**Skin Cancer Classification**

Kislay Kislay Mohammed Salman Taqi

Department of Mathematics Department of Mathematics

Stevens Institute of Technology Stevens Institute of Technology

Hoboken, United States Hoboken, United States

[kkislay@stevens.edu](mailto:kkislay@stevens.edu) [m4@stevens.edu](mailto:m4@stevens.edu)

*Abstract*—The most common type of cancer in the country is skin cancer. In the United States, more than 4 million cases of skin cancer are found each year. In order to reduce effort, time, and human life, an accurate automated system for skin lesion recognition is absolutely necessary for early detection. The goal of this project is to develop a machine learning model that can classify skin lesions at an early stage in order to accurately diagnose the illness and aid in clinical decision-making, increasing the likelihood that the condition can be treated before it spreads.

I. Introduction

Skin cancer is one of the most frequent cancers not only in the United States, but also worldwide, with about 10,000 people diagnosed with it in the United States every day. De- spite the fact that the number of Melanoma deaths is expected to rise by 22% in the coming year, early identification of the disease can lead to a 99% survival rate. Early detection of skin cancer is critical and can prevent further spread in some cases, such as melanoma and focal cell carcinoma. In any case, there are several factors that have a negative impact on detection accuracy. In recent years, the use of image processing and computer vision in healthcare and medical applications has increased significantly. In this study, we use deep learning models to detect and categorize cancer based on dermoscopic images of pigmented lesions.

II. Related Work

Skin cancer classification has been a focal point in medical image analysis, leveraging advancements in machine learning and computer vision techniques. Various studies have explored different methodologies and datasets to enhance classification accuracy and diagnostic capabilities.

One prominent study by Esteva et al. (2017) utilized a deep convolutional neural network (CNN) architecture trained on a large dataset of dermoscopic images. Their work demonstrated significant progress in classifying skin lesions into multiple categories, achieving performance on par with dermatologists in identifying melanoma and other common skin diseases. Furthermore, Tschandl et al. (2018) presented an extensive evaluation of different CNN architectures and data augmentation techniques using the International Skin Imaging Collaboration (ISIC) dataset. Their findings emphasized the importance of data preprocessing and augmentation in improving classification accuracy across various skin lesion types. Another notable contribution by Haenssle et al. (2018) highlighted the effectiveness of a machine learning-based algorithm in supporting dermatologists during clinical practice. Their AI system showed high sensitivity in detecting melanoma, aiding medical professionals in making accurate diagnostic decisions. Moreover, recent research by Codella et al. (2019) introduced a dataset focusing on challenging aspects of skin lesion analysis, emphasizing the need for robust models capable of handling diverse clinical scenarios. Their dataset incorporated different skin types, ages, and environmental conditions, posing significant challenges for classification algorithms. In summary, recent advancements in skin cancer classification have primarily revolved around deep learning techniques, dataset curation, and the integration of machine learning systems into clinical workflows. These studies showcase the potential of AI-driven approaches in assisting dermatologists and improving diagnostic accuracy in skin cancer detection.

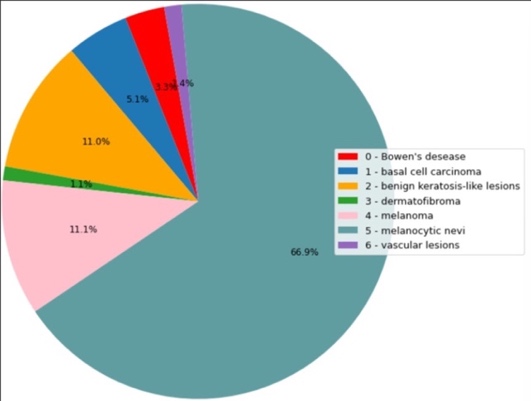
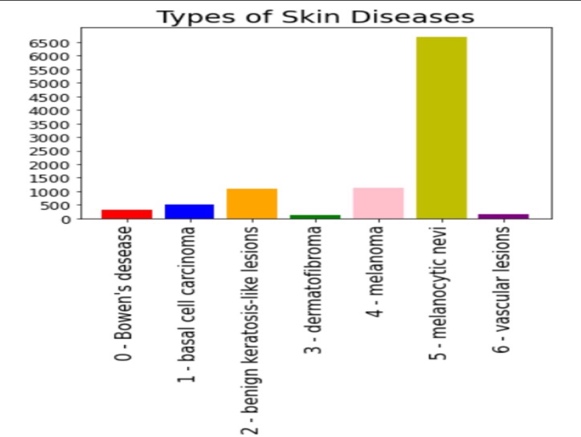
III. Description of Dataset

The Harvard database released accessible the HAM10000 (”Human Against Machine with 10000 training images”) dataset, which contains 10,015 dermatoscopic images, in June 2018. A metadata file including demographic information for each lesion is also provided. More than 50% of lesions are confirmed by histopathology (histo); the balance of the cases are confirmed by follow-up examination (follow up), expert consensus (consensus), or in-vivo confocal microscopy (con- focal). The HAM10000 dataset includes a file (HAM10000 metadata.csv) that contains additional dataset information, the most essential of which is the type of skin lesion displayed in each image. It is critical to comprehend the information contained in the metadata to determine which sections of the metadata may be used as a feature in our learning process. Here, we visualize the dataset’s metadata, specifically the

features age, gender, body localization, and cell type. Source of the Dataset: This dataset is a large collection of multi-sources dermatoscopic images of pigmented lesions from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for our project.

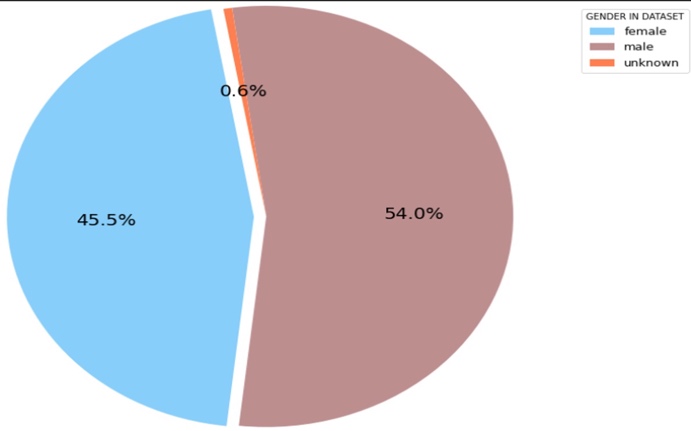
The URL for the data is as follows: https://www.kaggle.com/datasets/kmader/skin- cancer-mnist-ham10000

IV. Data Visualization

1) Types of Skin Diseases and their Counts

The following graphs show that the dataset is imbalanced, and the highest number of cases are that of ‘melanocytic nevi’ which are approximately 6500 in total. The lowest number of cases are that of ‘vascular lesions’ which are approximately close to 200 in total.

2) Distribution of skin cancer cases between male and female



The above pie-chart shows that 54% of the skin cancer patients are male and 45.5% of the skin cancer patients are female.

3) Distribution of age

The above plot shows the distribution of patients’ age spread over the entire dataset. The distribution is roughly a normal distribution with the maximum number of patients of age 45 and minimum number of patients with age 10.

