Project Report

Cancer Classification Using Deep Learning

1.1 Introduction:

Our goal in this study was to create a deep learning model that could correctly identify various cancer types using histopathology pictures. Accurate categorization is essential for treatment planning and patient care, and cancer diagnosis is a crucial responsibility in healthcare. By utilizing data science methods, we want to create a model that may help physicians correctly identify different forms of cancer.

1.2 Problem statement:

Our goal was to find a reliable way to classify different forms of cancer using histopathological scans. Patients may suffer major repercussions from misclassification or incorrect diagnosis, such as inappropriate treatment regimens or postponed therapies. Thus, in the area of cancer, creating a trustworthy and accurate classification model is crucial.

1.3 Data collection:

We collected two datasets for this project:

Dataset 1: Lung and colon cancer histopathological images
Dataset 2: Skin cancer images with nine different classes

These datasets provide a diverse range of cancer types, enabling us to train and validate our model effectively.

Dataset 1, Histopathological Images of Lung and Colon Cancer: The Department of Pathology at the University of Texas MD Anderson Cancer Center's public archive produced the data for this dataset. The collection is made up of histological pictures of tissue samples from individuals who had surgery between 2012 and 2015, including samples of lung and colon cancer. Using a digital slide scanner, the photos were taken and stored in a high-resolution format (PNG). Adenocarcinoma, large cell carcinoma, squamous cell carcinoma, small cell carcinoma, and colon adenocarcinoma were the five different forms of cancer that were identified on each picture according to the associated tissue sample.

Dataset 2, Skin Cancer ISIC: The International Skin Imaging Collaboration (ISIC), a nonprofit organization that offers a global forum for the sharing of information and skills in the area of dermatology, is where the data for this dataset was gathered. The dataset includes dermoscopic images of skin lesions from patients with a variety of skin conditions, including nine different types of skin cancer: seborrheic keratosis, dermatofibroma, melanocytic nevus, melanoma, basal cell carcinoma, actinic keratosis, benign keratosis, pyogenic granuloma, and vascular lesion. Digital cameras, cellphones, and other devices were used to take the pictures, which were then stored in a high-resolution format (JPEG). Based on the agreement of a panel of highly qualified dermatologists, the diagnosis of the associated skin tumor was indicated in each photograph.

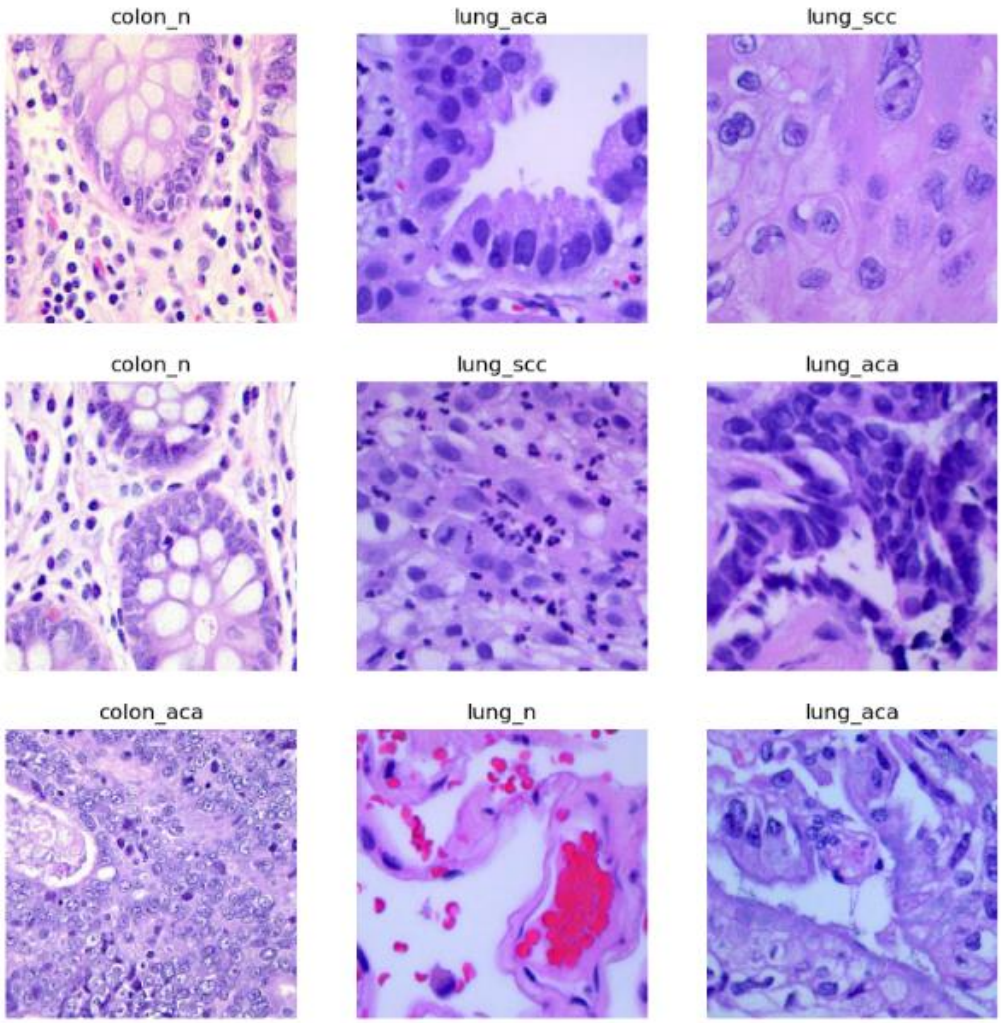
Data Set 1 comprises histological lung and colon cancer images, enabling the development of machine learning models for cancer cell identification. Such models facilitate early detection and diagnosis, enhancing patient treatment. Data Set 2 contains images of skin tumours, aiding in the creation of models for early detection and classification of skin cancer. These datasets serve as crucial resources for developing machine learning tools that assist professionals in cancer identification and treatment, ultimately improving patient outcomes.

2.1 Data analysis and exploration:

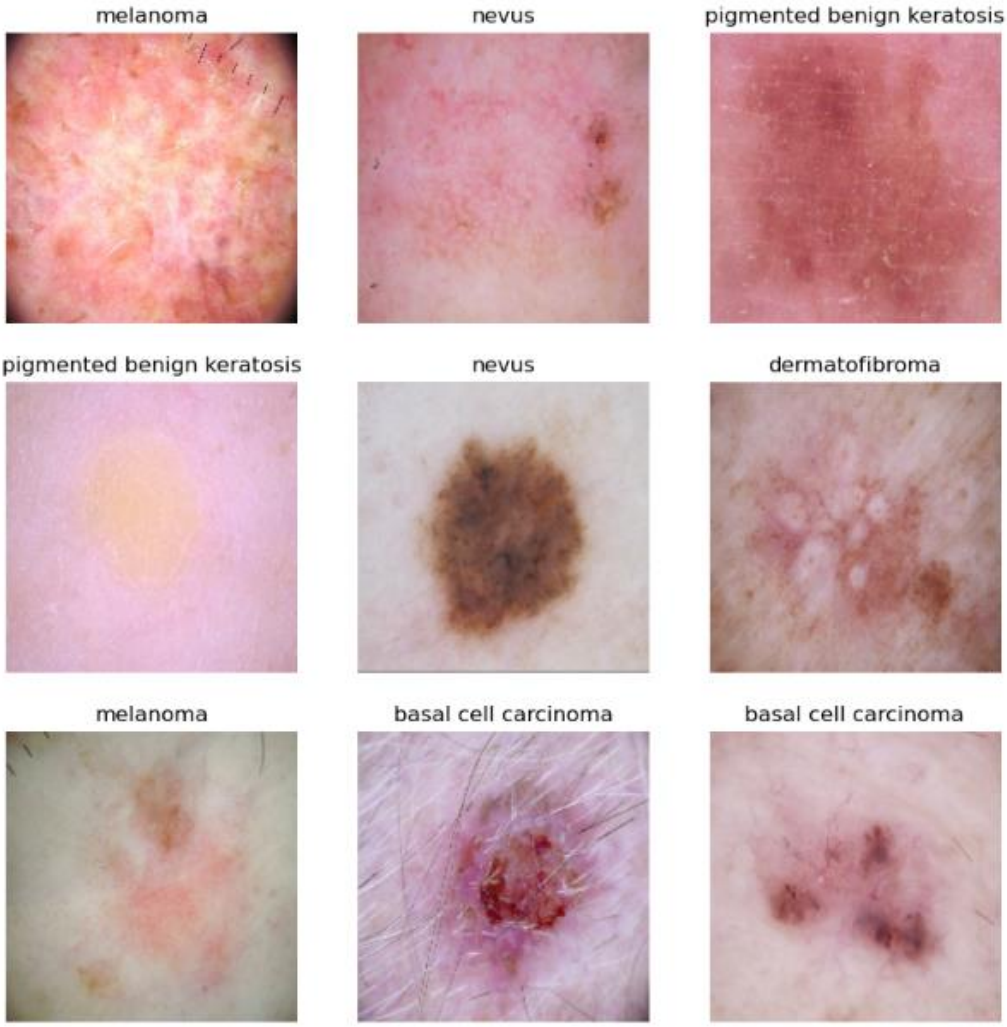
In the initial phase of our project, we performed exploratory data analysis (EDA) to gain insights into the structure and characteristics of the datasets.

Sample Images from the Dataset: To better understand the types of pictures available and their classifications, we start by displaying example photographs from the databases. Here are some sample photos from the corresponding datasets for each class.

Dataset 1: Lung and Colon Cancer Histopathological Images

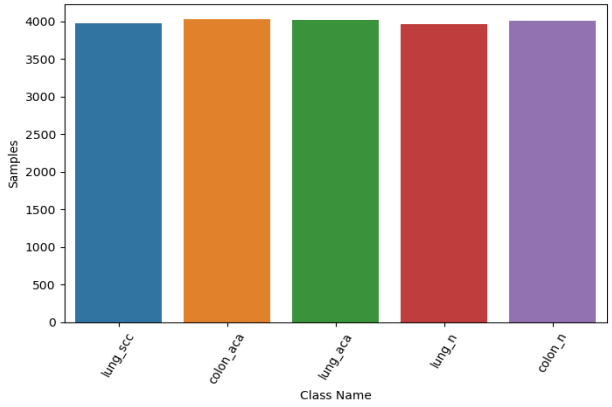


Dataset 2: Skin Cancer Images



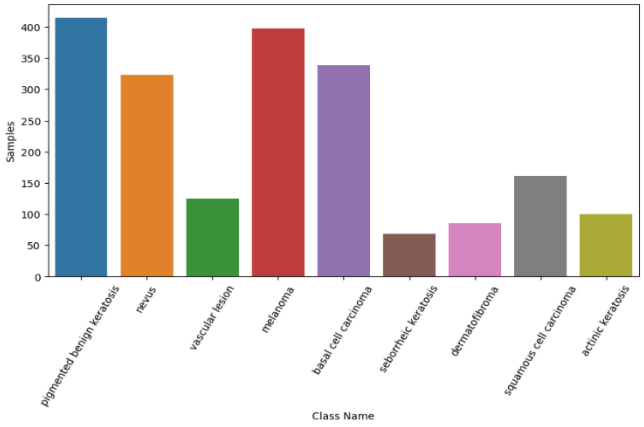
Class Distribution: We began by examining the distribution of classes within each dataset. For Dataset 1, which contained images from five cancer categories, we visualized the class distribution using bar plots. This allowed us to understand the balance or imbalance of samples across different cancer types.

Dataset 1:



	Class Name	Samples
0	lung_scc	3972
1	colon_aca	4029
2	lung_aca	4022
3	lung_n	3988
4	colon_n	4009

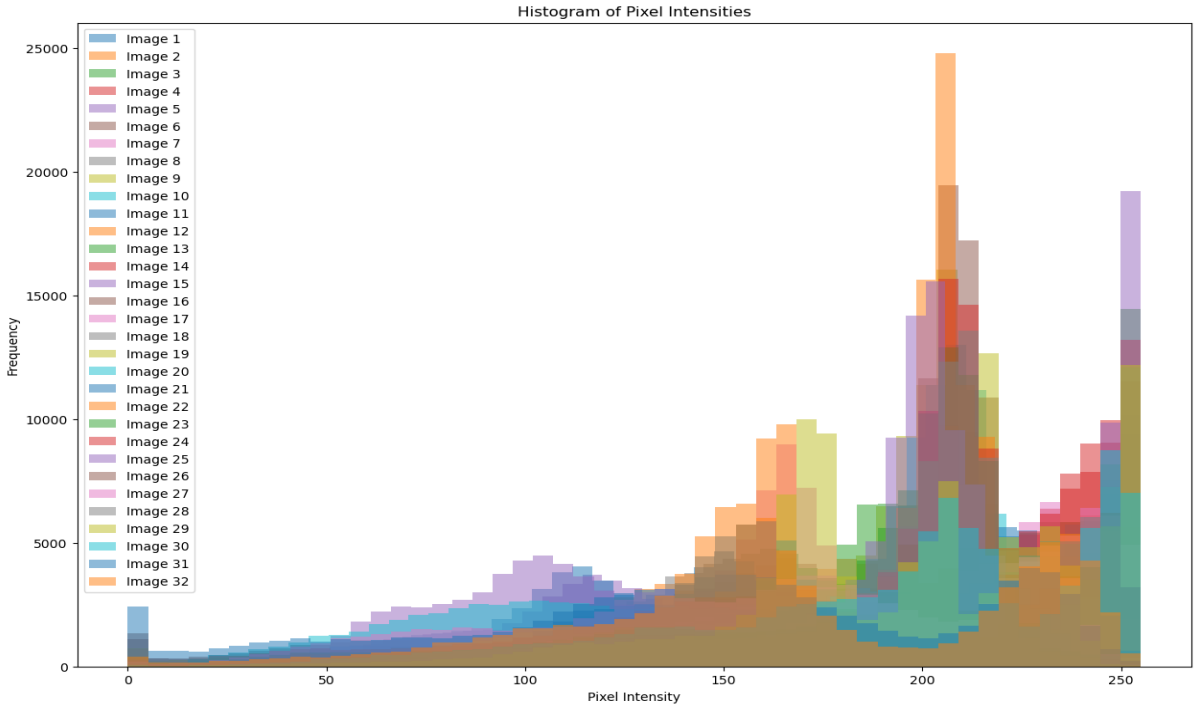
Dataset 2:



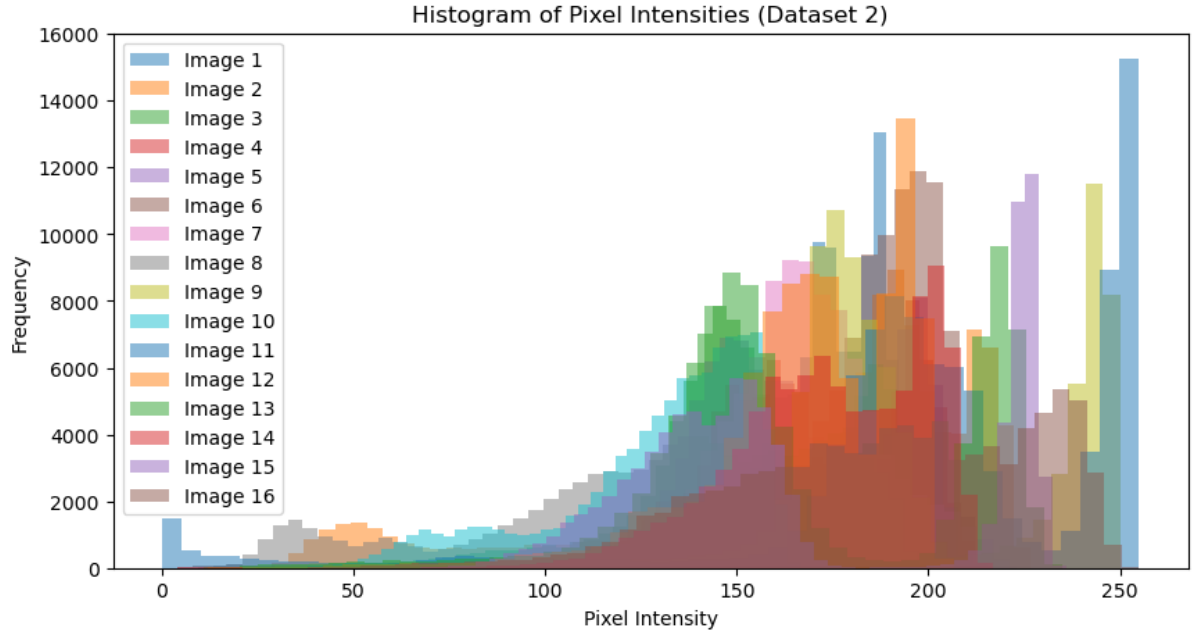
	Class Name	Samples
0	pigmented benign keratosis	415
1	nevus	323
2	vascular lesion	125
3	melanoma	398
4	basal cell carcinoma	339
5	seborrheic keratosis	69
6	dermatofibroma	86
7	squamous cell carcinoma	161
8	actinic keratosis	100

The Pixel Intensities Histogram: By looking at the distribution of pixel intensities in the pictures, we were able to further evaluate the data. Plotting pixel value histograms allowed us to find common intensity ranges and possible outliers in our investigation. Comprehending the pixel intensities' distribution aided us in selecting suitable pre-processing measures, such rescaling or normalization.

Dataset1:



Dataset 2:



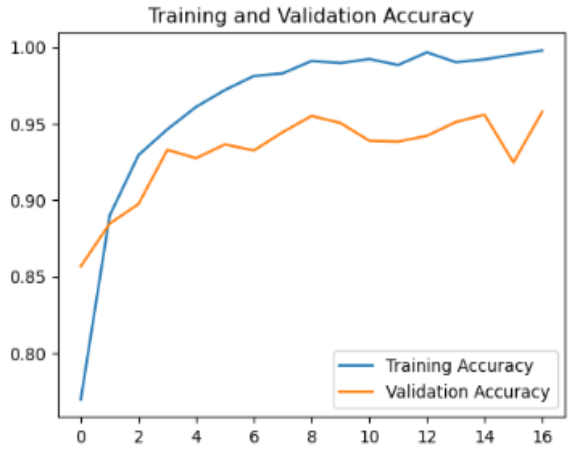
3.1 Approach:

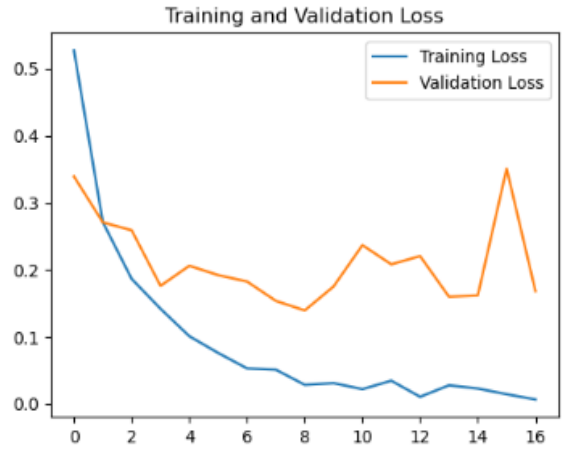
Model Development on the first dataset: Our classification model was created using the TensorFlow/Keras convolutional neural network (CNN) architecture. Multiple convolutional and pooling layers preceded fully linked layers in the model. Using Dataset 1, which includes pictures from five distinct cancer classifications, we trained the model.

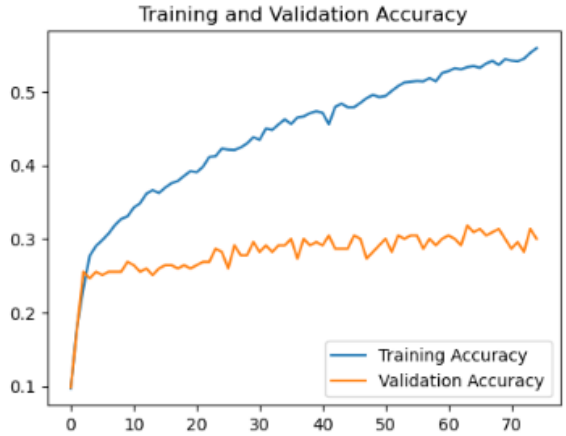
Transferred the learning on the second dataset: We used transfer learning to Dataset 2 in order to make use of the information obtained from dataset 1. By modifying the output layer to accommodate the nine classes found in Dataset 2, we improved the pre-trained model from the first dataset. We were able to apply the acquired features to the new dataset and reuse them because of this method.

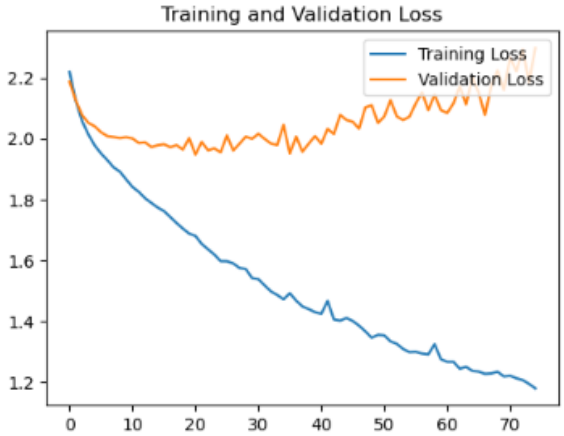
4.1 Results and analysis:

Model Performance on dataset 1: Our model performed well on Dataset 1, achieving over 99% accuracy on the training set and over 96% accuracy on the validation set. The model proved to be reliable in correctly diagnosing different forms of cancer.









5.1 Conclusion:

In summary, we successfully developed a deep learning model for the categorization of cancer through the use of histopathology pictures. On Dataset 1, our model showed excellent accuracy and offered insightful information on cancer diagnosis. But to enhance the transfer learning model's performance on Dataset 2, additional research and improvement are needed. All things considered, our experiment demonstrates how data science and deep learning may improve cancer detection and treatment.