DETECTION OF SKIN CANCER AND CREDIBILITY ASSESSMENT USING EXPLAINABLE AI

Authors
Khushi Naik, Chetan Sah, Bharati Jagdish Panigrahi, Tanaya Tamhankar
Link to Codebase

1 INTRODUCTION

Skin cancer, the most common cancer diagnosed worldwide, is causing serious health problems and its prevalence is increasing. The situation of skin cancer highlights the urgency of developing comprehensive diagnostic solutions. The deadliness of this cancer is due to mutations in the DNA of humans, who are generally exposed to excessively prolonged periods of sunlight. In the complex scenario of a skin cancer diagnosis, doctors begin the process with a preliminary visual examination, the most common form of skin examination Dermoscopy, a diagnostic technique aided by computer systems, to take high-resolution images of skin lesions to obtain detailed and microscopic understanding of possible malignancies. This large-scale study aims to go beyond the traditional limitations of skin cancer detection and develop sophisticated and subtle skin cancer detection systems. This initiative based on advanced machine learning algorithms aspires to not only increase the accuracy of analysis but also illuminate the black box of decision-making using SHAP. The research focuses on the development of a skin cancer detection system and incorporates Explainable AI (XAI) to improve the model's predictions.

Early detection of skin cancer is important for optimal treatment. Highlighting the importance of early detection, the project emphasizes its role in optimizing treatment outcomes. A robust skin cancer detection system that can distinguish between benign and malignant lesions is needed to facilitate timely intervention and improve patient prognosis. The dataset provided by SIIM-ISIC (Society of Imaging Informatics and Medicine) includes a wide range of images with contextual information such as age and gender. This dataset forms the foundation for training and validating the machine learning models, ensuring a comprehensive representation of skin conditions. To address conceptual ambiguity in deep learning models, descriptive AI techniques, especially SHapley Additive exPlanations (SHAP), are used. This increases the interpretability of the model's measures and provides insight into factors influencing classification. Using XAI the developed model not only provides accurate predictions but also provides interpretable insights. It can serve as a valuable tool for medical professionals in their decision-making process. By integrating the skin cancer detection system with the SHapley Additive exPlanations (SHAP), this research aims to contribute to further efforts to address the challenges associated with a skin cancer diagnosis.

Limitations of existing diagnostic methods call for the search for new methods for accurate detection and interpretation of skin cancer This report addresses this need using interpretable AI (XAI) methods and (XAI) integration, especially in SHapley Additive exPressions (SHAP) integration, to increase skin reliability and transparency cancer screening models. The chapters that follow will go into the methodology used in dataset collecting and preparation, as well as an examination of various deep learning models such as ResNet50, CNN, InceptionResNetv2, and VGG16. The reasoning for model selection, specifically the selection of ResNet50 as the best performer is discussed. This study embarks on an important insight between skin cancer diagnosis and interpretable AI, intending to move the field forward by providing medical professionals with diagnostic tools that are accurate, transparent and accessible.

2 RELATED WORK

1. Anand et al. (2022) presented an enhanced deep learning-based model for the diagnosis of skin cancer. The authors employed a transfer learning approach, modifying a pre-trained VGG16 model with additional layers and activation functions to improve diagnostic accu-

racy. This model demonstrates a high accuracy of 89.09%, outperforming existing state-of-the-art methods.

2. Hasan et al. (2021) investigated the application of Convolutional Neural Networks (CNNs) in detecting skin cancer. It highlights the utility of deep learning in medical diagnostics. Various pre-trained and self-built models, including VGG16, Support Vector Machine (SVM), ResNet50, and sequential models, are examined and compared based on their layer numbers, working processes, and accuracy. The results show VGG16 outperforming other models with a notable accuracy of 93.18%.

3. Nigar et al. (2022) addressed the challenge of delayed diagnosis of skin cancer due to the high similarity of skin lesion types in the early stages. It introduces an explainable artificial intelligence (XAI)–based skin lesion classification system aimed at improving accuracy in skin lesion classification. The proposed model demonstrates high performance with an impressive accuracy of 94.47%. Furthermore, the model employs the Local Interpretable Model-Agnostic Explanations (LIME) framework, enhancing the model's trustworthiness and applicability in clinical practice.

4. Demir et al. (2019) focused on the early detection of skin cancer using deep learning architectures, specifically ResNet-101 and Inception-v3. The study emphasizes the importance of early detection for successful treatment. The paper demonstrates the effectiveness of the ResNet-101 and Inception-v3 architectures in this classification task, achieving accuracy rates of 84.09% and 87.42%, respectively.

5. Knapič et al. (2021) explored the potential of Explainable Artificial Intelligence (XAI) methods in supporting decision-making in medical image analysis. Focusing on improving the comprehensibility of Convolutional Neural Network (CNN) decisions, the study applies three types of XAI methods: Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and Contextual Importance and Utility (CIU).

6. Abhvankar et al. (2021) presented the development of an application for detecting melanoma and non-melanoma skin cancer using deep learning algorithms, specifically Convolutional Neural Networks (CNN) and ResNet-50. The application also incorporates the use of ResNet-50 for detecting both melanoma and non-melanoma types of skin cancer, showcasing the potential of advanced deep-learning techniques in medical diagnostics.

7. Panthakkan et al. (2022) introduced a concatenated Xception-ResNet50 (X-R50) deep learning model for skin cancer detection, achieving a high prediction accuracy of 97.8% on the HAM10000 dataset. This model outperforms existing methods in early-stage skin cancer detection.

8. Jain et al. (2021) presented a comparative analysis of six different transfer learning networks for classifying various types of skin cancer. The networks evaluated include VGG19, InceptionV3, InceptionResNetV2, ResNet50, Xception, and MobileNet. Xception Net demonstrates the highest classification accuracy and efficiency suggesting its effectiveness in distinguishing between different types of skin lesions.

9. Budhiman et al. (2019) focused on melanoma skin cancer classification employing the ResNet architecture. The study aims to determine the most effective ResNet model for distinguishing between melanoma cancer and normal skin images. Various ResNet architectures, including ResNet 50, 40, 25, 10, and 7, were trained with augmented and under-sampled data. The study finds that the ResNet 50 model, without data augmentation, achieves the best results with a validation accuracy of 0.83 and an F1 score of 0.46.

10. Malo et al. (2022) explored the role of artificial intelligence in disease identification, particularly focusing on skin cancer detection using Convolutional Neural Networks (CNNs). The study emphasizes the increasing reliance of dermatologists on digitalised patient results for definitive skin cancer diagnosis. Employing deep learning algorithms, it assesses the potential of CNNs to differentiate between benign and malignant skin moles. The modified VGG-16 model demonstrates promising results, achieving an accuracy of 87.6%.

3 Scope Definition and Project Deployment

3.1 Scope

- The project focuses on utilizing deep learning techniques to provide accurate and reliable classification of skin lesions into benign and malignant categories.
- The project focuses on utilizing deep learning techniques to provide accurate and reliable classification of skin lesions into benign and malignant categories.
- It focuses on the credibility and dependability of AI systems in medical diagnostics.
- The model helps in the early detection of skin cancer which reduces future health problems.

3.2 Implementation

3.2.1 Dataset

The dataset used was provided by SIIM-ISIC (Society of Imaging Informatics and Medicine), see SII (2023), which included both images and supporting .CSV files comprising vital information such as age and gender.

3.2.2 Pre-processing

For consistency in model training, the input data was scaled to a standardized format. The dataset was then balanced using Scikit-Learn to address any potential class imbalance issues. To minimize data size while retaining data shape we used .npy files. Keras Image Data Generator was used to augment the image data.

3.2.3 Model Building

The following four models were selected based on their performance in the previous studies. We have taken code references from these sources, Swilam (2023) and Williams (2023).

- ResNet50
- VGG16
- InceptionResNet
- CNN

Amongst these models the best-performing model is ResNet50. It outperformed other models by avoiding over-fitting. The other models have fluctuating validation recall values. The AUC-ROC curve for ResNet50 constantly neared 1 throughout the training period, indicating that it acts as a stable classifier.

SHAP(SHapley Additive exPlanations) analysis was employed to assess the credibility of the ResNet50 model. The pink-coloured areas represented the regions where the model detected skin anomalies, classifying them as either benign or malignant. Medical experts can use these SHAP analysis results as a reference for prognosis which will help to diagnose and classify skin disorders.

4 Methodology

To perform Skin cancer detection, we have used the SIIM-ISIC dataset which is classified into 2 types of Skin cancer namely Benign and Malignant. Nowadays, skin cancer detection is a common proposal. However, since the medical profession is very highly sensitive, relying solely on model predictions to determine a person's cancer type is not a favourable approach. Models suffer from high false positive rates because of the lack of transparency, which impacts the diagnosis. Our approach is centred on using Explainable AI's SHapley Additive explanations (SHAP) to emphasize the image's attributes and positively influence the model's conclusions. To do this, we have carried out several steps as shown in Figure 1.

Figure 1: Steps performed for skin cancer detection using Explainable AI

4.1 Data collection and Pre-processing

One of the main concerns of applying deep learning for this task is the lack of training data. As mentioned above the SIIM- ISIC dataset was used for the task however, the number of samples that are now accessible is still insufficient, see Pacheco & Krohling (2019). There were only 584 photos of malignancy in the data set, compared to 32542 images of benign conditions. To produce a more extensive dataset for training the model and to minimize the amount of imbalance, datasets from 2019 and extra cancerous images were combined with the original dataset; however, the difference remained significant. Once the dataset was generated, we used Image Data Generator to create augmented data that were both zoomed in and out, as well as by scaling the images to 128 x 128 pixels for increasing diversity in the dataset. The augmented images and corresponding labels were saved as NumPy arrays for further use in training machine learning models. The combined dataset is split into training, validation, and test sets. The split ensures a balanced distribution of classes in each set.

4.2 Explanatory Data Analysis

When the data was prepared, we thoroughly examined it using Explanatory Data Analysis (EDA) to find patterns. It investigates the data and statistics of a dataset that includes characteristics such as patient ID, age, gender, diagnosis, and whether the cancerous growth is benign or malignant. The dataset contains information about lesions on the skin. Visualizations are created to understand the distribution of gender, age, and distribution of melanoma lesions. The number of men and women for each type of cancer is displayed in Figure 1. The gender and age distributions from the original dataset are displayed in Figure 2 and Figure 3, respectively.

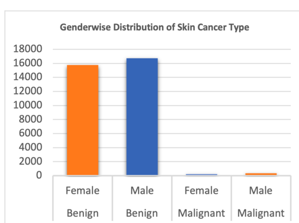

Figure 2: Gender Distribution by skin cancer type

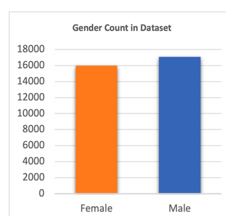

Figure 3: Gender Distribution in the dataset

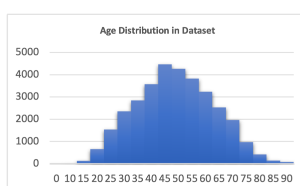

Figure 4: Age Distribution in the dataset

Figure 5: Images with skin cancer type Benign

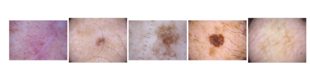

Figure 6: Images with skin cancer type Malignant

4.3 Model Building

Based on the Literature survey, we trained the data on various models that performed well for classifying skin cancer. For this, we trained a baseline Convolutional Neural Network and also used Transfer Learning with pre-trained models namely: VGG16, InceptionResnetV2, and Resnet50.

- Convolutional Neural Network: Convolutional neural networks (CNNs), a type of Deep Learning model, have outperformed conventional methods in many domains, most notably image and feature recognition, see Humayun et al. (2022). Additionally, they have been used successfully in the medical field, producing amazing outcomes and excellent performance in a range of difficult circumstances, see Gouda et al. (2022).
- VGG 16: One of the main advantages of using transfer learning with VGG16 for skin cancer classification is its ability to classify skin cancer with high accuracy from comparatively little data. This is because the network is already capable of identifying a wide range of characteristics, including texture, colour, and form, that are important for classifying skin cancer. Utilizing transfer learning in conjunction with VGG16 also has the benefit of preventing over-fitting, see Ibrahim et al. (2023).
- InceptionResnetV2: When it comes to object recognition, InceptionResnetV2 outperforms all other cutting-edge CNN models, exhibiting low computational overhead and excellent accuracy, see (Ahmadi Mehr & Ameri, 2022).
- ResNet50: Resnet-50 helps to improve general competence and prevent over-fitting, see Alwakid et al. (2022). The residual structure can address the issues of vanishing gradient explosion and network degradation, while the ResNet-50 network can converge more quickly, see Xue et al. (2021).

The models mentioned above were trained with an early stopping condition by employing a high number of training epochs and discontinuing training as soon as the model's performance on the validation Dataset no longer improved.

4.4 Model Analysis

Four metrics were used to examine the models once they had been trained: accuracy, AUC-ROC, recall, and loss on the validation set. The four metrics listed above were selected following the problem statement and dataset. Accuracy indicates the overall frequency of correct classification ML models. However, this measure is not sensitive, particularly when dealing with unbalanced datasets. Hence, Recall was employed to enhance the examination of the models' comparative performance. Recall provides the ratio of accurate "positive" predictions to the real number of "positive" cases—whether accurate or inaccurate), see (Bechelli & Delhommelle, 2022). Since the ROC curve depends only on sensitivity and specificity and is not influenced by prevalence, samples can be obtained regardless of the disease's prevalence in the community. A higher ROC value indicates that there is a low likelihood of the model misdiagnosing cancer, see Nahm (2022). A loss function uses a given set of variables to determine how well the model predicts the future.

4.5 Credibility Assessment with SHAP

The black-box nature of AI-based systems produces outstanding outcomes but without any justification. The proposed approach's conclusions are clarified and improved by the use of XAI techniques, such as the SHAP framework, which is based on SHapley values. SHapley values are useful in discovering and explaining asymmetry because they display the true contribution of each variable to the model prediction. It presents each variable's true influence, leading to a substantial reconstruction error, see Walia et al. (2022). The most optimal model was selected using the model analysis and SHAP was applied to that model to find features in the images that correspond to the model's prediction. This enables medical professionals to identify skin cancer early on.

5 Results and Discussion

The study focuses on using Explainable AI to emphasize the features influencing the decision-making process on the image and produce concise, comprehensible predictions that demystify the model's decision-making process. When attempting to determine which model performs the best, we take into account both the learning curves (loss, accuracy, recall, AUC-ROC) and confusion matrix metrics' (accuracy, precision, recall, F1 score). The best model will be the one that shows learning behaviour across epochs in addition to effective performance metrics.

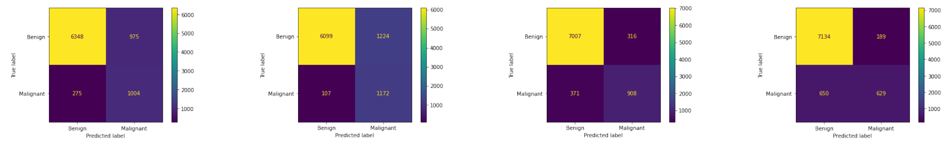

Figure 7: Confusion Matrices CNN, VGG16, InceptionResNetv2, ResNet50

Analyzing the confusion matrix for each of the four models in Figure 7 indicates that ResNet50 performs better in accurately predicting positive (malignant) instances, with a maximum precision of 76.89%. The F1 Score of InceptionResNetV2 is higher at 72.54%.

Table 1: Model Performance: Accuracy and Loss

	Accuracy		Loss	
	Train	Validation	Train	Validation
CNN	0.9248	0.9170	0.2394	0.2637
VGG-16	0.9744	0.9364	0.0737	0.1946
InceptionResNetV2	0.9669	0.9252	0.0902	0.2374
ResNet50	0.9088	0.9021	0.2481	0.2580

Table 2: Model Performance: AUC-ROC and Recall

	AUC-ROC		Recall	
	Train	Validation	Train	Validation
CNN	0.9252	0.9193	0.5926	0.5387
VGG-16	0.9891	0.9473	0.8804	0.8241
InceptionResNetV2	0.9865	0.9193	0.8375	0.6560
ResNet50	0.8832	0.9109	0.4923	0.3855

Table 1 and Table 2 display each model's learning curve. When all the criteria are taken into consideration, ResNet50 performs the best, according to the statistics shown in Table 1 and Table 2. ResNet performs best when comparing the outputs for the validation set for the AUC-ROC score, indicating that it can function better on imbalanced datasets, which is usually the case. While the VGG16 and InceptionResNetV2 exhibit over-fitting, the Convolutional Neural Network (CNN) Model displays less stable learning curves, as seen in Figure 8, Figure 9, Figure 10 and Figure 11.

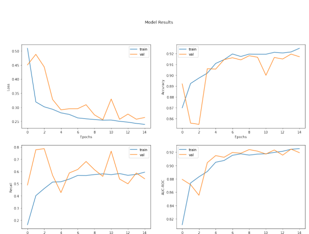

Figure 8: CNN

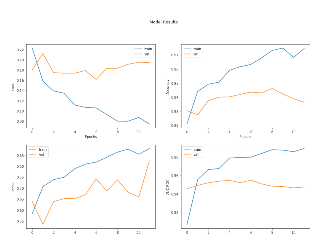

Figure 9: VGG16

Based on the combined assessment, ResNet50 might be considered the best model. Although it doesn't have the best individual scores in every category, it exhibits consistent learning without

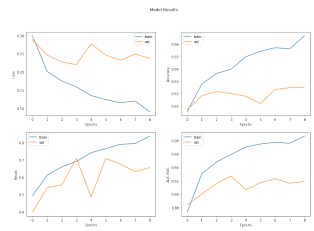

Figure 10: InceptionResNetv2

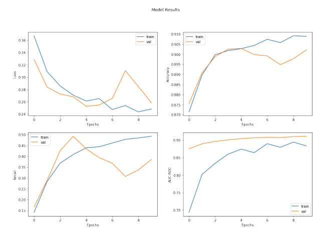

Figure 11: ResNet50

over-fitting and has an ideal balance across all assessed measures. This balance is essential in real-world applications where both the ability to generalize and the ability to correctly classify cases as shown by confusion matrix metrics are important.

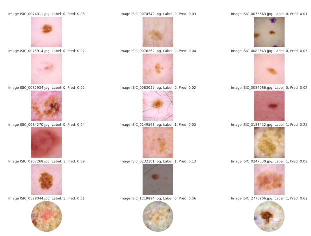

Figure 12: Predictions from the model ResNet50

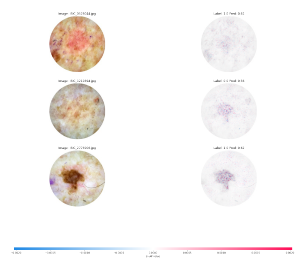

Figure 13: SHAP highlighting the features affecting the model's decision

After selecting the best model, we used SHapley Additive explanations (SHAP) to highlight the most significant features of the images. The predictions made by ResNet50 for every image are displayed in Figure 12. These images and predictions are then used as input to the SHAP model, which suggests features influencing the model's decision-making process. The impact level of the feature is indicated by the colours pink and blue. Each pixel's contribution to the image's classification into a particular class increases with its pinkness, whereas its contribution to the image's non-classification increases with its blueness.

6 Conclusion

The results obtained for skin disease classification using the SIIM-ISIC dataset indicated that ResNet50 outperformed other prominent models such as CNN, InceptionResNetv2, and VGG16. SHAP (SHapley Additive exPlanations) for model interpretation provided useful insights into the ResNet50 decision-making process. The model's information aids medical professionals in their patient prognosis, serving as a reference point based on the model's credibility.

7 Future Work

- The system can be used for continuous updating and monitoring of the model based on incoming data to assure its relevance and correctness over time.
- More advanced data augmentation techniques can be used to increase the size of the dataset which can help to improve model generalization by exposing it to a wider range of input variations.
- Ensemble methods can be used to enhance overall model performance.

Author Contributions

- Khushi Naik(NetID:kknaik, 1150203947): CNN, InceptionResNetv2, Report(Related Work, Formatting)

- Bharati Jagdish Panigrahi(NetID:bpanigra, 1342997035): ResNet50, VGG16, Report(Methodology, Results)
- Chetan Sah(NetID:csah, 1486289026): Pre-processing, SHAP, Report(Conclusion, Future Work)
- Tanaya Tamhankar(NetID:ttamhank, 1611549003): Data Collection, EDA, Report(Introduction, Scope)
- Equal Contribution for Poster Presentation
- Each member reviewed five articles on Skin Cancer Detection using Deep Learning.

REFERENCES

SIIM-ISIC Melanoma Classification Competition. Kaggle, 2023. URL `https://www.kaggle.com/competitions/siim-isic-melanoma-classification`.

N. Abhvankar, H. Pingulkar, K. Chindarkar, and A. P. I. Siddavatam. Detection of melanoma and non-melanoma type of skin cancer using cnn and resnet. *Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1–6, 2021.

Reza Ahmadi Mehr and Ali Ameri. Skin cancer detection based on deep learning. *Journal of biomedical physics & engineering*, 12(6):559–568, 2022.

Ghadah Alwakid et al. Melanoma detection using deep learning-based classifications. *Healthcare*, 10(12), 2022.

V Anand, S Gupta, A Altameem, S R Nayak, R C Poonia, and A K Saudagar. An enhanced transfer learning based classification for diagnosis of skin cancer. *Diagnostics*, 12(7), 2022.

Solene Bechelli and Jerome Delhommelle. Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images. *Bioengineering*, 9(3), 2022.

A. Budhiman, S. Suyanto, and A. Arifianto. Melanoma cancer classification using resnet with data augmentation. *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 17–20, 2019.

A. Demir, F. Yilmaz, and O. Kose. Early detection of skin cancer using deep learning architectures: resnet-101 and inception-v3. *2019 Medical Technologies Congress (TIPTEKNO)*, pp. 1–4, 2019.

W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi. Detection of skin cancer based on skin lesion images using deep learning. *Healthcare*, 10(7), 2022.

Mohammed Rakeibul Hasan, Mohammed Ishraaf Fatemi, Mohammad Monirujjaman Khan, Manjit Kaur, and Atef Zaguia. Comparative analysis of skin cancer (benign vs. malignant) detection using convolutional neural networks. *Journal of Healthcare Engineering*, pp. 17 pages, 2021.

Mamoona Humayun, R. Sujatha, Saleh Naif Almuayqil, and N. Z. Jhanjhi. A transfer learning approach with a convolutional neural network for the classification of lung carcinoma. *Healthcare*, 10, 2022.

A. Ibrahim, M. Elbasheir, S. Badawi, A. Mohammed, and A. Alalmin. Skin cancer classification using transfer learning by vgg16 architecture (case study on kaggle dataset). *Journal of Intelligent Learning Systems and Applications*, 15, 2023.

S Jain, U Singhania, B Tripathy, E A Nasr, M K Aboudaif, and A K Kamrani. Deep learning-based transfer learning for classification of skin cancer. *Sensors (Basel)*, 2021.

Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3:740–770, 2021.

D. C. Malo, M. M. Rahman, J. Mahbub, and M. M. Khan. Skin cancer detection using convolutional neural network. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0169–0176, 2022.

Francis Sahngun Nahm. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean journal of anesthesiology*, 75(1):25–36, 2022.

N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo. A deep learning approach based on explainable artificial intelligence for skin lesion classification. *IEEE Access*, 10:113715–113725, 2022.

Andre G. C. Pacheco and Renato A. Krohling. Recent advances in deep learning applied to skin cancer detection. 2019.

Alavikunhu Panthakkan, S.M. Anzar, Sangeetha Jamal, and Wathiq Mansoor. Concatenated xception-resnet50 — a novel hybrid approach for accurate skin cancer prediction. *Computers in Biology and Medicine*, 150, 2022.

Mayar Mohamed Swilam. Keras siim-isic melanoma classification resnet. Kaggle, 2023. URL `https://www.kaggle.com/code/mayarmohamedswilam/keras-siim-isic-melanoma-classification-resnet`.

Savita Walia, Krishan Kumar, Saurabh Agarwal, and Hyunsung Kim. Using xai for deep learning-based image manipulation detection with shapley additive explanation. *Symmetry*, 14, 2022.

Garrett Williams. Vgg16 model. Kaggle, 2023. URL `https://www.kaggle.com/code/garrettwilliams90/vgg16-model`.

Jiaqi Xue, Chentian Ma, Li Li, and Xuan Wen. Multiple effnet/resnet architectures for melanoma classification. *2021 International Conference on Computer Engineering and Application (ICCEA)*, 2021.