Skin Lesion Classification using a Sequential Convolutional Neural Network and HAM10000

Austin Wort*, Mahdi Hassen*, Marcus Uy*, Nathan Vu*, Katrina Kezele*
*Faculty of Engineering, Toronto Metropolitan University, Toronto, Canada
{austin.wort, mahdi.hassen, marcus.uy, nam.vu, katrina.kezele}@torontomu.ca

***Abstract*—Skin cancer is a global health concern, with prevention, early detection, and accurate diagnosis playing an extremely crucial role in the prognosis of a patient. Typically, skin lesions are manually examined by dermatologists who use a set of techniques to diagnose skin lesions, however, this leaves the margin of error in the influence of experience and expertise, which could vary greatly. This paper presents a novel skin lesion classification application that allows users to input their images to receive an informed prediction of their diagnosis. This application is built on a custom sequential convolutional neural network (CNN) using the TensorFlow/Keras Python library(s), and trained using the open-source HAM10000 dataset, which consists of 10,000 dermatoscopic images of pigmented lesions, specifically actinic keratoses, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevi, melanoma, and vascular skin lesions [1]. Although this application is not to be used as a replacement for seeking a diagnosis from a medical professional, it is intended to help confirm diagnoses or alert the user to seek medical attention.**

***Index Terms*—Skin cancer, convolutional neural network (CNN), TensorFlow/Keras, HAM10000, dermatoscopic.**

I. Introduction

Skin cancer is one of the most common forms of cancer worldwide, however, the prevalence of skin cancers can be somewhat mitigated by avoiding certain risk factors (i.e., limiting skin exposure to high UV rays). However, if individuals are predisposed to skin cancers based on genetics or other factors, this type of cancer still holds to be one of the most optimal patient prognoses if diagnosed correctly and early on enough, prior to metastasis. The distinction between a benign lesion and a malignant one is pivotal for determining further courses of action and consequently, an effective and successful prognosis for the patient. Dermatologists are typically specialists who are responsible for the diagnosis of abnormal skin lesions, either through a standard ABCDE (Asymmetry, Border, Colour, Diameter, and Evolving) check, or analyzing dermoscopic images. This manual analysis, however, leaves an unnecessary error margin in correctly diagnosing on the basis of subjectivity and variability, leaving the patient at potential risk in terms of mortality. This volatile factor highlights the need for automated, consistently reliable diagnostic tools.

Within the past decade, promising advancements in the realm of computer vision (CV) and deep learning (DL) have demonstrated astute results in medical image analysis. Leveraging convolutional neural networks (CNNs), a type of deep learning, for analyzing dermatoscopic images, and Tensorflow, an open-source framework for deep learning, presents the potential to increase diagnostic accuracy, providing the essential tools for implementation, training, and fine-tuning such models. Thus, developing a CNN-based model using Tensorflow is an ideal choice for medical image testing. Furthermore, due to the growing interest in medical imaging, many skin lesion datasets exist today. As part of the 2018 ISIC Challenge, the HAM10000 dataset was released to garner friendly competition and push the boundaries of machine learning. This dataset contains 7 different classes: Dermatofibroma (df), Vascular lesions (vas), Benign keratosis-like lesions (bkl), Melanocytic nevi (nv), Actinic keratoses (akiec), Basal cell carcinoma (bcc), and Melanoma (mel). Of these 7, the latter 2 are among the most dangerous forms of skin cancer. Successful identification of these malignant types of skin lesions are therefore critical to timely and accurate diagnosis, enabling appropriate treatment and improving patient outcomes.

In this paper, a comprehensive approach is taken in developing a DL model based on the HAM10000 dataset for classifying skin lesions as one of 7 classes. Tensorflow provides the groundwork for building and training the model using its high-level API, Keras. Multiple machine learning and data analytics tools were employed to organize the dataset and help train the model, such as numpy, pandas and sci-kit learn. Finally, OpenCV was used to to execute data augmentation functions to further diversify and balance the dataset. By leveraging modern computer vision and deep learning tools, we were able to create a multi-layered sequential model that can categorize skin lesions with high accuracy, paving the way for more proactive skin cancer detection.

II. Methodology

The approach of this project began with selecting a sequential CNN as the model of choice, given that the application requires a single input, and produces a single output, and this particular CNN model provides ease of use and simplified debugging. Following this selection, the Adam optimization algorithm was additionally selected to refine and improve performance of the proposed model. This algorithm combines an adaptive gradient algorithm (AdaGrad) with root mean square propagation (RMSProp), and was chosen due to its efficacy and efficiency [2]. Although these decisions fit the application for the proposed architecture, the sequential CNN model, and any model for that matter, can only provide accurate medical predictions given that it is trained on high-quality, diverse data. The HAM10000 (Human Against Machine with

10000 training images) provides exactly that, with 10000 dermatoscopic images available for use in training DL models. It is important to note however, that while HAM10000 does include images spanning seven classifications, it is unevenly distributed, as shown in Figure 1. To reasonably balance the dataset, underrepresented classifications were extrapolated using data augmentation. After augmenting these classifications, the new dataset is much more balanced, as shown in Figure 2. The generalized steps of the proposed methodology can be seen as follows:

A. Model Development

The sequential CNN developed for this project was built from the ground up and trained in the same manner. While this approach can be somewhat more challenging and problematic, building and training the model from scratch in this manner actually serves our purpose better; Given the problem domain and dataset for this project, the proposed model must be able to detect and classify images from highly specific features that a generalized model derived using transfer learning—leveraging the weights of a pre-trained model to streamline retraining—might not capture effectively.

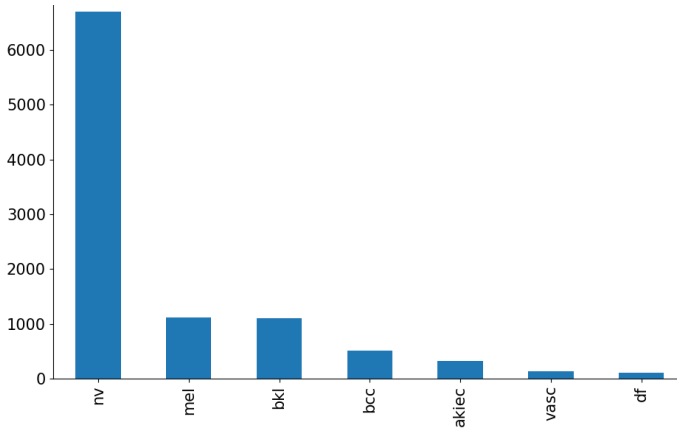

Fig. 1. Original HAM10000 Data

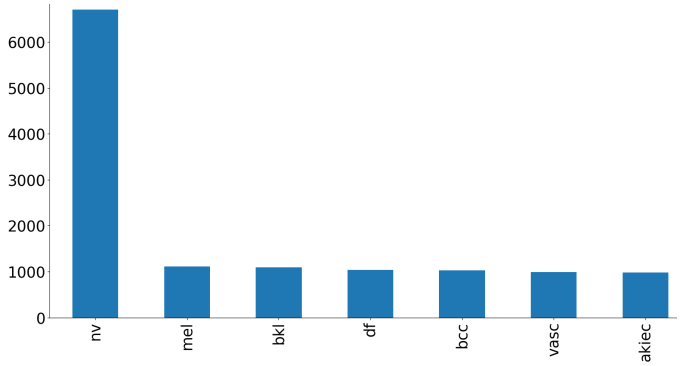

Fig. 2. Augmented HAM10000 Data

B. Dataset Preparation

Given that the provided images are all of high quality, have a standardized aspect ratio, and have been augmented to diversify and reduce bias in the sample data, the dataset is effectively prepared for training. However, data augmentation has a dual purpose in the training process. Given that the proposed model is relatively small, as is the dataset, it becomes extremely easy for the model to overfit during training, where it becomes increasingly familiarized with sample data, effectively memorizing it [3]. This affects the generalized performance of the model, leading to worse performance with unseen data. Augmenting the sample data works to prevent, or at least reduce the occurrence of overfitting, to provide a model that will perform better on generalized, unseen data.

C. Model Training

The sequential CNN model was systematically fed individual images from the preprocessed and normalized HAM10000 dataset. The training duration of the model was optimized to proceed until 25 epochs, at which point training and validation loss were both found to be minimized, given by Figure 3. After this point in training, the model begins to overfit and the validation loss increases as epochs progress.

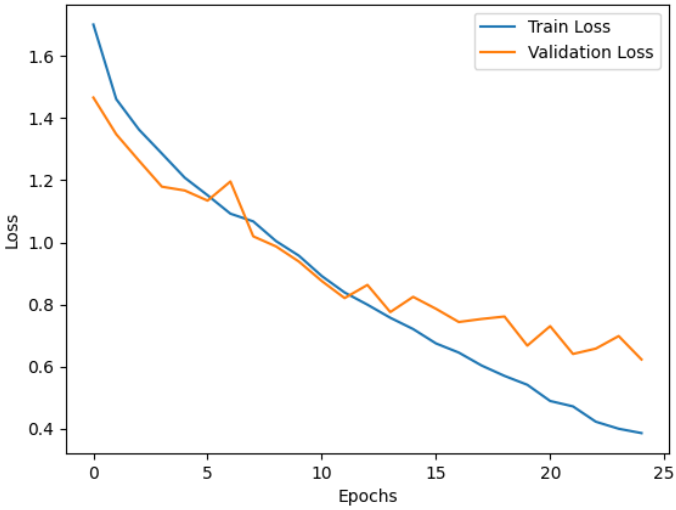

Fig. 3. Training and Validation Loss Plot

D. Model Deployment with User Interface

The final, trained classification model was imported into an executable, allowing users to input their own images for classification. The user interface presents the image to the user, along with a calculated confidence score and a predicted classification (akiec, bcc, bkl, df, nv, mel, or vasc) along with any other potential conditions linked to the determination.

III. Results

After fine-tuning training parameters and dataset metrics, the final, trained model was able to achieve a calculated accuracy of 79.52%, over 1200 samples, 25 epochs of training, and a batch size of 32 (Model accuracy was able to reach

a maximum of 83% using 40 epochs of training, however validation loss due to overfitting was far too great for this to be considered a generalized model). The finalized loss value was found to be 0.63. These results show the proposed model to be relatively accurate in its predictions and trained appropriately. Confidence scores calculated by the model to be presented to the user typically reflect a high-level of accuracy, displaying confidence scores of +90%, but do however suffer from bias arisen from imbalance in the dataset; Despite our best efforts to extrapolate underrepresented classifications through data augmentation, the model more commonly performs significantly worse on these underrepresented classes, reflected in the confidence score which can reach below 50% in some of our trials. Regardless of this however, the accuracy reflected in the majority of the classifications displays the efficacy of our model, and implies that, given sufficient data, classification performance and confidence scores can be significantly improved upon. The test and validation accuracy of the trained model can be seen in Figure 4.

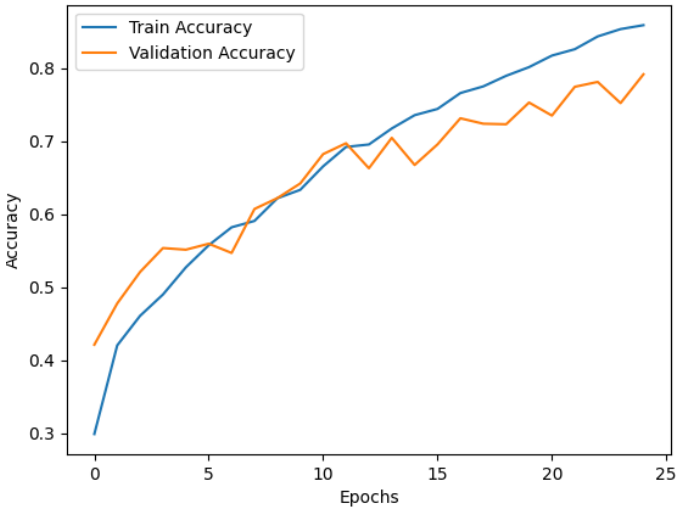

Fig. 4. Training and Validation Accuracy Plot

As discussed earlier, 25 epochs of training was determined to be the limit for training duration, as exceeding this would expose the model to overfitting as validation loss increasingly diverges upward from training loss. Minimizing this behaviour was crucial to ensuring that the application provided the user with a reasonably accurate classification for a given image.

Testing the functionality of the application, it was imperative that performance and confidence scores not be affected by inputting an image the model had been exposed to prior, thus the application was only tested using images that were unseen to the model—manually removed from the dataset and labeled for validation—to ensure that it would perform well on generalized data. An example of the application in use is given in Figure 5.

IV. Discussion

The performance of the application demonstrates the efficacy of the training process, providing users with clear and

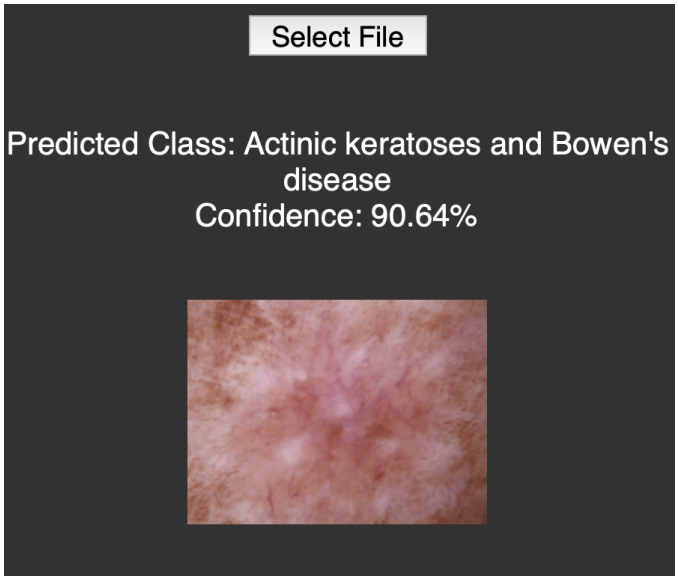

Fig. 5. Example Application Classification on Selected HAM10000 Image

interpretable results and a descriptive classification of the given image with a high-level of confidence. Providing users with this information allows them to make informed decisions about their health, and determine if seeking medical attention may be necessary or not. It is important to reiterate however, that this model is not intended to replace the opinion of a medical professional, but rather act as a tool in helping to confirm diagnoses, or as an indicator for users to seek professional medical attention. Conversely, although this model has an intended dual purpose for both professional medical and personal use, the actual performance is expected to be more favorable in medical settings. This is due to the nature of the HAM10000 dataset, as it is comprised of professional quality dermatoscopic images that a user typically would not have access to outside of a hospital; Any image taken by the user with potentially either their phone or by an actual camera would almost certainly not contain a level of detail high enough to be classified accurately by the model, given that many types of skin lesions may appear extremely similar but can be differentiated through detection of highly specific features.

Despite these limitations, the proposed model and application still serve as a promising implementation of a custom trained sequential CNN for use in medical environments, and holds the potential to serve as a personal diagnostic tool given further refinements.

V. Conclusions and Future Work

In conclusion, a custom sequential convolutional neural network was succesfully developed and trained on the HAM10000 skin lesion dataset using the Adam optimization algorithm, all through use of the TensorFlow/Keras Python library(s). The model was able to achieve a 79.52% level of accuracy over 25 epochs of training to minimize validation loss and simultaneously maximize accuracy. The trained model was

implemented into a custom application with a user interface that allows users to input their own images for classification by the model, displaying a predicted class, any potential accompanying conditions, and a calculated confidence score.

To further improve this model, the most prominent issue to be improved upon is overfitting; Although the trained model was not subject to overfitting, this is only due to the training process being stopped early to avoid it. To further reduce the prominence and effects of overfitting, we could minimize the loss function and simplify the model by means of regularization, or use cross validation—selecting tuned parameters to be fed back into the model before retraining—to achieve this [4]. An additional consideration for improving the model's performance would be employing the use of larger, more evenly distributed datasets that would provide sufficient diversity of data to train the model further. An example of such datasets is the SLICE-3D dataset which conists of 400000 dermoscopic images collected from dermalogical centers spanning the globe [5]. Datasets of this scale were not selected for this project due to insufficient hardware to support the efficient training of a model on such a dataset. Finally, the exploration of different, alternative model architectures could serve beneficial to improving the performance of our application, as the proposed model is relatively simplistic in design, and at the very least could certainly be further refined through parametric tuning and architectural reformation.

ACKNOWLEDGMENT

We, the authors, would like to acknowledge the support of Dr. Omar Falou in providing us with both the foundational knowledge in computer vision to succeed in our development efforts, but also for the opportunity to work on this project.

REFERENCES

[1] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, no. 1, Aug. 2018, doi: https://doi.org/10.1038/sdata.2018.161.

[2] Jason Brownlee, "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning," Machine Learning Mastery, Jul. 02, 2017. https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

[3] IBM, "What is Overfitting? — IBM," www.ibm.com, 2024. https://www.ibm.com/topics/overfitting

[4] "How to Avoid Overfitting," KDnuggets. https://www.kdnuggets.com/2022/08/avoid-overfitting.html

[5] N. R. Kurtansky et al., "The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection," Scientific Data, vol. 11, no. 1, Aug. 2024, doi: https://doi.org/10.1038/s41597-024-03743-w.