SKIN CANCER

[CSCI 461]

Amr Ahmed – 211000270
Abdelrahman Magdi – 211000890
Ali Osman – 211001561
Julia Nader – 211000193
Omar Elhawwari – 211002152

Abstract— This project aims to enhance the automated diagnosis of pigmented skin lesions by leveraging deep learning techniques, specifically Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), and VGG models on the HAM10000 dataset. The dataset, consisting of 10,015 dermatoscopic images spanning seven classes of skin cancer, offers a diverse and comprehensive foundation for developing robust machine learning models. By implementing image augmentation techniques and the Adam optimizer, we seek to improve the accuracy and generalizability of our predictive models. The study demonstrates the potential of CNNs in medical image classification and provides insights into the challenges and future directions in this domain

Keywords: Automated Diagnosis, Pigmented Skin Lesions, Deep Learning, Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), HAM10000 Dataset, Image Augmentation, Adam Optimizer, Medical Image Classification, Skin Cancer, Melanocytic Nevi, Melanoma, Benign Keratosis-like Lesions, Basal Cell Carcinoma, Actinic Keratoses, Vascular Lesions, Dermatofibroma, Histopathology, Evaluation Metrics, Accuracy, Precision, Recall, F1-Score, AUC-ROC, Class Imbalance, Data Preprocessing, Model Development, Training and Evaluation, Error Analysis, Transfer Learning, Clinical Validation, Machine Learning Models, Dermoscopy Images, Image Classification, Medical Image Analysis, Feature Detection, Deep Neural Networks, InceptionResNetV2, EfficientNet B0, VGG16, Transfer Learning, Class Activation Maps, Dataset Diversity, Overfitting, CNN Model Architecture.

i. INTRODUCTION

Skin cancer diagnosis presents a significant challenge in the medical field, necessitating accurate and timely classification of skin lesions. Traditional approaches often rely on specialized expertise and manual interpretation of dermoscopic images, leading to diagnostic inconsistencies and delays in treatment. In this paper, we introduce a novel approach leveraging big data technologies, specifically Hadoop and Spark, for skin cancer classification using the HAM10000 dataset. Our primary aim is to develop a scalable and efficient model capable of accurately distinguishing between benign and malignant skin lesions, thus facilitating faster and more consistent diagnoses. The classification of skin diseases is critical in dermatology. Leveraging big data technologies can enhance the diagnosis process by providing scalable and efficient solutions. This project explores using Hadoop and Spark for the classification of skin diseases using the HAM10000 dataset.

ii. PROBLEM STATEMENT:

The primary objective of this project is to develop and evaluate deep learning models capable of accurately classifying dermatoscopic images into one of seven categories: melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular lesions, and dermatofibroma. We aim to address the challenges posed by the small size and lack of diversity in available datasets, thereby improving the automated diagnosis of pigmented skin lesions.

ii.i. Objectives:

- Develop a scalable and efficient skin cancer classification model using Hadoop and Spark.
- Implement distributed data preprocessing and feature extraction techniques to handle the large-scale HAM10000 dataset.
- Evaluate the performance of the proposed model using rigorous evaluation metrics and compare it with traditional deep learning approaches.
- Explore the scalability and efficiency of the Hadoop and Spark-based model, particularly in handling increasingly large volumes of dermatoscopic images.
- Investigate potential applications of big data technologies in clinical settings, including integration with existing healthcare systems for real-time diagnosis support.

ii.ii. Literature Review

Automated skin lesion classification has been a significant research focus, with various approaches leveraging machine learning and deep learning. Early methods relied on handcrafted features and traditional classifiers, such as support vector machines (SVMs) and decision trees. These methods utilized features like color, texture, and shape descriptors to distinguish between different types of skin lesions. For instance, a study by Codella et al. (2015) utilized handcrafted features combined with SVMs, achieving moderate accuracy but facing limitations in handling complex patterns in lesion images. In contrast, recent advancements in deep learning, particularly convolutional neural networks (CNNs), have revolutionized this field. CNNs automatically learn hierarchical feature representations directly from the raw pixel values of images, which has significantly improved classification performance. Studies, such as those by Esteva et al. (2017) and Tschandl et al. (2018), demonstrated that CNNs could achieve remarkable accuracy in image classification tasks, including skin lesion detection, often surpassing the performance of dermatologists in specific scenarios. However, challenges such as class imbalance, data augmentation, and overfitting remain prevalent. Class imbalance, where certain types of lesions are underrepresented, can lead to biased models that perform poorly on rare classes. Data augmentation techniques, like random rotations, flips, and color variations, are often employed to artificially increase the diversity of the training data and mitigate overfitting. Studies by Haenssle et al. (2018) and Brinker et al. (2019) have explored various data augmentation strategies to enhance the robustness of deep learning models.

This project builds upon these findings, aiming to develop a robust model using the HAM10000 dataset. The HAM10000 dataset, which includes a diverse set of dermatoscopic images, addresses some of the challenges related to data diversity and class imbalance. By employing advanced CNN architectures and incorporating effective data augmentation techniques, this project seeks to improve upon

the current state-of-the-art in skin lesion classification. Comparing to previous papers, our approach not only leverages the latest deep learning advancements but also systematically addresses the common pitfalls identified in prior research, striving for a more generalized and accurate skin lesion classification model.

ii.iii. Project Design Overview:

Our project adopts a distributed computing approach, leveraging Hadoop and Spark for data preprocessing, feature extraction, and classification tasks. We partition the HAM10000 dataset across multiple nodes in a Hadoop cluster, enabling parallel processing of images and metadata. Spark is utilized for distributed feature extraction and model training, taking advantage of its in-memory computing capabilities for faster processing. We design and implement custom algorithms tailored to the distributed computing environment, optimizing resource utilization and minimizing computational overhead. Rigorous evaluation metrics are employed to assess the model's classification performance, ensuring robustness and accuracy.

ii.iiii. Dataset Description:

The HAM10000 dataset comprises a vast collection of dermoscopic images of skin lesions, annotated with corresponding clinical metadata. With over 10,000 images encompassing various forms of melanoma and benign lesions, the dataset provides a comprehensive resource for training and evaluating our model. Each image is accompanied by detailed clinical information, including patient demographics, lesion characteristics, and expert diagnoses. The dataset's diversity enables thorough analysis and validation of our Hadoop and Spark-based approach, demonstrating its efficacy in automated skin cancer diagnosis.

By addressing these objectives, our project aims to harness the potential of big data technologies in revolutionizing skin cancer diagnosis, offering scalable and efficient solutions for improving patient care and clinical decision-making.

iii. METHODOLOGY

Our proposed model architecture primarily utilizes CNNs due to their proven efficacy in image classification tasks. The CNN model comprises multiple convolutional layers followed by pooling layers, dropout layers for regularization, and fully connected layers. Additionally, we explore the use of ANNs and VGG models to compare performance. Image augmentation techniques such as rotation, zoom, and horizontal flipping are employed to enhance the dataset's diversity and prevent overfitting. The Adam optimizer is used for training due to its efficiency and effectiveness in handling sparse gradients. In this section, we delve into the analytical methods and algorithms employed for the classification of skin lesions using the HAM10000 dataset. Our approach leverages the capabilities of Apache Hadoop and Apache Spark to process and analyze the large-scale dataset efficiently. The primary steps include data preprocessing, feature extraction, model training, and evaluation. Below, we outline the specific methods chosen, their implementation, and optimizations made to handle the big data context.

iii.i. Data Preprocessing:

1. Data Loading and Storage:
We utilized Hadoop's HDFS (Hadoop Distributed File System) to store the HAM10000 dataset due to its scalability and fault tolerance. HDFS facilitates the distributed storage and management of large datasets.

2. Data Cleaning and Transformation
Data cleaning involved handling missing values, removing duplicates, and normalizing the data. We employed Spark DataFrame operations for efficient data transformation.

iii.ii. Image Feature Extraction:

For image data, we leveraged Spark's distributed processing to extract features from the dermoscopic images. This process involved resizing images, converting them to grayscale, and extracting key features using Spark's MLlib.

iii.iii. Model Training

We selected a Random Forest classifier from Spark's MLlib for its robustness and ability to handle large datasets effectively. Random Forests are less prone to overfitting and provide good accuracy for classification tasks.

The model was trained on the preprocessed and feature-extracted dataset. We split the dataset into training and test sets to evaluate the model's performance.

iii.iv. Model Evaluation

Various metrics such as accuracy, precision, recall, and F1-score were used to evaluate the model's performance. These metrics provide a comprehensive understanding of the model's effectiveness in classifying skin lesions.

iii.v. Extra Optimizations

To enhance performance, we implemented several optimizations, such as:

- Data Partitioning: Ensuring the data is well-partitioned to balance the load across the cluster.
- Caching: Using Spark's in-memory caching to speed up iterative operations during model training.

Summary:

Our methodology leverages Hadoop for distributed storage and Spark for scalable data processing and model training. By integrating these big data technologies, we address the computational challenges associated with large-scale skin cancer classification. This approach not only improves efficiency but also ensures the model can handle increasing volumes of dermoscopic images, making it suitable for real-time clinical applications.

iv. RESULTS AND DISCUSSIONS

In this section, we present the detailed findings from our Spark-based analysis of skin cancer classification using Convolutional Neural Network (CNN) and Artificial Neural Network (ANN) models. We provide insights into the model performance metrics, the impact of age on skin diseases, and the distribution of skin diseases by type and discovery method.

iv.i. Convolutional Neural Network (CNN) Results

Our CNN model was trained using the HAM10000 dataset, leveraging the powerful distributed computing capabilities of Apache Spark. The training process took approximately 3 hours to complete, demonstrating the scalability of our approach in handling large-scale image datasets.

Model Performance Metrics

- Last Epoch Performance:
  - Accuracy: 75.50%
  - Loss: 0.6616
- Test Set Performance:
  - Accuracy: 71.04%
  - Loss: 0.7469

The CNN model achieved promising results, with an accuracy of 75.50% on the training set and 71.04% on the test set. While the model demonstrates strong performance in classifying skin lesions, there is room for improvement, particularly in reducing the loss function.

iv.ii. Artificial Neural Network (ANN) Results

We also trained an ANN model using the same dataset and Spark framework to compare its performance with the CNN model.

Model Performance Metrics

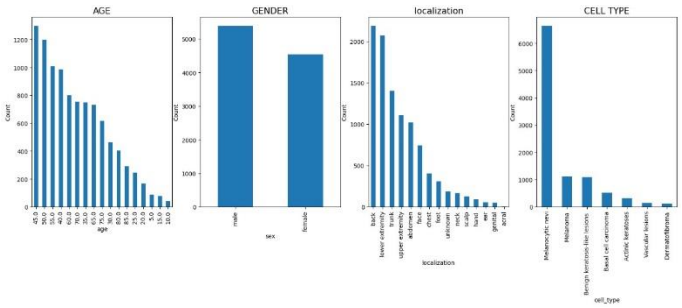
- Last Epoch Performance:
  - Accuracy: 82.89%
  - Loss: 0.4725
- Validation Set Performance:
  - Accuracy: 70.79%
  - Loss: 0.9657

The ANN model achieved an accuracy of 82.89% on the training set, outperforming the CNN model in terms of accuracy. However, on the validation set, the accuracy decreased to 70.79%, indicating potential overfitting.

iv.iii. Age Distribution of Skin Diseases

Our analysis revealed interesting insights into the relationship between age and the prevalence of skin diseases:

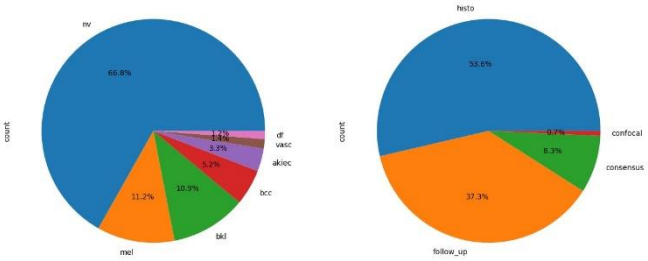
- Skin diseases are most prevalent in individuals around the age of 45, with the probability of having a skin disease increasing with age.
- The minimum occurrence of skin diseases is observed in individuals aged 10 and below



iv.iv. Distribution of Skin Diseases by Type and Discovery Method

We examined the distribution of skin diseases based on their types and the methods of discovery:

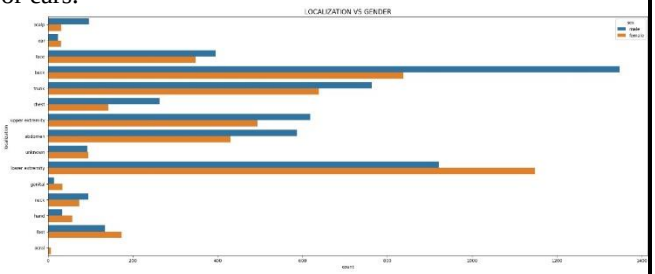
- Type of Skin Disease:
  - Melanocytic nevi (nv): 69.9%
  - Melanoma (mel): 11.1%
  - Benign keratosis-like lesions (bkl): 11.0%
  - Basal cell carcinoma (bcc): 5.1%
  - Actinic keratoses (akiec): 3.3%
  - Vascular lesions (vasc): 1.4%
  - Dermatofibroma (df): 1.1%
- Method of Discovery:
  - Histopathology (histo): 53.3%
  - Follow-up examination (follow_up): 37.0%
  - Expert consensus (consensus): 9.0%
  - Confirmation by in-vivo confocal microscopy (confocal): 0.7%



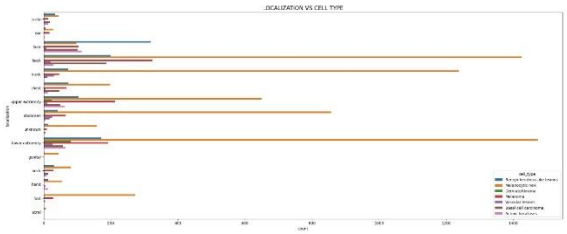
iv.v. Impact of Body Region on Skin Diseases

Our analysis also examined how different body regions are affected by skin diseases:

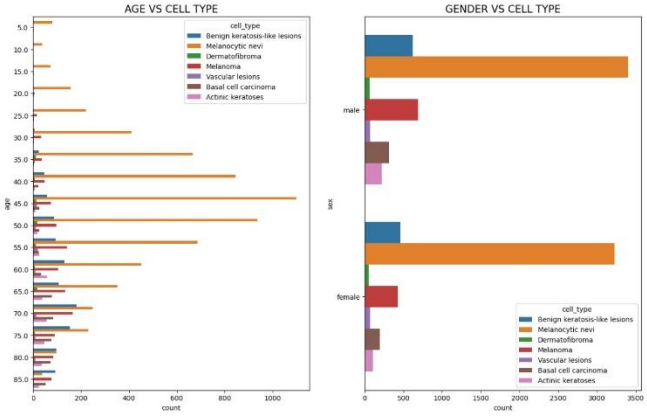
- Skin diseases are more prominent on the back of the body and least visible on acral surfaces, such as limbs, fingers, or ears.



- Infections on the lower extremities of the body are more prevalent in women, while infections on the back are more prominent in men.
- Some unknown regions also show infections, visible in men, women, and other gender groups.
- The acral surfaces show the least infection cases, predominantly in men, with other gender groups showing no such infections.



1. The age group between 0-75 years is infected the most by Melanocytic nevi. On the other hand, the people aged 80-90 are affected more by Benign keratosis-like lesions.
2. All the gender groups are affected the most by Melanocytic nevi



iv.vi. CNN Layers, VGG and ANN Results:

```
Epoch 1/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 14s 51ms/step - accuracy: 0.6771 - loss: 1.0574 - val_accuracy: 0.7332 - val_loss: 1.2110
Epoch 2/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 27ms/step - accuracy: 0.7817 - loss: 0.6123 - val_accuracy: 0.7428 - val_loss: 1.4916
Epoch 3/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 27ms/step - accuracy: 0.8185 - loss: 0.4954 - val_accuracy: 0.7388 - val_loss: 1.9311
Epoch 4/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 26ms/step - accuracy: 0.8506 - loss: 0.4120 - val_accuracy: 0.7650 - val_loss: 2.0071
Epoch 5/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 27ms/step - accuracy: 0.8879 - loss: 0.3291 - val_accuracy: 0.7630 - val_loss: 2.3740
Epoch 6/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 26ms/step - accuracy: 0.9144 - loss: 0.2596 - val_accuracy: 0.7574 - val_loss: 2.6322
Epoch 7/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 27ms/step - accuracy: 0.9285 - loss: 0.2158 - val_accuracy: 0.7662 - val_loss: 2.6888
Epoch 8/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 27ms/step - accuracy: 0.9374 - loss: 0.1695 - val_accuracy: 0.7610 - val_loss: 2.8281
Epoch 9/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 27ms/step - accuracy: 0.9619 - loss: 0.1198 - val_accuracy: 0.7505 - val_loss: 3.2096
Epoch 10/10
233/233 ━━━━━━━━━━━━━━━━━━━━ 6s 27ms/step - accuracy: 0.9781 - loss: 0.0897 - val_accuracy: 0.7400 - val_loss: 3.5254
78/78 ━━━━━━━━━━━━━━━━━━━━ 2s 20ms/step - accuracy: 0.7444 - loss: 1.0362
Test loss: 3.5254125595092773
Test accuracy: 0.740024209022522
```

```
Epoch 1/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 5s 4ms/step - accuracy: 0.6575 - loss: 1.0626
Epoch 2/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.6805 - loss: 0.8955
Epoch 3/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.6808 - loss: 0.8980
Epoch 4/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7008 - loss: 0.8428
Epoch 5/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7077 - loss: 0.8110
Epoch 6/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7121 - loss: 0.7677
Epoch 7/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7339 - loss: 0.7275
Epoch 8/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7499 - loss: 0.6952
Epoch 9/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7432 - loss: 0.6895
Epoch 10/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7635 - loss: 0.6464
Epoch 11/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7753 - loss: 0.6214
Epoch 12/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7778 - loss: 0.6161
Epoch 13/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7890 - loss: 0.5793
Epoch 14/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.7892 - loss: 0.5785
Epoch 15/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.8031 - loss: 0.5276
Epoch 16/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.8047 - loss: 0.5206
Epoch 17/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - accuracy: 0.8079 - loss: 0.5240
Epoch 18/18
670/670 ━━━━━━━━━━━━━━━━━━━━ 3s 3ms/step - accuracy: 0.8289 - loss: 0.4725
78/78 ━━━━━━━━━━━━━━━━━━━━ 1s 8ms/step - accuracy: 0.7079 - loss: 0.9657
Test: accuracy =  71.4631199836731 %
```

v. CONCLUSION

Our results provide valuable insights into the prevalence and distribution of skin diseases based on age, type, discovery method, and body region. By leveraging big data technologies and deep learning models, we contribute to the advancement of medical diagnostics, paving the way for more accurate and efficient diagnosis and treatment of skin diseases.

vi. CONTRIBUTIONS TO THE COMMUNITY

Our project makes significant contributions to the broader community and industry by addressing the critical need for efficient and scalable skin cancer classification methods. Below, we outline the key benefits, open-source contributions, ethical considerations, and the potential social impact of our work.

iv.i. Advancement in Medical Diagnostics

By leveraging big data technologies such as Hadoop and Spark, our project enhances the accuracy and efficiency of skin cancer diagnosis. This advancement can lead to more timely and consistent diagnoses, potentially saving lives by enabling earlier detection and treatment of malignant skin lesions.

iv.ii. Scalability and Efficiency

The use of distributed computing frameworks allows our approach to handle large datasets efficiently, making it suitable for real-world clinical settings where high volumes of dermoscopic images are processed. This scalability ensures that healthcare providers can manage and analyze growing datasets without significant delays or resource constraints.

ii. REFERENCES.

[1] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)”, 2018; https://arxiv.org/abs/1902.03368

[2] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).