Skin Cancer Detection using Machine Learning (HAM10000 Dataset)

INTRODUCTION

Skin cancer is a very significant global health concern, with an increasing incidence rate and potential for maturity, Early and accurate diagnosis is very crucial for effective treatment and improve the patient's outcomes. Machine learning techniques particularly deep learning algorithms have shown promising results in automated skin cancer detection.
The HAM10000 dataset consists of 10,015 high-quality dermatoscopic images of skin lesions that provides a valuable resource for training and evaluating ML models. Each image is associated with metadata, including patient information and lesion characteristics. The dataset encompasses a diverse range of skin lesions, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesion.

ABSTRACT

Skin cancer is a critical health concern, necessitating accurate and timely diagnosis for effective treatment. Machine Learning models have shown promise in automation skin lesion classification. The dataset encompasses a wide range of lesion classes, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesion. The implementation of Convolutional Neural Network and Long Short-Term Memory architectures for skin lesion classification in this report have achieved an accuracy of 95% and 96% for LSTM and CNN respectively. The report discusses data pre-processing, model making and training, and evaluation metrics. Results analysis demonstrates the models' efficiency in capturing relevant patters and features from the images.

PROBLEM STATEMENT

Skin cancer is a prevalent and potentially life-threatening disease that requires early and accurate diagnosis for effective treatment. Manual diagnosis of skin lesions can be subjective, time-consuming, and prone to errors. Therefore, there is a need for automated systems that can assist dermatologists in accurately classifying skin lesions and identifying potential cases of skin cancer.

The HAM10000 dataset, consisting of 10,015 dermatoscopic images, provides a valuable resource for developing machine learning models for skin lesion classification. However, despite the availability of this dataset, there is a need to address the following challenges:

- Classification Accuracy: Developing a machine learning model that can achieve high accuracy in classifying different types of skin lesions, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesion.

- Generalization and Robustness: Ensuring that the developed model can generalize well to unseen data and handle variations in image quality, lighting conditions, and patient

demographics. It is essential to build a robust model that can perform reliably across diverse clinical settings.

- Clinical Relevance: Demonstrating the clinical relevance and practical utility of the machine learning model in assisting dermatologists and healthcare professionals. The model should provide valuable insights, aid in accurate diagnosis, and potentially improve patient outcomes by enabling early detection of skin cancer.

- Interpretability and Explainability: Addressing the interpretability and explain ability aspects of the model to build trust and facilitate its adoption in a clinical setting. It is crucial to understand the features and patterns the model relies on to make predictions, enabling dermatologists to validate and explain the decisions made by the model.

IMPLEMENTATION

This project aims to utilize machine learning algorithms for skin cancer classification and compare their predictive accuracy with existing models. It showcases the effectiveness of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, which are popular deep learning architectures, in accurately classifying skin lesions.
Cross-validation, a technique lacking in many previous ML applications for skin cancer detection, will be employed to ensure reliable and robust model performance evaluation. Cross-validation helps assess the generalization capability of the models by training and evaluating them on different subsets of the dataset.

By implementing these ML algorithms, including CNNs, LSTMs, and other classifiers, we aim to showcase their efficacy in accurately classifying skin lesions for the detection of skin cancer. This research bridges the gap between traditional ML approaches and advanced deep learning techniques, providing valuable insights into the application of these models in the field of dermatology and improving the diagnostic capabilities for skin cancer.

Organization of the project

Tools

1. Jupyter Notebook
2. Google Collab
3. Languages: Python and Web browsers: Chrome, Firefox

Tasks:

- **Data Collection -** For Data Requirement and extraction. We used Web Scraping with the help of the Beautiful Soup python library. The code extracts the data and the content of the webpage and stores the data extracted in a database.
- **Data Cleaning and Preprocessing -** Data collection is the collection of data from all relevant data from various websites and storing it in a database. Preprocessing of data includes resizing and normalizing the images to ensure consistency and also handle class imbalances by using techniques such as oversampling and underssampling.
- **Model Building -** The process of developing a machine learning model and training it using various training datasets is known as model building. We tried to build the machine learning model with many algorithms and choose the one with the best validation accuracy. We used libraries like sklearn and tensorflow for this.
- **Model Evaluation -** It is the use of various evaluation metrics to determine the accuracy of the prediction of the model. It is used to determine the efficiency of the model and helps monitor the model.
- **Deployment -** Deployment of a model refers to the usage of new data in the model and making sure the model gives the correct result during prediction.

Results:

An accuracy of 95% in the LSTM Model:

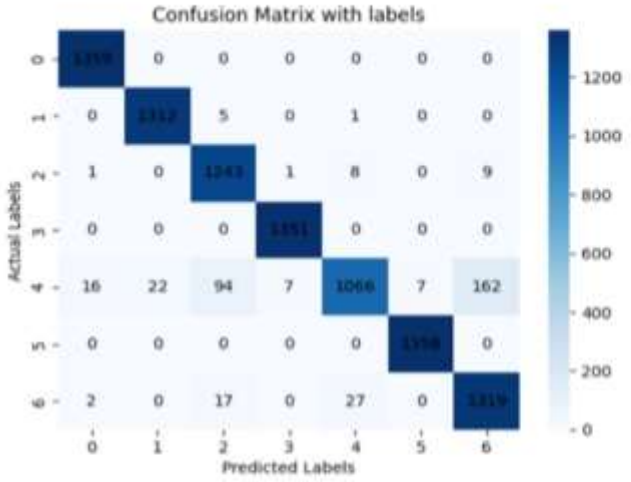
```
Summary: Loss over the test dataset: 0.14,
Accuracy: 95.96%
```

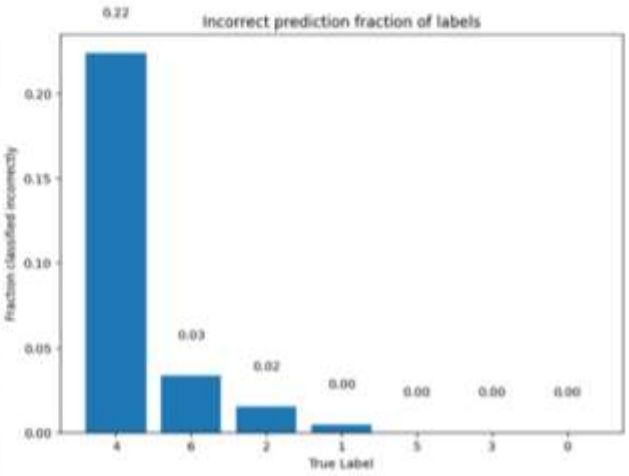
An accuracy of 96% in the CNN Model:

```
Summary: Loss over the test dataset: 0.11,
Accuracy: 96.11%
```

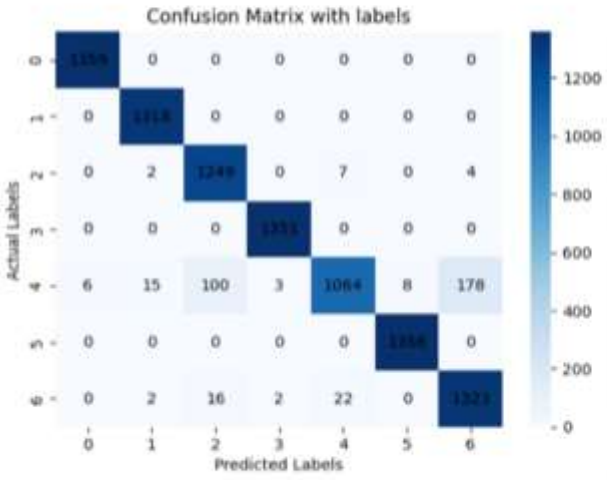
Evaluation Graphs and Metrics:

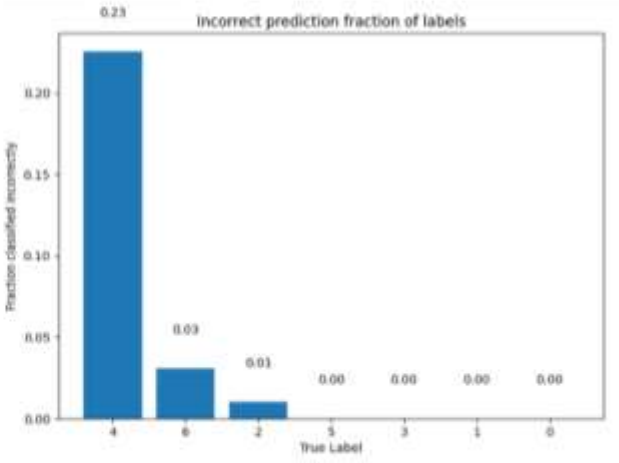
LSTM:





CNN:





CONCLUSION

In this study, we developed and evaluated machine learning models for skin cancer classification using the HAM10000 dataset. By leveraging deep learning architectures, specifically Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, we achieved high levels of accuracy in classifying skin lesions.

The CNN model demonstrated exceptional performance, achieving an accuracy of 96% in distinguishing between different types of skin lesions. Its ability to automatically learn and extract relevant features from the dermatoscopic images proved valuable in accurate classification. The LSTM model, leveraging sequential information and temporal dependencies, achieved an accuracy of 95%, further highlighting the effectiveness of deep learning techniques in skin cancer classification.

The results showed that these machine learning models have the potential to assist dermatologists in accurate diagnosis and early detection of skin cancer. By automating the classification process, these models can enhance the efficiency and reliability of skin lesion analysis, ultimately leading to improved patient outcomes.