especially in healthcare. Our project should strive for explainability in its models, ensuring that dermatologists can understand and interpret the AI's diagnostic suggestions.
- Intellectual Property and Research Integrity: Ethically, it is important to respect intellectual property rights and acknowledge the use of any patented technology or concepts. Our project used open-source tools and datasets, but further development, especially for commercial purposes, must consider potential patent infringements.
- Potential for Misuse: Any technology, particularly in the domain of personal health data, can be misused. It's ethical to consider safeguards against the potential misuse of our technology, such as unauthorized access to or manipulation of medical data.
- Impact on Professional Practice: Introducing AI in medical diagnostics could raise concerns about the impact on professional practice for dermatologists. It's ethical to ensure that such technology is integrated in a way that supports and enhances, rather than diminishes, the role of medical professionals.

In conclusion, while our project does not directly engage with materials or designs that are ethically problematic, it necessitates careful consideration of data privacy, bias and fairness, reliability, transparency, intellectual property, the potential for misuse, and its impact on professional practice.

7. PROJECT MANAGEMENT

Our initial plan for the "Skin Lesion Classification with Computer Vision" project was structured into several distinct phases over two semesters. These phases included a literature review, dataset collection and preprocessing, model selection and training, testing, and finally implementation. Key milestones were set for each phase with specific deadlines to ensure timely progress. The final project plan, while retaining the core structure of the initial plan, underwent several adjustments due to unforeseen challenges and insights gained during the project's progression.

Extended Preprocessing Phase:

While working on the preprocessing phase, we realized that we underestimated the complexity of preprocessing the ISIC2019 dataset. Techniques like black corner removal and hair masking with DullRazor required more time than anticipated for fine-tuning, leading to an extension of this phase.

Model Selection and Training Adjustments:

Initially, we planned to implement and test a limited set of CNN models. However, after further review of the literature and preliminary results, we expanded our scope to include more advanced models like EfficientNet, which required additional time for training and optimization.

Testing and Evaluation Revisions:

The testing phase was extended to incorporate a more comprehensive evaluation strategy. This involved not just assessing individual model performance but also exploring various ensemble techniques to improve overall accuracy and reliability.

Implementation and Integration:

The final implementation phase was adjusted to allow for additional integration testing, ensuring that the system worked seamlessly and could be easily integrated into existing healthcare workflows.

Lessons Learned:

Flexibility and Adaptability:

One of the key lessons was the importance of being flexible and adaptable. The ability to revise our plan in response to technical challenges and new insights was crucial in maintaining project momentum and achieving our objectives.

Time Management and Planning:

We learned that more generous time allowances should be made for complex tasks like data preprocessing and model training. Better anticipation of potential delays could lead to more realistic timeline projections.

Collaboration and Communication:

Regular team meetings and open communication channels were vital in managing the project efficiently. This ensured that everyone was aligned on objectives, progress, and changes to the plan.

Risk Management:

We learned to identify potential risks early and develop contingency plans. This proactive approach helped us navigate through technical challenges more smoothly.

Documentation and Tracking:

Keeping detailed records of our progress, decisions, and changes to the plan was invaluable. This practice not only helped in maintaining an overview of the project status but also served as a learning tool for future projects.

Stakeholder Engagement:

Regular updates and feedback sessions with our supervisor and potential end-users provided valuable insights that shaped the project's direction. Engaging with stakeholders early and frequently is something we would emphasize more in future projects.

8. CONCLUSION AND FUTURE WORK

The project, "Skin Lesion Classification with Computer Vision," has achieved noteworthy advancements in diagnosing skin cancer through improved skin lesion classification using CNN architectures. Our goal to harness the power of advanced convolutional neural networks (CNNs) for the accurate classification of skin lesions has been largely achieved, culminating in a system that not only elevates the standards of medical imaging but also paves the way for practical applications in healthcare.

Our approach, integrating state-of-the-art CNN architectures like ResNet, ResNext, and EfficientNet, coupled with comprehensive preprocessing and data augmentation, has proven effective. The meticulous removal of black corners, hair masking using the DullRazor algorithm, and the application of the Shade of Gray method have significantly enhanced the quality of the dataset, leading to more reliable model training and classification results. Furthermore, balancing the dataset through augmentation techniques has addressed the challenge of class imbalances, a critical aspect in ensuring the model's accuracy and generalizability.

Despite these successes, we acknowledge certain limitations. The most notable is the dataset's scope, which, while extensive, may not fully represent the global diversity of skin types and conditions. This limitation suggests an opportunity for future research, emphasizing the inclusion of a more varied and comprehensive dataset that encompasses a wider range of skin conditions across different demographic groups. Another constraint was the computational resources required for training sophisticated models. Future iterations of this project could explore more efficient computational techniques or leverage emerging technologies to optimize the training process.

Looking ahead, expanding the dataset to incorporate a broader spectrum of skin types and conditions is paramount. This expansion would not only enhance the model's accuracy but also its applicability on a global scale. Furthermore, exploring the integration of these models into clinical settings, along with the development of user-friendly interfaces, will be crucial steps towards making this technology a staple in medical diagnostics.

Moreover, the potential integration of this technology into mobile health applications represents a promising direction for preventive healthcare. Such applications could democratize access to preliminary skin cancer screenings, potentially saving lives through early detection.

In conclusion, this project stands as a testament to the potential of AI and machine learning in revolutionizing medical diagnostics. It highlights the synergy between technological innovation and healthcare, offering a glimpse into a future where AI not only supports but enhances medical expertise, ultimately leading to improved patient outcomes and a new frontier in preventive care.

For future work, our aspirations extend beyond the current scope, aiming to increase the impact and broaden the applicability of our research in the area of medical diagnostics and

patient care. At the forefront of our future endeavors is the expansion of the dataset. The diversity of skin types and conditions in our dataset is a critical factor that directly influences the model's accuracy and global applicability. This expansion will not only enhance the model's accuracy but also its fairness and effectiveness across different demographic groups worldwide. Collaborating with international dermatological databases and institutions will be key to achieving a comprehensive and inclusive dataset.

The development of user-friendly interfaces for medical professionals is another significant aspect of our future work. This algorithm can be converted into an app. We hope to encourage the adoption of our technology in the healthcare sector and so increase its impact by making it user-friendly and easily accessible. Moreover, the potential for integrating our technology into mobile health applications represents a transformative opportunity in preventive healthcare. Such applications could make skin cancer screenings more accessible to the general public, potentially leading to earlier detection and treatment. Developing these applications with a focus on user experience, privacy, and data security will be key to their success and widespread adoption.

Furthermore, we want to make sure that our models are impartial and morally sound. The moral implications of utilizing AI in healthcare must be addressed as the technology continues to gain momentum in this area. This entails protecting the privacy of data, getting permission before using it, and reducing bias in the models. We intend to provide standards and best practices for the moral application of artificial intelligence in dermatology, making sure that our technology is applied sensibly and fairly.

Finally, we acknowledge the significance of interdisciplinary cooperation in fostering creativity and realizing comprehensive solutions. We will get a variety of viewpoints and knowledge by interacting with professionals in domains like bioinformatics, data science, healthcare policy, and patient advocacy. By working together, we will be able to tackle the complex issues surrounding the detection and management of skin cancer, which will ultimately result in more thorough and efficient solutions.

9. APPENDIX

Appendix A. Individual Model Results

						Best Model						TTA
Model Name	**Preprocess?**	**Augmentation?**	**Optimizer**	**#Epochs**	**Batch Size**	**Model Accuracy**	**Model Loss**	**Val Accuracy**	**Val Loss**	**Test Accuracy**	**Test Loss**	**Accuracy**
ResNet	Yes	Yes	Adam	20	32	0.99	0.04	0.60	2.00	0.62	1.88	
ResNet	Yes	No	Adam	25	16	0.98	0.03	0.64	1.72	0.62	1.63	
ResNet	Yes	Yes	RMSprop	20	32	0.74	0.76	0.63	1.5	0.66	1.3	
ResNet	No	No	RMSprop	20	16	0.88	0.52	0.65	1.27	0.63	1.35	
ResNet	Yes	No	RMSprop	20	16	0.97	0.09	0.66	1.32	0.67	1.08	0.54
ResNet	Yes	Yes	Lion	25	32	0.94	0.14	0.47	2.60	0.47	2.10	
ResNet	Yes	No	Lion	25	16	0.38	1.90	0.35	2.30	0.34	1.80	
ResNeXt50	Yes	Yes	Adam	25	32	0.95	0.14	0.67	1.62	0.62	2.14	
ResNeXt50	Yes	No	Adam	20	16	0.98	0.20	0.7	1.77	0.64	1.87	
ResNeXt50	Yes	Yes	RMSprop	25	16	0.97	0.15	0.71	1.82	0.67	1.99	
ResNeXt50	Yes	No	RMSprop	25	16	0.97	0.07	0.67	1.88	0.68	1.09	0.69
ResNeXt50	Yes	Yes	Lion	25	16	0.65	1.28	0.46	1.76	0.46	1.63	
ResNeXt50	Yes	No	Lion	25	16	0.54	1.13	0.42	8.33	0.32	1.97	
EfficientNet B3	Yes	Yes	Adam	20	32	0.99	0.03	0.39	3.88	0.5	3.18	
EfficientNet B3	Yes	No	Adam	25		0.96	0.37	0.65	1.85	0.55	1.96	
EfficientNet B3	Yes	Yes	RMSprop	20	16	0.98	0.06	0.36	4.49	0.59	2.74	
EfficientNet B3	Yes	No	RMSprop	25	16	0.97	0.09	0.67	1.67	0.6592	2.29	0.6642
EfficientNet B3	Yes	Yes	Lion	20	16	0.64	1.15	0.46	2.03	0.44	1.54	
EfficientNet B3	Yes	No	Lion	25	16	0.54	1.17	0.5	1.92	0.31	2.01	

Appendix B. Ensemble Model Results

Soft Voting Accuracy: 0.735

class	precision	recall	f1score	support
Melanoma (MEL)	0.65	0.67	0.66	904
Melanocytic nevus (NV)	0.82	0.83	0.83	2575
Basal cell carcinoma (BCC)	0.65	0.86	0.74	664
Actinic keratosis (AK)	0.46	0.69	0.55	173
Benign keratosis (BKL)	0.75	0.35	0.48	524
Dermatofibroma (DF)	0.59	0.36	0.45	47
Vascular lesion (VASC)	0.92	0.48	0.63	50
Squamous cell carcinoma (SCC)	0.77	0.39	0.52	125
Macro average	0.70	0.58	0.61	5062
Weigthed Average	0.74	0.73	0.73	5062

Hard Voting Accuracy: 0.704

class	precision	recall	f1score	support
Melanoma (MEL)	0.55	0.70	0.62	904
Melanocytic nevus (NV)	0.79	0.81	0.80	2575
Basal cell carcinoma (BCC)	0.67	0.79	0.73	664
Actinic keratosis (AK)	0.50	0.59	0.54	173
Benign keratosis (BKL)	0.78	0.27	0.40	524
Dermatofibroma (DF)	0.59	0.27	0.38	47
Vascular lesion (VASC)	0.92	0.44	0.59	50
Squamous cell carcinoma (SCC)	0.72	0.27	0.40	125
Macro average	0.69	0.52	0.56	5062
Weigthed Average	0.72	0.70	0.69	5062

Weighted Hard Voting Accuracy: 0.721

class	precision	recall	f1score	support
Melanoma (MEL)	0.60	0.67	0.64	904
Melanocytic nevus (NV)	0.83	0.80	0.82	2575
Basal cell carcinoma (BCC)	0.62	0.85	0.72	664
Actinic keratosis (AK)	0.46	0.62	0.53	173
Benign keratosis (BKL)	0.68	0.38	0.49	524
Dermatofibroma (DF)	0.59	0.40	0.48	47
Vascular lesion (VASC)	0.85	0.46	0.60	50
Squamous cell carcinoma (SCC)	0.68	0.43	0.53	125
Macro average	0.67	0.58	0.60	5062
Weigthed Average	0.73	0.72	0.72	5062

Appendix C. Source Code Link

https://github.com/eyluloyku/skin-lesion-classification

10. REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). OpTUNA: a next-generation Hyperparameter Optimization Framework. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1907.10902

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C., Lu, Y., & Le, Q., V. (2023c). Symbolic Discovery of Optimization Algorithms. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.06675

Codella, N., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N. K., Kittler, H., & Halpern, A. C. (2017). Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1710.05006

Codella, N., Rotemberg, V., Tschandl, P., Çelebi, M., Dusza, S. W., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. A., Kittler, H., & Halpern, A. C. (2019). Skin Lesion Analysis Toward Melanoma Detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv (Cornell University)*. https://arxiv.org/abs/1902.03368

Combalia, M., Codella, N., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Halpern, A. C., Puig, S., & Malvehy, J. (2019). BCN20000: Dermoscopic lesions in the wild. *arXiv (Cornell University)*. http://export.arxiv.org/pdf/1908.02288

Fu, Y. (2020). *Keras documentation: Image classification via fine-tuning with EfficientNet*. https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/

Gessert, N., Nielsen, M., Shaikh, M., Shimizu, S., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, *7*, 100864. https://doi.org/10.1016/j.mex.2020.100864

Hassan, S. R., Afroge, S., & Mizan, M. B. (2020, June). Skin lesion classification using densely connected convolutional network. In 2020 *IEEE Region 10 Symposium (TENSYMP)* (pp. 750-753). IEEE. https://doi.org/10.1109/tensymp50017.2020.9231041

Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2018, December). Skin cancer classification using deep learning and transfer learning. In *2018 9th Cairo international biomedical engineering conference (CIBEC)* (pp. 90-93). IEEE.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1412.6980

Lee, T., Ng, V., Gallagher, R. P., Coldman, A. J., & McLean, D. I. (1997). Dullrazor®: A software approach to hair removal from images. *Computers in Biology and Medicine*, *27*(6), 533–543. https://doi.org/10.1016/s0010-4825(97)00020-6

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr52688.2022.01167

Tieleman, T. and Hinton, G. Lecture 6.5 — RMSProp, COURSERA: Neural Networks for Machine Learning. Technical report, 2012.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2016). Aggregated Residual Transformations for Deep Neural Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1611.05431