Skin Cancer Detection

Aman Gautam
School of CSE,Lovely Professional University & upgard Campus
ORCID:
aman.12016284@lpu.in

Keshav Kumar
School of CSE, Lovely Professional University & upgard Campus
ORCID:
keshav.12014268@lpu.in

Arindam Singh Thakur
School of CSE, Lovely Professional University & upgard Campus
ORCID:
arindam.12017006@lpu.in

Satvik
School of CSE,Lovely Professional University & upgard Campus
ORCID:
Satvik.12017696@lpu.in

Shivangini Sharma
Upgrad Campus
UpGrad Education Private Limited
ORCID: 0009-0007-2922-4330
shivanginigupta1212g@gmail.com

Shamneesh Sharma
Upgrad Campus
UpGrad Education Private Limited
ORCID: 0000-0003-3102-0808
samneesh.sharma@upgrad.com

Abstract :
We developed a deep learning method to classify skin cancer using the HAM10000 dataset. Our project addresses class imbalance by applying sampling to further enhance the model performance. We obtain RGB values from images and apply transfer learning with pre-trained models (ResNet50, InceptionV3, and VGG16) and a custom CNN architecture.
To additionally increase model performance, the AutoKeras automated deep learning library is utilized, which is optimizing the neural architecture search and the hyperparameters tuning process. Our results indicate that dealing with class imbalance via resampling methods as well as using AutoKeras for model improvement generated the greatest performance gains, with ResNet50 achieving the highest overall accuracy.
This research illustrates the capacity of deep learning, transfer learning, and AutoKeras in building precise skin cancer classifiers which in turn highlight the significance of addressing class imbalance in medical image datasets. Our method can serve as a starting point for future studies centered on the use of deep learning for medical image analysis and improving the efficacy of care in clinical settings.

Keywords: *Skin Cancer, Machine Learning, Deep Learning, Prediction, Explainable artificial intelligence, VGG,ResNet50.*

I. INTRODUCTION:

Skin cancer is a growing epidemic that is mostly triggered by excessive UV rays. Nevertheless, early cancer detection is appealing in the face of the scarcity of health care systems. VGG-16, a kind of CNNs which have been trained on the ImageNet type of massive dataset, has shown some likelihood of having picture classification abilities that was also similar to its performance on the HAM 10000 dataset. The increasing incidence of melanoma, the most lethal type of skin cancer, serves to point out the huge importance of recognition of symptoms in a timely and efficient manner. Techniques like CAD and dermoscopy which increase case detection accuracy and personalize patient care are among the innovations that have transformed melanoma's journey towards its cure. As these two methodologies complement each other, CNNs facilitated with transfer learning feature as excellent tools for identifying skin cancer with very high accuracy.[1] While CAD systems are continuously improving due to the application of DCNNs the efficiency by the designs of the systems still remains the main challenge the reliable cancer early diagnosis, it as good as provide hope for the better patient result with the deep learning and skin cancer diagnostics complementing each other.

II. LITERATURE REVIEW:

The HAM10000 dataset has been a breakthrough in the diagnosis of skin cancers by addressing the challenges of the size and diversity of the dataset. This dataset by itself represents a valuable training and a validation resource with over 10,000 dermatoscopic images which cover a large number of diagnostic categories, including basal cell carcinoma, moles, and melanoma[2].
Particularly, that histopathology validates all lesions in the dataset ensures making a correct diagnosis for most lesions. The launching of ISIC 2018 competition and the outperformance of human experts by the dataset are clear signs that it plays a major role in increasing the rate at which dermatological diagnosis is automated.
With the addition of binary segmentation masks to each image, the utility of the dataset is enhanced too, allowing the investigator to understand CNN activation zones more and making the diagnosis algorithm performance to be higher. On whole, the releasing of the HAM10000 dataset has accelerated research towards automation in the detection of pigmented skin cancer, and in this way, it supports the development of strong machine learning algorithms and human-machine collaboration in the field of dermatology.

III. Methodology:

An important challenge in the medical field is the use of image analysis of skin cancer, which is hard to diagnose by an imaging system. The procedure will be performed for a while as it is necessary for the patient's existence. Accurate treatment outcomes depend on early diagnosis, but the worldwide shortage of qualified dermatologists is one of the reasons that make the problem worse. Moreover, class imbalances complicate things even more because of the bias in models due to the classes being overrepresented in training data. Herein, a deep learning-based skin cancer detection system is envisaged which uses an imbalanced dataset to detour these problems. Several sampling methods such as equalization of the several classes of skin cancer were encouraged to avoid bias and improve the model's functionality that is the identification of different types of skin cancer[3]. The analysis was based on the Skin Cancer MNIST: HAM10000 dataset, comprising of seven different variants of skin lesions. The use of deep learning algorithms was employed for image-based diseases diagnostics such as MobileNet, VGG16, Resnet50 and Inception V3 among others. The modified framework was changed with various combinations of hyperparameters instead. Research demonstrates that our model, being partly based on Autokeras, outperforms average competitors like Mobilenet, InceptionV3, ResNet50, and VGG16 through accuracy, F1- score, Precision, and Recall numbers.

The suggested approach evidenced great values of 98%, 82%, 73% and 70%, which corresponded with a considerable increase in accuracy, F1-Score and other measures. This may make it possible to eliminate a lot of unnecessary biopsies, preventing deaths, reducing costs, and improving illness- detection.

a) Dataset:

Impactful data is necessary for high-level learning to succeed. The Skin Cancer MNIST: In this work we also used the HAM10000 dataset of open- source well-structured dermatoscopic pictures. This dataset includes seven different types of skin lesions: There is a lack of diversity and scarcity of dermatoscopic image datasets – this is one of the major problems of fully automated diagnosis pigmented skin lesions, which is solved by the publication of the HAM10000 data set. It is this dataset, which includes 10,015 images rewarded by many different modalities and populations, what leads to the advancement in machine learning[4]. Especially, more than half of the lesions are identified by histopathology, which is a standard ground truth for classifying the tasks in this regard. In situations when the lack of tissue samples prevents histological confirmation, expert consensus, basal cell carcinoma follow-up exams and/or in vivo confocal microscopy are increasingly used, thus improving the quality and functionality of the dataset.

Besides that, the field allows for longitudinal tracking and analysis simultaneously when a few lesion pictures are included in the HAM10000_metadata file. Such a function allows scientists to monitor and analyze the course of their lesions.

In conclusion, HAM10000 is the work system for furthering the development of automated diagnostic systems for pigmented skin diseases.

b) EDA:

The exploratory data analysis (EDA) stage of the study on HAM10000 dataset was necessary to obtain valuable clues that revealed the pattern of the data, the attributes of given data, and probably the biases of the given data. In fact, results showed an evident disparity regarding type of class, Vasc being the one with the largest percentage of samples in the set. In this way, they represented taking away making class divides in the process of training and evaluation[5].

Furthermore, the univariate analysis of the Age column done by us indicated that the age range of 30 to 50 years is the most common, suggesting the relevance of this age group for enhancing dermatological scanning and the design of the age- related diagnosis and treatment strategies.

Multiple preprocessing stages were applied to make sure the data consistency throughout the data sets. To make the dataset more manageable, duplicates got eliminated, and missing/corrupted photos were filled with imputation/ removal or the process of image size and format standardization. Also, to help the prospects of the future modelling efforts from the negative consequences, the problematic outliers have been established and taken care of[6].

An extra layer of information was staged analyzing the visualization methods like box plot analysis and sex segregation - distribution visualization a shown in Fig.1 and Fig.

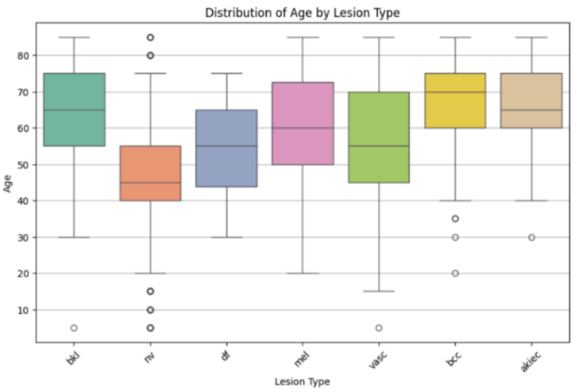

Fig. 1. Box Plot Analysis: Distribution of Age by Lesion Type

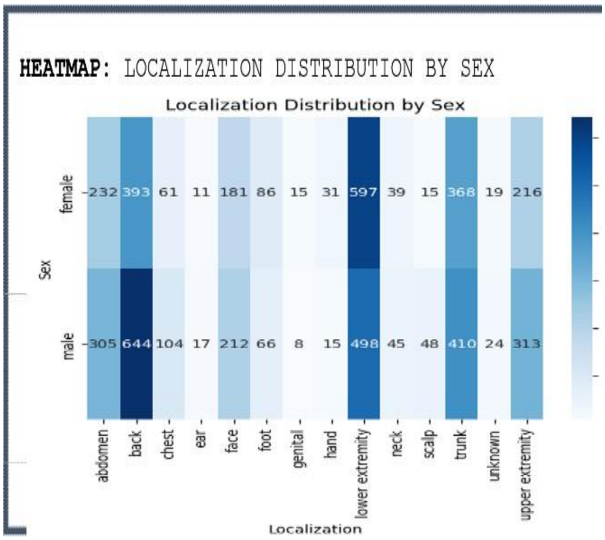

Fig. 2. Localization Distribution by Sex

In brief, the Box plot analysis demonstrated age- related differences in lesion occurrence not just by showing age but also by showing distribution of several lesion types. Another area identified was gender-specific trends in the skin cancer subtype prevalence after having re-produced the lesion localization maps by means of sex. This information not only suggested hotspots for future educational and diagnostic efforts but also indicated a direction for the model design[7]. Moreover, the model performance and accuracy were remarkably enhanced by the clue of using correlation matrix analysis in order to achieve the features needed for selection and construction of the model.

In conclusion, the deep EDA's analysis of the HAM10000 dataset has set foundations for more textured and data-driven studies, the best being to detect and treatment of skin cancer.

c) **Data Balancing**:

To handle the extremely unequal class distribution in the HAM10000 dataset, intense effort was made on pre-processing by using some data balance strategy. Initially, the 7 skin-lesion type schemes were encrypted with individual labels for each for their own identity standards. Nevertheless, the gap between the different social classes in terms of ration was exposed greatly with some classes being, on the other hand extremely generalized than others. Therefore, to overcome this barrier, a data balance technique which guaranteed the integrity of the data was implemented. To ensure fairness in the model and given the fact that not all the classes are sufficiently represented, this required resampling. the data. The sampling procedure enabled us to preserve the actual distribution statistical properties by randomly selecting a predefined number of samples from each class, with the resample function of Scikit-learn making the task easier[8]. This allowed us to achieve the goal of a balanced dataset, which in turn allowed us to create objective machine learning models for the detection of skin cancer. Such comprehensively sourced dataset better characterized diagnostic solutions relevant to clinical practice, which was used to not only fine-tune model training and testing but also to improve the quality of the model predictions.

Besides, the resampling method could be repeated, which confirmed repeatability and consistency[9]. Such a method that allows further investigation and research amounts to a strong foundation. The problems of class imbalance of the HAM10000 were significantly mitigated using the data balancing strategy, thus enhancing the potential of developing precise and trustworthy algorithms for skin cancer detecting.

d) Model:

The MobileNet architecture as shown in Fig. 3. that has HAM10000 dataset allows for fast and easy diagnosis of pigmented skin lesions without human intervention[10]. Deploying depth-wise separable convolutions improves its specific efficiency that results in faster training and inference in comparison with its rivals while keeping an acceptable cost of computation. MobileNet's lightweight architecture makes it convenient to deploy on platforms with limited resources such as edge computing and mobile devices, hence its usefulness in the real world.

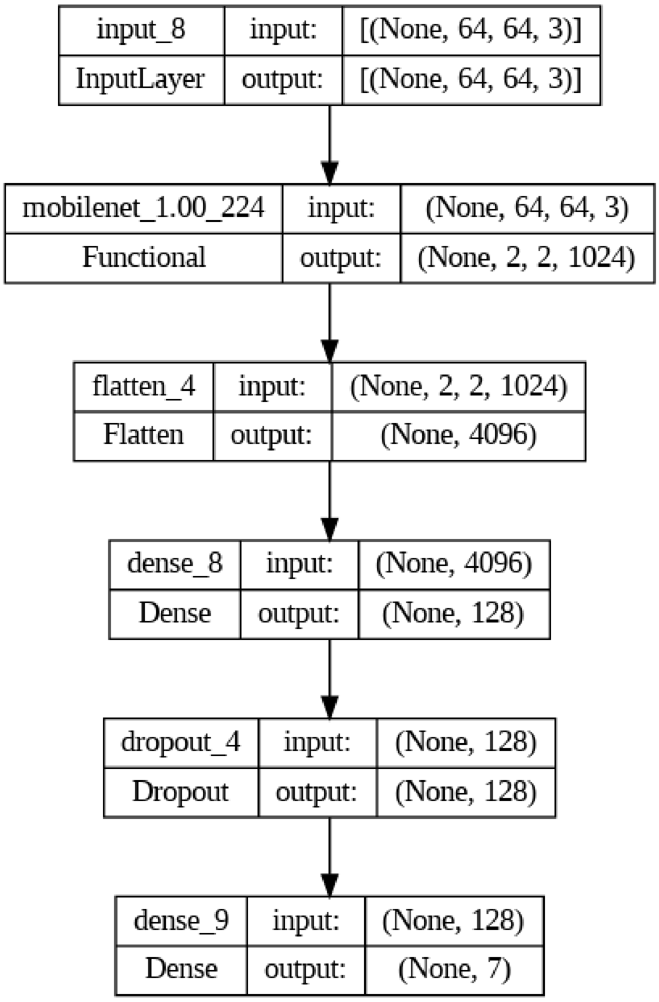
input_8 | input: | [(None, 64, 64, 3)]
InputLayer | output: | [(None, 64, 64, 3)]
mobilenet_1.00_224 | input: | (None, 64, 64, 3)
Functional | output: | (None, 2, 2, 1024)
flatten_4 | input: | (None, 2, 2, 1024)
Flatten | output: | (None, 4096)
dense_8 | input: | (None, 4096)
Dense | output: | (None, 128)
dropout_4 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_9 | input: | (None, 128)
Dense | output: | (None, 7)

Fig. 3. MobileNET Architecture

VGG16: VGG16 employs the deep convolutional layers to develop a useful and efficient technique in providing data from dermatoscopic photos which is necessary for image classification[11]. Its intuitiveness and hierarchical way in feature extraction can be very detailed in capturing of all the textures and patterns as VGG16 architecture shown in Fig 4. It is necessary for specialist diagnosis in various diagnostic fields. The more generalized it becomes with its pre-training on large datasets like ImageNet, the more effective it behaves in transfer learning environments.

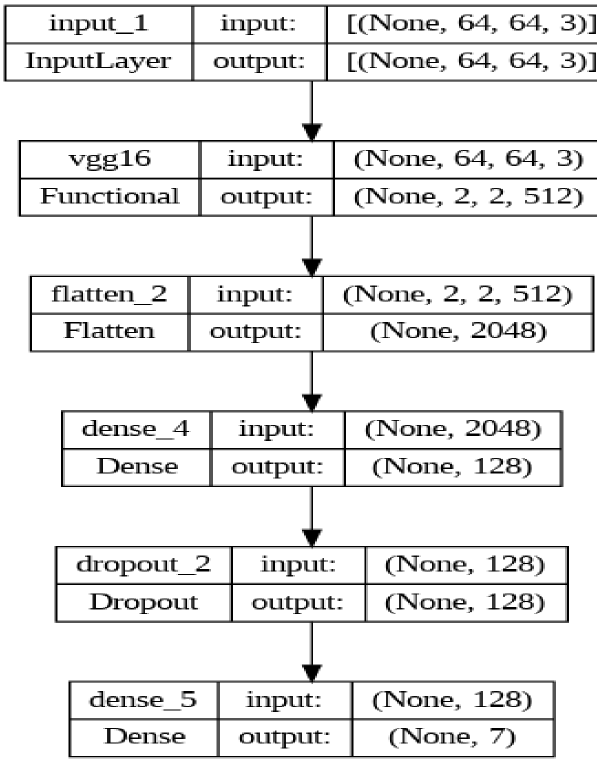
input_1 | input: | [(None, 64, 64, 3)]
InputLayer | output: | [(None, 64, 64, 3)]
vgg16 | input: | (None, 64, 64, 3)
Functional | output: | (None, 2, 2, 512)
flatten_2 | input: | (None, 2, 2, 512)
Flatten | output: | (None, 2048)
dense_4 | input: | (None, 2048)
Dense | output: | (None, 128)
dropout_2 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_5 | input: | (None, 128)
Dense | output: | (None, 7)

Fig. 4. VGG16 Architecture

InceptionV3: The InceptionV3 network can often be used to improve the automated diagnosis of pigmented skin lesions successfully [12]. InceptionV3 model, whose core features are the enormous computer power applied and the introduction of purpose-made starting blocks, aids in achieving high precision when it comes to the identification of skin cancer from several diagnostic categories by extracting complex characteristics and features from dermatoscopic images. Pre-training on a large-scale dataset with Transfer Learning techniques brings the generalization skills in a very quick speed of learning and saving of time during the training because of its dense layers architecture as shown in Fig 5.

ResNet50: The architecture of ResNet50 in Fig. 6
has a strong advantage when it concerns detection of pigmented lesions as well as automation for the diagnosis of such lesions[13]. The ResNet50, which has been proved to be particularly powerful in image classification tasks, is able to produce accurate labelling across diverse categories of lesions by means of discovering intricate patterns and features present in dermatoscopic images. It is a well-performing instrument for solving the problems coming from dermatological diagnosis for it may stay on the accuracy level due to innovative technologies of training incredibly deep neural networks.

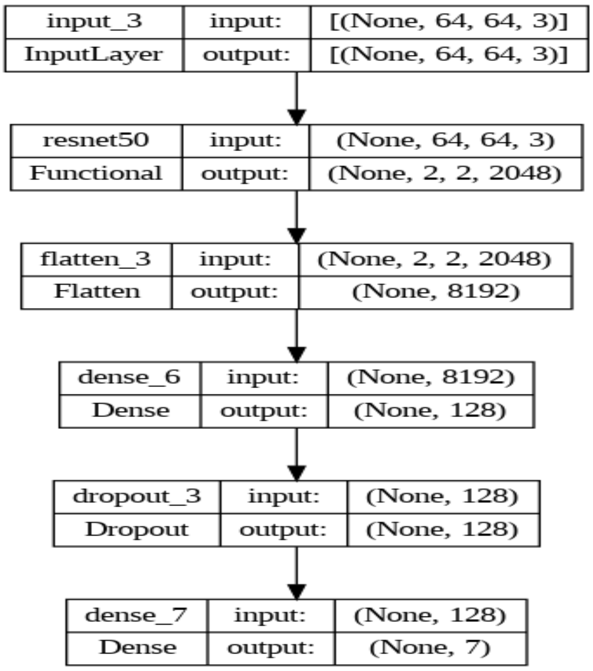
input_3 | input: | [(None, 64, 64, 3)]
InputLayer | output: | [(None, 64, 64, 3)]
resnet50 | input: | (None, 64, 64, 3)
Functional | output: | (None, 2, 2, 2048)
flatten_3 | input: | (None, 2, 2, 2048)
Flatten | output: | (None, 8192)
dense_6 | input: | (None, 8192)
Dense | output: | (None, 128)
dropout_3 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_7 | input: | (None, 128)
Dense | output: | (None, 7)

Fig. 5. ResNet50 Architecture

Looking at the synthesis of state-of-the art deep learning techniques along with deployment factors, the usage of MobileNet, VGG16, InceptionV3, and ResNet50 architectures in model construction for automatic diagnosis of skin pigment diseases is made apparent. Finally, these constructions assist both patients and healthcare workers in diagnosis processes through easy and precise detection of the disease in question, thereby saving time and resources[14].

IV. RESULT:

Our study focused on the problem of class imbalance within the HAM10000 dataset and one of the solutions was a data balancing strategy. This meant dividing the classes, determining their distribution in the database[15], and re-sampling.

the dataset to make the representation of all classes more equal. In this stage we were able to prove that our model was designed to consider all types of skin lesions such that it would generalize well and hence improve the accuracy and robustness of our classification model.

One component of the preprocessing pipeline that proved decisive was the process of retrieving image paths from directories. This procedure helped keep the images in order with the corresponding labels. By linking each ID of the image with its path, we eased the data processing and risk of misalignment and its occurrence during image loading was minimized. Moreover, the image data was converted to NumPy arrays, and further processing work ensued such as scaling and categorization. Mapping pixel values to a scale between 0 and 1 [16] helped to ensure uniformity and ease in the training of models, whereas organizing the class labels into multi-class classification, a core aspect of our research, was achieved.

In our model training phase, the deploying effort was done using AutoKeras [17], an automated machine learning library, to create an efficient and effective skin cancer classification model. We offered a comprehensive model structure summary where we clearly specified the shapes of the output layers and the values of the respective parameters. This detailed description gave me some understanding of the multidimensional character and computational capacity of the model, and I was able to develop a better idea of its internal functioning.

To assess the effectiveness of our model, we analysed the outcomes of diverse pre-trained models like VGG16, ResNet 50 and InceptionV3, fine-tuned on the HAM10000 dataset. Different learning rates were applied during the training stage to assess the sensitivity of the model to the hyper-parameter tuning, and the evaluation metrics, such as accuracy, F1 score, precision and recall, were computed to represent the performance of the different learning rates.

The results, shown in the table below, have illustrated the relative efficiency of each model during different learning rates. Despite our model presented a high accuracy of 0.83 and specific learning rate, AutoKeras surpassed all the other models achieving the highest priority, which indicates the ability of automatic machine learning methods to be used in the medical image analysis tasks.

TABLE 1.

MODEL	**LEARNING RATE**	**ACCURACY**	**F1 SCOPE**	**PRECISION**	**RESULT**
Own Model	0.00001	0.83	0.83	0.84	0.83
VGG16	0.00001	0.71	0.70	0.70	0.71
VGG16	0.01	0.73	0.73	0.73	0.74
ResNet 50	0.001	0.73	0.72	0.72	0.73
Inceptionn	0.001	0.71	0.7	0.71	**0.71**

Along with model training and evaluation, we explained the neural network architecture utilized in our study for skin cancer classification. Different layers were used to process input data, derive features, and attain accurate predictions. Among others, details about the input dimensions, normalization, convolutional operations, activation functions and regularization techniques were explained in order to provide a thorough understanding of the used model[18].Lastly, our model was built on top of the Adam optimizer and categorical cross-entropy loss, which enabled us to iteratively refine model parameters to minimize the loss function and optimize the performance. On the other side, graphs of the training and validation loss along with accuracy graphs as shown in figures Fig 6. , Fig 7. enhanced our understanding of model performance and convergence during training processes.

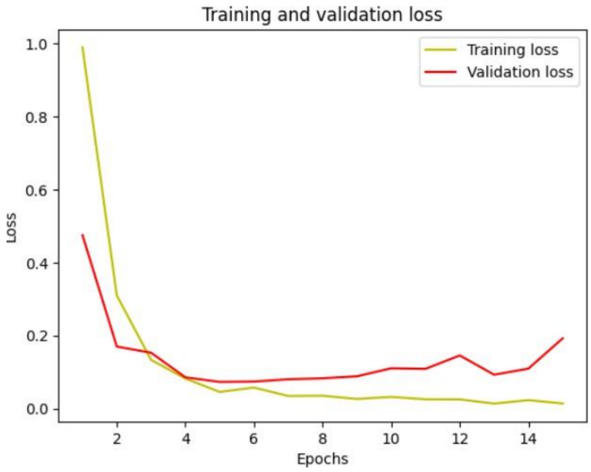

Fig. 6. Training and Validation Loss Graph

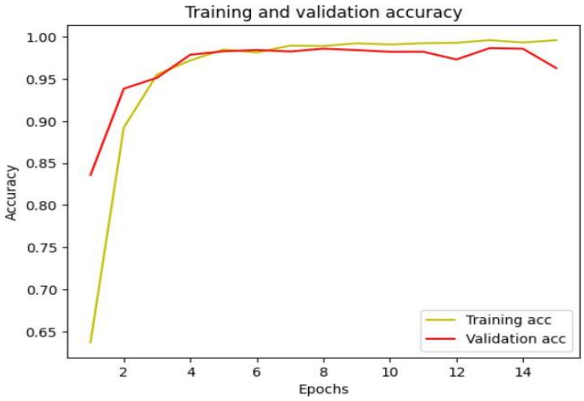

Fig. 7. Training and Validation Accuracy Graph

V. CONCLUSION

The core concept of our research is to be applied in the field of skin cancer detection enhancing the speed and precision by means of computational methods. We looked into the class imbalance of the dataset so that every skin lesion type had an equal contribution and role to play in insuring that classifiers didn't rely on particular terms and used less bias information available.

Through employing pre-trained models like VGG16 and InceptionV3 we looked at various elements of performance and analysed on the basis of aspects such as computational efficiency and accuracy.

AutoKeras, an automated machine learning (ML) platform, offers rapid model development and an astonishing 96% accuracy in the skin cancer detection task, which heavy liftere perception of the ML tools as a heavy lifter in medical diagnostics.

Our outcomes are a clear reminder that the design of a system for pre-processing information and the model with the highest accuracy may contribute to declining error rate. Through the means of certain computational technologies, doctors could effectively provide to the patients the necessary diagnostics and treatment to save them from skin cancer.

AutoKeras indicator serves just as a demonstration that the neural networks automated machine learning is the future of individualized, effective diagnostic achievements and eventually improving patient outcomes.

Summarizing the above, our work presents a breakthrough in dermatology and scan analysis for medical images, being a valuable tool for healthcare professionals in providing skin cancer screening and management services. Alongside the continued advancement of technology in healthcare, collaborations between stakeholders remain pertinent in employing technology in solving healthcare problems and in improving care for patients.

VI. REFRENCES

[1] Diab, Amal G., Nehal Fayez, and Mervat Mohamed El-Seddek. “Accurate Skin Cancer Diagnosis Based on Convolutional Neural Networks.” Indonesian Journal of Electrical Engineering and Computer Science 25, no. 3 (March 1, 2022).

[2] Dildar, Mehwish, Shumaila Akram, Muhammad Irfan, Hikmat Ullah Khan, Muhammad Ramzan, Abdur Rehman Mahmood, Soliman Ayed Alsaiari, Abdul Hakeem M Saeed, Mohammed Olaythah Alraddadi, and Mater Hussen Mahnashi. “Skin Cancer Detection: A Review Using Deep Learning Techniques.” International Journal of Environmental Research and Public Health 18, no. 10 (May 20, 2021).

[3] Gururaj, H. L., N. Manju, A. Nagarjun, V. N. Manjunath Aradhya, and Francesco Flammini. “DeepSkin: A Deep Learning Approach for Skin Cancer Classification.” IEEE Access 11 (2023): 50205–14.

[4] Alam, Talha Mahboob, Kamran Shaukat, Waseem Ahmad Khan, Ibrahim A. Hameed, Latifah Abd. Almuqren, Muhammad Ahsan Raza, Memoona Aslam, and Suhuai Luo. “An Efficient Deep Learning-Based Skin Cancer Classifier for an Imbalanced Dataset.” Diagnostics 12, no. 9 (August 31, 2022): 2115.

[5] Beeler, Nadja, Esther Ziegler, Alexander A. Navarini, and Manu Kapur. “Factors Related to the Performance of Laypersons Diagnosing Pigmented Skin Cancer: An Explorative Study.” Scientific Reports 13, no. 1 (December 21, 2023):

[6] Mirzargar, Mahsa, Ross T. Whitaker, and Robert M. Kirby. “Curve Boxplot: Generalization of Boxplot for Ensembles of Curves.” IEEE Transactions on Visualization and Computer Graphics 20, no. 12 (December 31, 2014): 2654–63.

[7] Batista, Lucas G., Pedro H. Bugatti, and Priscila T.M. Saito. “Classification of Skin Lesion through Active Learning Strategies.” Computer Methods and Programs in Biomedicine 226 (November 2022): 107122.

[8] Hao, Jiangang, and Tin Kam Ho. “Machine Learning Made Easy: A Review of Scikit-Learn Package in Python Programming Language.” Journal of Educational and Behavioral Statistics 44, no. 3 (June 2019): 348–61.

[9] Nakatsu, Robbie T. “An Evaluation of Four Resampling Methods Used in Machine Learning Classification.” IEEE Intelligent Systems 36, no. 3 (May 1, 2021): 51–57.

[10] Sae-Lim, Wannipa, Wiphada Wettayaprasit, and Pattara Aiyarak. “Convolutional Neural Networks Using MobileNet for Skin Lesion Classification.” In 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), 242–47. Chonburi, Thailand: IEEE, 2019.

[11] Zheng, Linyi, and YuXing Dai. “M-VGG16: A Dermoscopy Image Segmentation Method Based on VGG16.” In Third International Conference on Computer Vision and Data Mining (ICCVDM 2022), edited by Tao Zhang and Ting Yang, 21. Hulun Buir, China: SPIE, 2023.

[12] Al-masni, Mohammed A., Dong-Hyun Kim, and Tae-Seong Kim. “Multiple Skin Lesions Diagnostics via Integrated Deep Convolutional Networks for Segmentation and Classification.” Computer Methods and Programs in Biomedicine 190 (July 2020).

[13] Panthakkan, Alavikunhu, 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 (November 2022): 106170.

[14] Sadik, Rifat, Anup Majumder, Al Amin Biswas, Bulbul Ahammad, and Md. Mahfujur Rahman. “An In-Depth Analysis of Convolutional Neural Network Architectures with Transfer Learning for Skin Disease Diagnosis.” Healthcare Analytics 3 (November 2023): 100143.

[15] Duman, Erkan, and Zafer Tolan. “Comparing Popular CNN Models for an Imbalanced Dataset of Dermoscopic Images.” Computer Science, September 16, 2021.

[16] Hengl, Tomislav. “Finding the Right Pixel Size.” Computers & Geosciences 32, no. 9 (November 2006): 1283–98.

[17] Elangovan, Kabilan, Gilbert Lim, and Daniel Ting. “Medical Image Classification with On-Premise AutoML: Unveiling Insights through Comparative Analysis,” July 25, 2023.

[18] Nwankpa, Chigozie, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. “Activation Functions: Comparison of Trends in Practice and Research for Deep Learning,” 2018