Convolutional Neural Network (CNN ) Based Approach for Multiclass Classification of skin lesions from HAM10000 dermatoscopic image

Aadarsh Kushwaha   
*Msc Data Analytics*  
CCT College DublinDublin, Ireland  
2023398@student.cct.ie

*Abstract*

The aim of this research is to create and assess a Convolutional Neural Network (CNN) model tailored for the multiclassification of skin lesions, utilizing the HAM10000 dataset from ISIC 2018 (Tschandl, 2018), which consists of dermatoscopic pictures. To help dermatologists diagnose and treat skin diseases more quickly, this study uses deep learning techniques to accurately classify skin lesions into several diagnostic groupings by using Keras and TensorFlow Deep Learning Platform. The training set comprises 10,015 dermoscopic images labeled across seven diagnostic categories: dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), vascular lesions (vasc), actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (mel), melanoma (mel), melanoma (mel), and melanocytic nevi (akiec). Using Big Data technology like Apache Hadoop for distributed computing for large dataset (foundation, 2006-2024) and Python Api of Apache Spark called PySpark in order to achieve Parallel Processing (Foundation, 2019), Convolutional Neural Networks (CNNs) on the Jupyter notebook environment, a multiclass classifier model was created. Despite the small amount of training data, image data augmentation techniques were employed. Multiclass classifiers using a 12-layer model architecture trained for 150 epochs and achieved 84% accuracy rate.

*CCS Concepts*

Data Preprocessing, Statistical study, Data splitting, CNN Model Architecture, Model Training, Data Augmentation technique, Model Evaluation, Fine tuning, and Optimization

Keywords

Convolutional Neural Networks (CNN), Skin Lesion Classification, Dermatoscopic Images, Multiclass Classification, HAM10000 Dataset

# Introduction

Can a CNNs model effectively classify skin lesions from dermatoscopic images into multiple diagnostic categories with high accuracy and reliability, using the HAM10000 dataset?

All the experiment is carried out Jupyter notebook using Big Data Hadoop, Big data refers to a collection of structured, unstructured, and semi-structured datasets, posing challenges for storage and processing within conventional database systems. Managing these vast datasets entails addressing hurdles in storage, analysis, transfer, searchability, and more. It has default batch size of 128MB. Since doesn’t perform well on small amount of data set. So for this research HAM10000 dataset is of 2.8GB which is not too good but workable on that. Hadoop is used for storage on distributed system (CloudDuggu, 2023). Apache Spark is a rapidly expanding open-source cluster computing tool with broad applicability. It offers an extensive range of APIs in Java, Scala, Python, and R, alongside an engine that facilitates general execution (cloudduggu, 2023). Here in research I have python Api of Spark as PySpark for processing huge dataset on Hadoop cluster in order to achieve parallel computing.

Of all the cancers that affect people worldwide, skin cancer is the most common form that affects people of all genders. Both melanoma and non-melanoma skin cancers, including keratinocyte skin cancers such as squamous cell carcinoma and basal cell carcinoma, have become more common in the last few decades. World Health Organization (WHO) data indicates that each year, between 2 and 3 million instances of non-melanoma skin cancer and about 132,000 cases of melanoma skin cancer are detected worldwide. Surprisingly, skin cancer accounts for one out of every three cancer cases diagnosed annually. The increasing prevalence of both melanoma and non-melanoma skin cancers is mostly attributed to behavioural modifications and increased exposure to ultraviolet (UV) light when engaging in outdoor activities (Kreuter, 2023). Approximately 90% of non-melanoma skin cancers are linked to UV exposure. Despite the favourable prognosis associated with early detection, over 5,400 individuals succumb to advanced non-melanoma skin cancer globally each month. (Gururaj, et al., 2023).

Skin lesions can be categorized into two types: primary and secondary. Primary skin lesions represent irregular skin conditions that may be present from birth or develop gradually over time. Alterations or aggravations to primary skin lesions can lead to the formation of secondary skin lesions. These secondary lesions typically originate from the scabbing process that occurs when a mole is scraped until it bleeds (Dhivyaa, 2020). Dermatologists advise one of three treatments for affected skin, based on the type of lesion: home care, medication, or surgery. Despite appearing harmless, some skin lesions can carry significant risks, potentially indicating malignancy and requiring surgical treatment. Melanoma, the most dangerous form of skin cancer, can become life-threatening once it spreads, although it can be effectively treated in its early stages. Therefore, precise diagnosis of skin abnormalities is vital to protect patients' health and ensure timely medical intervention. (Abbas, 2021).

In the proposed work, Convolutional Neural Networks (CNN) is used to accurately classify pigmented skin lesions in dermoscopic images to detect malignant skin lesions as early as feasible.

The proposed study employed the HAM10000 dataset, which comprises 10,015 images featuring a diverse array of dermoscopic images showcasing pigmented skin lesions commonly encountered in various contexts. Similar to many medical datasets, the HAM10000 dataset exhibits significant class imbalances, posing a notable challenge in the present research. The images in the dataset have a resolution of 600 × 450 pixels and are saved in JPEG format. Initially, the images undergo manual cropping and centring around the lesion, alongside adjustments for visual contrast and colour accuracy. Each image and patient record contain seven distinct attributes: patient age, gender, lesion ID (a unique identifier for specific lesion types), image ID (a unique number identifying each image), DX type for technical validation, geographic location of the skin lesion, and a diagnostic category for the skin lesion, aiding in the classification and diagnosis of various skin conditions (Tschandl, 2018). Type of DX are given in image below:

# Literature review

(Nazia Hameed, 2020) has proposed a Multi-Class Multilevel (MCML) classification method inspired by the "divide and conquer" strategy. This approach combines machine learning and deep learning techniques. They also introduced DSNet, an automatic semantic segmentation network for skin lesions. To optimize the network and reduce the parameter count, they utilized separable depth-wise convolution. (Johnpaul, 2020) has suggested a model capable of generating high-resolution feature maps to aid in retaining spatial information within the image. Utilizing two distinct datasets, the authors advocated for the application of Random Forest and Decision Tree algorithms for classifying skin types.

Data augmentation strategies were proposed by (Bhoi, 2021) to equalize different types of lesions across picture ranges. Their model, which relied on LSTM and MobileNet V2 techniques, demonstrated efficacy in the identification and categorization of skin diseases while utilizing little computational resources. Using CNN transfer learning, (Schaefer, 2020) demonstrated how image size affects the classification of skin lesions. They proved that, in terms of performance, image cropping is superior to resizing. Moreover, the best classification performance was obtained using a straightforward ensemble method that included the output of three CNNs that were fine-tuned and images that had been cropped at six different sizes. By fine-tuning weights and biases, (Zhang, 2020) documented the effectiveness of a CNN optimized with an enhanced whale optimization technique, with the goal of minimizing discrepancies between network output and desired output.

To sum up, a great deal of work has been done by a number of researchers using a variety of machine learning and deep learning techniques, frequently using a variety of datasets, to improve the categories for skin lesions. The main objective of this work is to categorize skin lesions using the HAM10000 dataset, which has only been used by a few number of academics. This effort aims to produce much better results by classifying skin lesions into seven groups using various machine learning and convolutional neural network techniques.

# Proposed methodology

Identifying skin cancer in its initial stages poses a significant challenge for dermatologists. The widespread adoption of deep learning techniques, as illustrated in Figure 2, aids in the classification of seven types of skin cancer images. Among various methodologies, Convolutional Neural Networks (CNN) have emerged as the foremost approach for object detection and classification tests.

## Dataset Description

In terms of diagnostic skin lesion categories, the data set included seven different classes.

1. Actinic Keratoses [akiec]: These are forms of squamous cell carcinoma that are non-invasive and can be treated locally without resorting to surgery. (The dataset comprises 327 images.)
2. Basal Cell Carcinoma [bcc]: This type of epithelial skin cancer seldom spreads but, if untreated, can pose a threat to life. (There are 514 images included in the dataset.)
3. Benign Keratosis-like Lesions [bkl]: Conditions such as seborrheic keratoses, lichen-planus like keratoses, and solar lentigo fall under "benign keratosis," resembling seborrheic keratosis or sun lentigo with regression and inflammation. (The dataset encompasses 1099 images.)
4. Dermatofibroma [df]: These skin lesions can be benign growths or a response to minor trauma. (There are 115 images provided in the dataset.)
5. Melanoma [mel]: Melanoma is a cancerous growth originating from melanocytes and can manifest in various forms. If detected early, it can be treated with a simple surgical procedure. (The dataset contains 1113 images.)
6. Melanocytic Nevi [nv]: These are harmless melanocyte neoplasms that can vary in appearance and size. Different variants may exhibit notable differences from a dermatoscopic standpoint. (The dataset consists of 6705 images.)
7. Vascular Lesions [vasc]: Examples of benign or malignant angiomas include cherry angiomas, angiokeratomas, and pyogenic granulomas. (There are 142 images available in the dataset.)

## Sampling the Dataset

Because of the dataset's disorganized picture structure, it is necessary to arrange each image according to its folder using the image 1D and file path. The most important factors for this sorting procedure are found to be the image ID and the diagnostic (dx) label. The dataset analysis indicates that the maximum count of 6705 is found in nv skin lesions, while the smallest sizes are represented by the counts for df and vasc skin lesions, which are 115 and 142, respectively. As such, using a dataset of 100\*7 photos for model training and selecting 100 images per class are not enough to provide the best classification accuracy. In order to overcome this constraint, more data will be produced, and data augmentation methods will be used to improve the diversity of the dataset and enhance model performance

## Data Visualization

The statistical insight got from the given data are as mentioned below:

1. In the provided dataset, most samples are from the age bracket of 45, totaling 1299 instances, while there are 400-464 samples collected from individuals aged 80-85. Additionally, skin lesions have been documented in children below the age of 10 years (group, 2023).
2. We observe that the male category is more heavily represented, with approximately 1000 more images (representing a 10% higher proportion) compared to the female category. While ideally, we would aim for equal representation, the relatively small imbalance suggests it can be addressed with lower priority. Additionally, there are 57 images with an unknown sex label, indicating instances where sex annotation was unavailable. Given the small percentage of data with unknown sex labels, it may be tempting to discard them.

## Data augmentation

Data augmentation involves creating additional "data" to enrich the training dataset. In this study, the proposed approach utilized the resample function from sklearn to generate a random number of data samples from the existing dataset. By employing data augmentation, the machine learning models are more likely to capture diverse characteristic features, enhancing their ability to distinguish between classes. Following augmentation, each class comprises 1000 images, resulting in a training dataset of 1000\*7 images, totalling 7000 images. To address the finite number of data points in the CNN model (with each class containing 1000 images), random transformations such as rotations and shearing were applied during training. Importantly, each epoch retained the same number of images as the original dataset, thereby mitigating the risk of overfitting through data augmentation

## Feature Engineering, Image resizing and Data spliting

All images undergo resizing to a 32x32 pixel NumPy array to capture the pixel values before being processed by various machine learning models. The utilization of the to\_categorical method from Keras facilitates the conversion of target labels from string format in the dataset to one-hot encoded vectors, thereby enabling easy interpretation of predictions from the integer class labels produced by the Keras CNN model. Following dataset augmentation, a total of 7000 images (comprising 1000 images from each class) are employed for training the machine learning model. Of these, 5250 images constitute 75% of the training set, while the remaining 1750 images constitute the testing set. Pixel values are subsequently normalized to a range of 0 to 1 by dividing the X label by 255, given that the maximum pixel value is 255.

# Training and Testing

I devised a multiclass classification model to categorize skin lesions and various types of skin cancer. Utilizing Convolutional Neural Networks (CNNs), adept at image analysis, enabled the discernment of weights and biases for different objects within images, facilitating their distinction. CNNs employ various layers, notably convolution and pooling layers, to extract features and reduce image size, respectively.

To optimize the training model, we experimented with different configurations, adjusting neuron counts and convolutional/pooling layer numbers. The resulting models demonstrated superior performance. Our CNN architecture includes sequential layers with four convolutional-pooling pairs, followed by flattening and dense layers. A dropout layer aids in preventing overfitting by randomly deactivating neurons during training.

Our CNN model customarily resizes images to 32x32 dimensions, commencing with a 256-filter 3x3 convolutional layer, succeeded by a 2x2 max-pooling layer. Subsequent layers employ 128 and 64 filters with ReLU activation. The model is compiled using categorical crossentropy loss, Adam optimizer (Musstafa, 2021), and softmax activation for multi-class classification, achieving optimal results. The Adam optimizer, renowned for its rapid convergence to the global optimum, stands as a prominent choice in deep learning optimization and was employed in training our models. Our model's final layer utilizes either the SoftMax activation function for multi-class classification or the Sigmoid activation function for binary classification.

The configuration is depicted in the accompanying figure, with the Adam optimizer favored for its rapid convergence. The final layer employs SoftMax activation for multi-class classification (Mandal, 2024).

We employed a split ratio of 75:25 for training and validation data in the multiclass classifier. Following each epoch, we logged and graphed the models' accuracy and loss. Accuracy denotes the proportion of correct predictions, while loss serves as a performance metric. We utilized the Categorical Cross-Entropy loss function to assess model performance. Each epoch utilized a batch size of 128, with training conducted over 150 epochs. Overfitting was not observed during training and validation. Initially, at 100 epochs, the accuracy stood at 78.39%, increasing to 84.11% as the epochs progressed to 150.

Relu Activation function:

y=max(0,x)

Adam Optimizer formula:

A black and white math equations

Description automatically generated with medium confidence

# Result

The multiclass classifier achieves an accuracy of 84.11% after 150 epochs, as shown in Figure 5. Analyzing the accuracy curves, we observe the training loss, which is the model's loss on the training data, and the validation loss, which is the loss on the testing data. Initially, both losses are high as the model hasn't learned accurate predictions. As training progresses, losses decrease. However, divergence between training and validation losses indicates overfitting, where the model becomes too specific to the training data. To mitigate overfitting, training epochs can be reduced, or regularization techniques applied. The optimal training epochs and regularization coefficient vary based on the dataset and model.

The graph illustrates the training and validation accuracy of a deep learning model. Training accuracy pertains to the model's accuracy on training data, while validation accuracy refers to its accuracy on validation data, typically more challenging. Both accuracies improve with more epochs, indicating learning. However, validation accuracy declines post-100 epochs, contrasting the training accuracy's rise, signalling overfitting. Overfitting arises when a model becomes overly attuned to training data, resulting in overly specific predictions. This can lead to inferior performance on novel data.

The bar chart illustrates the proportion of misclassifications among six distinct classes. On the x-axis, the true label is indicated, while the y-axis displays the fraction of erroneous predictions. Notably, all instances of "df" and "vasc" are correctly classified. However, "mel" and "nv" exhibit misclassifications, accounting for 26% and 33% respectively, of their total sample data, despite an equal sample size of 1000 for each type of skin lesion.

The classification report for a CNN model presents various metrics for each class:

Precision: The ratio of true positives to all predicted positives.

Recall: The ratio of true positives to all actual positives.

F1-score: The harmonic mean of precision and recall.

Support: The number of samples in each class. Additionally, the report includes the overall accuracy of the model, as well as the macro and weighted averages of the metrics across all classes.

The formulas for these metrics are as follows:

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-score = 2 \* (Precision \* Recall) / (Precision + Recall)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Macro average = (Precision + Recall + F1-score) / 3

Weighted average = (Precision \* Support + Recall \* Support + F1-score \* Support) / Total support (developers, 2007-2024)

This model exhibits high precision and recall for all classes, indicating accurate classification of most samples. Moreover, the overall accuracy of the model is high, demonstrating its strong performance on the dataset.

# Discussion

The CNN Deep Learning models were trained and evaluated on a Ubuntu operating system running on an Intel i9 processor with 8GB of RAM within a virtual box under the umbrella of a MAC OS environment, utilizing an i9 Intel processor with 16GB of RAM. The models were created with Jupyter Notebook (Sphinx, 2015) and Python 3.7.9, with Keras (group., n.d.), TensorFlow, and Numpy, Hadoop, PySpark (Foundation, 2024), Scikit-learn as dependencies.

The key findings of research on skin lesion classification utilizing dermatoscopic images, Convolutional Neural Networks (CNNs), and the HAM10000 dataset have significant implications for medical diagnostics and the development of automated systems for skin cancer detection. However, there are several limitations and areas for further investigation.

1. Implications: The primary discoveries from studies focusing on skin lesion classification through dermatoscopic images, Convolutional Neural Networks (CNNs), and the HAM10000 dataset carry substantial implications for medical diagnostics and the progression of automated systems for skin cancer detection. Nonetheless, there exist several constraints and avenues for additional exploration. Employing CNNs in skin lesion classification holds promise for enhancing diagnostic precision, assisting dermatologists in timely identification and treatment. Automated classification systems have the potential to streamline healthcare services by shortening diagnostic durations, thus enhancing access to diagnostic facilities, particularly in underserved regions lacking dermatologist access.
2. Limitations: Although extensive, the HAM10000 dataset may contain demographic, skin type, and lesion type biases, potentially limiting result generalizability. Variable lesion annotations may introduce inconsistencies in model training. Models trained on this dataset may not effectively generalize to diverse populations or underrepresented lesion types. Despite promising performance, CNNs require further clinical validation to ascertain their real-world efficacy and influence on patient outcomes.
3. Contradicting Viewpoints and Research Gaps: Research findings on CNN-based skin lesion classifiers may differ, highlighting the necessity for standardized evaluation protocols. Some studies may prioritize enhancing CNN model interpretability, potentially impacting their performance. Further investigation is needed to determine the transferability of pre-trained CNN models to new datasets or clinical contexts. Integrating additional modalities like patient history or genetic data with dermatoscopic images may improve accuracy but poses challenges in data integration and model refinement.

# Conclusion and Future Work

In this study, we introduced a deep neural network utilizing CNN architecture tailored for skin cancer and lesion detection. Our model achieved a promising accuracy of 84.11%. We observed that the model's accuracy is significantly influenced by the class distribution in the training set. Customized CNN techniques were assessed based on metrics such as Accuracy, Precision, Recall, and F1-Score. Pre-processing of images, feature-target separation, and data augmentation were conducted prior to training/testing. While our customized CNN exhibited improved accuracy, overfitting was observed, which can be mitigated by reducing the number of training epochs. These findings suggest superior classification performance for the HAM10000 dataset. Comparative analysis with existing studies demonstrated a 51% reduction in loss and fewer errors, reaffirming the efficacy of our proposed approach.

Future research endeavours may involve exploring alternative learning models, such as transfer learning, which utilizes pre-existing well-performing models to enhance accuracy rather than starting from scratch. Additionally, there is a need to focus on refining pre-processing techniques and reduce overfitting to effectively address imbalanced data in skin lesion/cancer analysis. Because overfitting leads to incorrect classification in new data. Studies like this are pivotal in advancing early detection and treatment of skin cancers due to their simplicity and precision. Moreover, such research can facilitate the development of practical tools, such as applications for physicians and patients, streamlining the diagnostic process and expediting treatment for this potentially debilitating disease.

# Reference

Abbas, S., 2021. *A novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm.* [Online]   
Available at: https://doi.org/10.7717/peerj-cs.390  
[Accessed 25 03 2024].

Bhoi, A. K., 2021. *Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM.* [Online]   
Available at: https://www.mdpi.com/1424-8220/21/8/2852  
[Accessed 26 03 2023].

cloudduggu, 2023. *Apache Spark Introduction.* [Online]   
Available at: https://www.cloudduggu.com/spark/introduction/  
[Accessed 23 2 2024].

CloudDuggu, 2023. *Big Data Overview.* [Online]   
Available at: https://www.cloudduggu.com/hadoop/big-data/  
[Accessed 23 2 2024].

developers, s.-l., 2007-2024. *sklearn.metrics.classification\_report.* [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification\_report.html  
[Accessed 31 03 2024].

Dhivyaa, C. R., 2020. *Skin lesion classification using decision trees and random forest algorithms.* [Online]   
Available at: https://link.springer.com/article/10.1007/s12652-020-02675-8  
[Accessed 20 03 2024].

Foundation, A. S., 2019. *Unified engine for large-scale data analytics.* [Online]   
Available at: https://spark.apache.org/  
[Accessed 25 02 2023].

Foundation, T. A., 2024. *PySpark Overview.* [Online]   
Available at: https://spark.apache.org/docs/latest/api/python/index.html  
[Accessed 4 2 2024].

foundation, T. A. S., 2006-2024. *Apache Hadoop.* [Online]   
Available at: https://hadoop.apache.org/  
[Accessed 20 03 2023].

group., K. G., n.d. *Getting started with Keras.* [Online]   
Available at: https://keras.io/getting\_started/  
[Accessed 2 2 2024].

group, N. s. c. b. a., 2023. *Jupyter Notebook.* [Online]   
Available at: localhost:8888/notebooks/Desktop/CA1\_SEM2\_Image\_Classification\_CNN/CA1\_SEM2\_CNN\_Image\_Classification.ipynb  
[Accessed 26 03 2023].

Gururaj, H. L., Manju, N. & Nagarjun, A., 2023. *DeepSkin: A Deep Learning Approach for Skin Cancer Classification.* [Online]   
Available at: https://ieeexplore.ieee.org/abstract/document/10122533  
[Accessed 31 03 2024].

Johnpaul, P., 2020. *Skin lesion classification using decision trees and random forest algorithms.* [Online]   
Available at: https://link.springer.com/article/10.1007/s12652-020-02675-8  
[Accessed 25 03 2023].

Kreuter, P. D. A., 2023. *Skin Cancer: Recent Advances in Diagnosis, Treatment, and Prevention.* [Online]   
Available at: https://www.mdpi.com/journal/cancers/special\_issues/0W34NG34VF  
[Accessed 03 31 2024].

Mandal, M., 2024. *Introduction to Convolutional Neural Networks (CNN).* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/  
[Accessed 23 03 2024].

Musstafa, M., 2021. *Optimizers in Deep Learning.* [Online]   
Available at: https://musstafa0804.medium.com/optimizers-in-deep-learning-7bf81fed78a0  
[Accessed 16 03 2024].

Nazia Hameed, A. M. S. M. K. G. M. H., 2020. *Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques.* [Online]   
Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417419306797?via%3Dihub  
[Accessed 25 03 2024].

Schaefer, G., 2020. *Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification.* [Online]   
Available at: https://www.sciencedirect.com/science/article/abs/pii/S0169260719311460?via%3Dihub  
[Accessed 25 03 2023].

Sphinx, 2015. *Project Jupyter Documentation.* [Online]   
Available at: https://jupyter.org/  
[Accessed 2 1 2024].

Tschandl, P. R. C. &. K. H., 2018. *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.* [Online]   
Available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T  
[Accessed 15 02 2024].

Zhang, N., 2020. *Skin cancer diagnosis based on optimized convolutional neural network.* [Online]   
Available at: https://www.sciencedirect.com/science/article/abs/pii/S0933365719301460?via%3Dihub  
[Accessed 26 03 2023].