**Material & Method:**

**HAM10000 Dataset**

The HAM10000 dataset is a large dataset containing 10,000 skin pigmented lesion images across seven different classes that are crucial for diagnosing skin cancer. This dataset was created by Tschandl and his team to overcome the inadequacy of smaller datasets used for classification. The images in this dataset have different resolutions and were organized, cleaned, and prepared to train a neural network effectively due to their diverse nature. The final version of the dataset contains 10,015 images and is available for academic research purposes on the ISIC archive. An expert pathologist in dermoscopy confirmed the accuracy of the dataset's ground truth.

The data set is composed of two primary parts. The first is a metadata file that contains particular details about cancerous lesion images, including the skin lesion's location, patient age and gender, lesion diagnosis, and the skin lesion directory. The second and more critical part of the data set consists of visual files. The goal of this study is to classify skin lesions based entirely on digital images, using the visual files in the data set.

The seven important diagnosis classes in the HAM10000 dataset are:

* Melanocytic nevi
* Melanoma
* Benign keratosis-like lesions
* Basal cell carcinoma
* Actinic keratoses
* Vascular lesions
* Dermatofibroma

The dataset contains 10,015 images in total, with 327 images in the "akiec" class, 541 images in the "bcc" class, 1,099 images in the "bkl" class, 155 images in the "df" class, 6,705 images in the "nv" class, 1,113 images in the "mel" class, and 142 images in the "vasc" class.Top of Form

**Actinic Keratosis (akiec):**

Actinic Keratosis is a prevalent and non-invasive type of carcinoma that is considered to be an early indicator of skin cancer rather than an actual cancer. It is a subtype of squamous cell carcinoma that can be treated locally without requiring surgery. If left untreated, it may develop into an invasive squamous cell carcinoma. This type of skin lesion is commonly found on the face and is caused by prolonged exposure to UV light.

**Basal Cell Carcinoma (bcc):**

Basal cell carcinoma is a type of skin cancer that originates in the basal cells responsible for producing new skin cells. It is the most common form of skin cancer and is more likely to occur in areas exposed to direct sunlight, such as the head and neck. The cancer usually appears as pink growths, recurrent sores, or red patches on the skin. The lesions typically develop slowly and do not spread easily.

**Benign Keratosis-Like Lesions (bkl):**

The BKL category in the database comprises of three types of lesions that are not cancerous. These lesions are Lichenoid Keratosis, Solar Lentigo, and Seborrheic Keratosis. Lichenoid Keratosis is a harmless skin condition that typically appears as a small, single, grey-brown lesion on the chest and upper limbs. Solar Lentigo is a type of hyper-pigmented infection that varies in size from a few millimeters to over one centimeter. Seborrheic Keratosis is a benign condition that usually appears on the back, collar, scalp, and chest. It is reddish-brown or grayish-brown in color and generally does not require extensive treatment.

**Dermatofibroma (df):**

Dermatofibroma is a common skin condition that typically affects adolescents or older individuals, with no significant gender difference. Clinically, dermatofibroma appears as firm nodules, patches, or bumps with a smooth surface and a color that can vary from light brown to dark brown, reddish-purple, or yellow. These benign skin lesions commonly appear on the upper arm, upper back, and lower leg.

**Melanocytic Nevi (nv):**

There are seven subclasses in the database that encompass all of the benign skin tumours known as melanocytic nevi, which can have various appearances. These nevi are caused by the growth of melanocytes, the pigment-producing cells of the skin, and are often a result of exposure to UV radiation during early childhood.

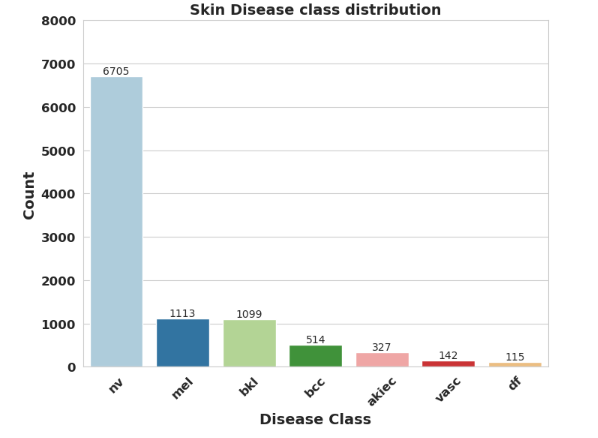
**Vascular Lesions (vasc):**

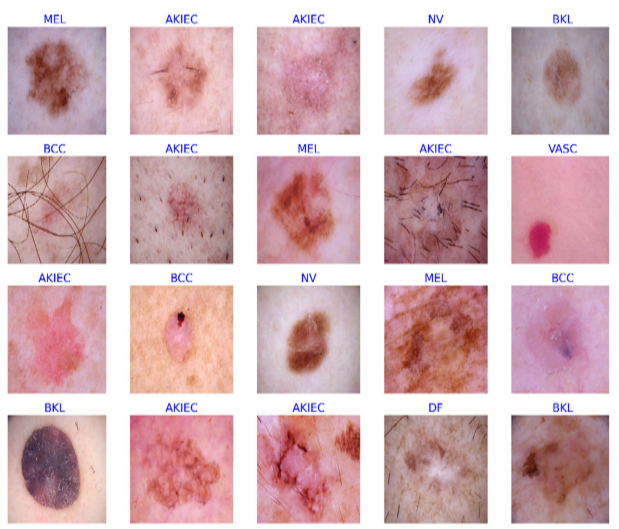
Most cases of vascular lesions (vasc) are hereditary, but they can also develop later in life and are typically benign. These lesions can take on a variety of appearances and are characterized by the formation of sores on the skin and surrounding tissues. They are sometimes referred to as birthmarks.

**Melanoma (mel):**

Melanoma is a cancerous growth that can take on a variety of forms, originating from malignant melanocytes. If detected early, it can be treated with a simple surgical procedure. Melanomas may be either invasive or non-invasive and are most commonly found on sun-exposed areas of the body such as the face, trunk, hands, neck, and legs. The identification of melanoma is based on irregularly shaped patches with uneven borders and varying colors, typically larger than 6 mm and prone to growth. Failure to treat melanoma can result in it spreading to other organs and causing death.

The HAM10000 dataset includes seven distinct classes, as described previously, and the number of images in each class is provided in the graph. However, the distribution of images is imbalanced, which can affect the performance of machine learning models trained on the dataset. To address this issue, data augmentation techniques such as oversampling are employed, as detailed in the pre-processing section.

****

****

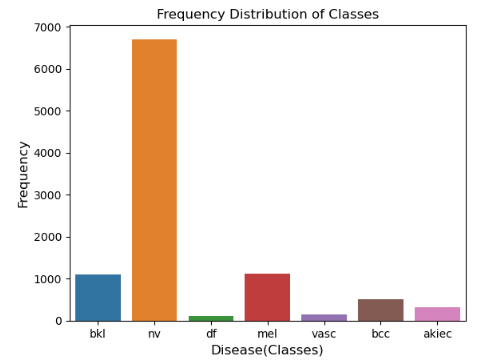
**Proposed Methodology:**

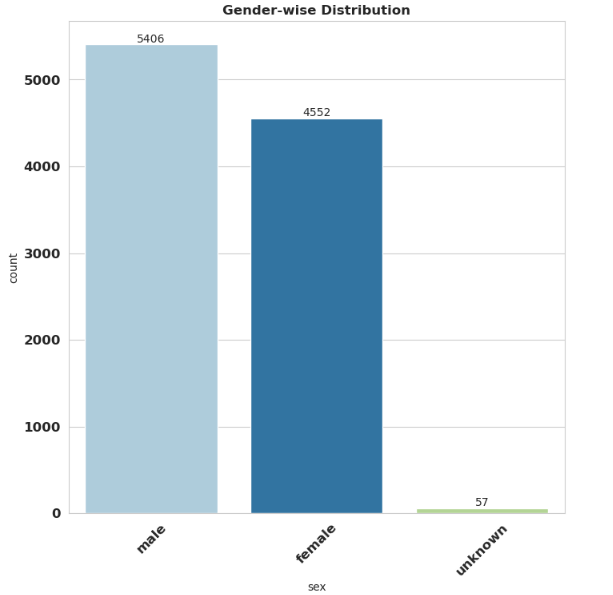
**Data Preprocessing:**

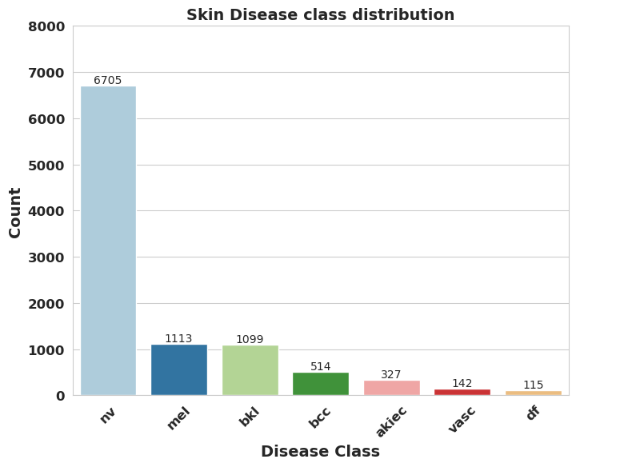
**Image Preprocessing:** Image Pre-processing is a crucial step in working with clinical image data, particularly for training convolutional neural network (CNN) based systems. In this study, the authors aimed to improve the generalizability of their CNN design and reduce time-consuming preprocessing steps. They applied two popular techniques: image resizing and image normalization. Image scaling was used to account for the variations in image size and intensity. Image normalization was necessary as some images in the dataset were obtained from different sources and exhibited variations in pixel intensity due to undesirable artifacts. To overcome the problem of variations in image contrast, the authors normalized the contrasts of the training images during the training process. This was achieved by dividing each pixel value by 255 and setting the image intensity values to range between [-1 and 1].

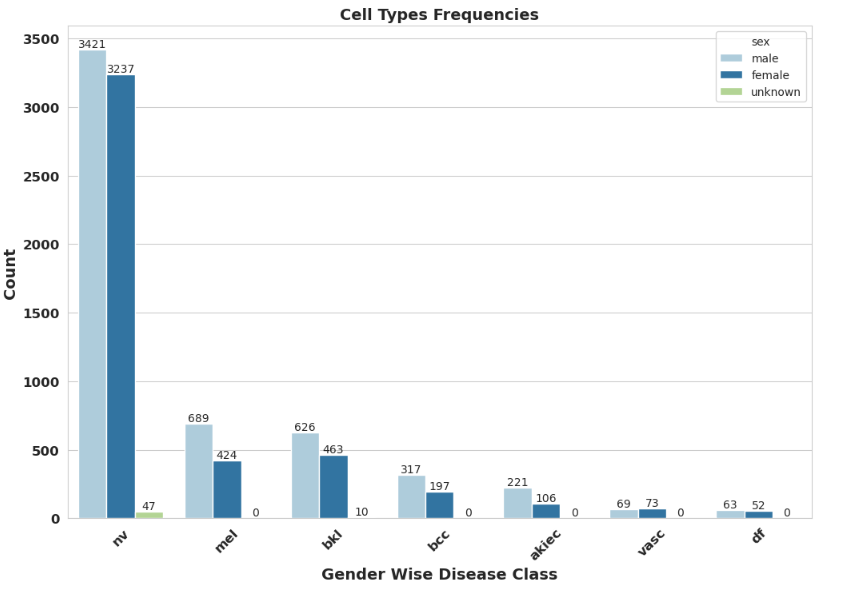
**Meta-Data Preprocessing** **(Data Cleaning & Normalization):** This paragraph describes the metadata pre-processing step. This step involves removing any missing data from the clinical information. Since most demographic features are categorical variables represented as strings or categories, they are converted to the categorical data format using one-hot encoding. For each level of a categorical feature, a new variable is created. Each category is mapped with a binary variable containing either 0 or 1, with 0 representing the absence and 1 representing the presence of that category. Numeric demographic features, such as age, are also normalized.

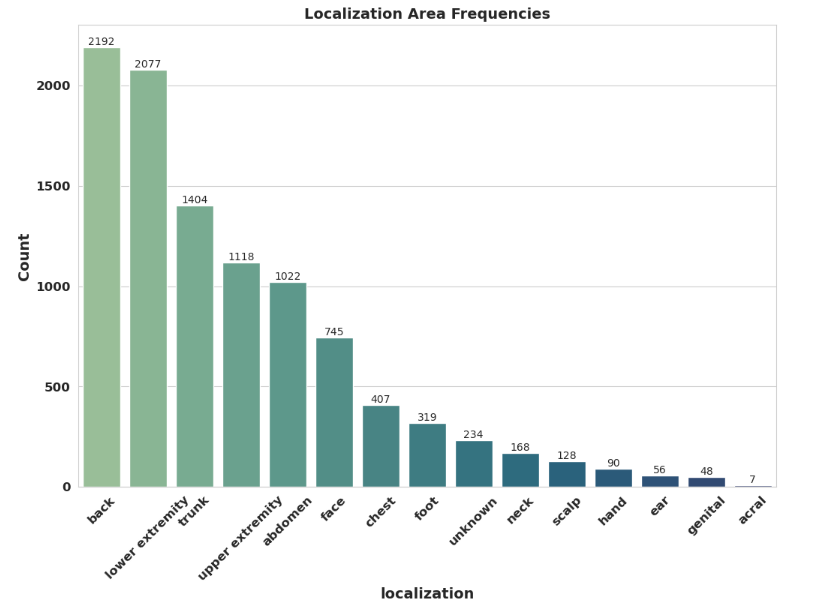
**Data Visualization:** Data visualization is the process of creating visual representations of data using Python libraries such as Matplotlib, Seaborn, Plotly, and others. Data visualization helps in understanding the patterns and relationships within the data by converting raw data into visual representations such as graphs, charts, histograms, scatterplots, and more. It allows us to gain insights into the data and present it in a more understandable and visually appealing way. Python's data visualization libraries provide a wide range of tools and techniques to create high-quality visualizations that can help in making better data-driven decisions, for data visualization of meta data and images matplotlib and seaborn are used.

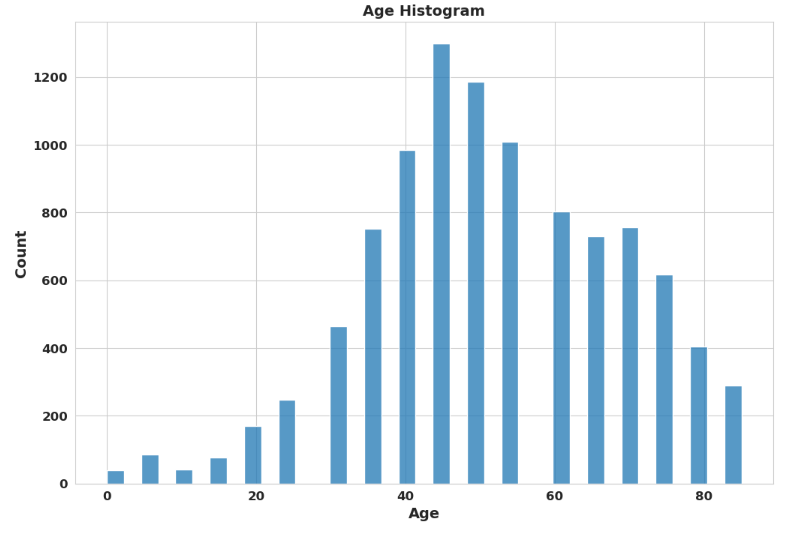
****

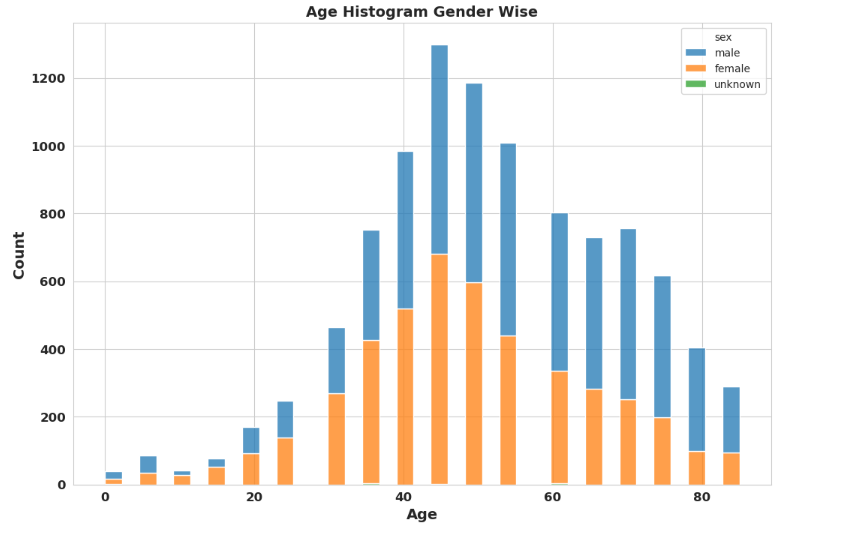
****

****

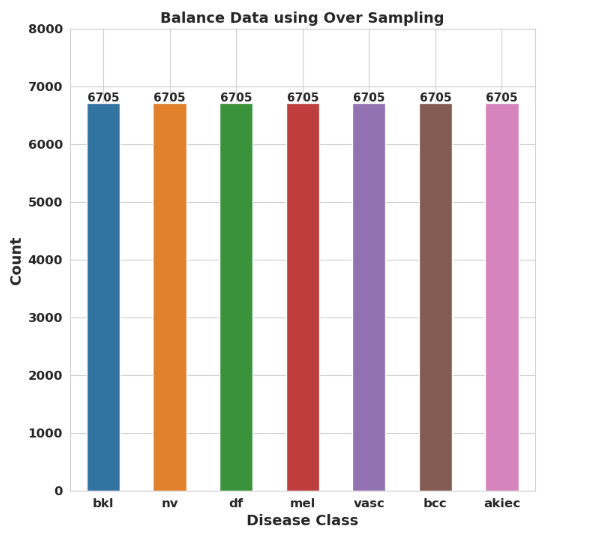
****







**Data Balancing:** Data balancing refers to the process of adjusting the class distribution of a dataset to reduce class imbalance. Class imbalance occurs when one or more classes in a dataset have significantly fewer instances compared to other classes. This can lead to biased model predictions and poor performance, especially in machine learning algorithms that are sensitive to class distribution, as you can see in the class distribution graph the data is highly imbalance ,so for data balancing we have adopt data over sampling technique to balance the class distribution. Before data balancing there are total 10015 images ,after balancing image goes increase to 46935 and each class have 6705 images. Data balancing is an important step in preprocessing for machine learning tasks, especially in medical or financial domains where the cost of false predictions can be high.



**Data Splitting:** Data splitting is the process of dividing a dataset into two or more subsets for the purpose of training and testing a machine learning model. The most common approach is to split the dataset into a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate the model's performance on unseen data.In addition to the training and testing sets, it is common to split the data further into a validation set. The validation set is used during the training process to monitor the model's performance and to fine-tune the model's hyperparameters.

Data splitting is an important step in machine learning because it helps to prevent overfitting, which occurs when a model is trained too well on the training data and is not able to generalize well to new, unseen data. By evaluating the model's performance on a separate testing set, we can get a better estimate of how well the model will perform on new data. In out dataset we split the complete image data set in to training, testing with the ratio of 80 & 20. There are total 46935 images ,out of this 37548 images are used as training and 9387 images are used as testing data. The training data. Further 20% from the training data is used as validation data while training the model.

**Image Downsampling:** The dataset consists of two main components. The first is a metadata file containing specific information about cancer lesion images, such as the location of the skin lesion, patient age and gender, diagnosis of the lesion, and the directory for the skin lesion. The second and more important section of the dataset is made up of visual files. The objective of this study is to classify skin lesions based solely on digital images. To accomplish this, the data file was reorganized to include only the lesion type and image file directory, and each lesion's textual label was transformed into a digital value between 0 and 6.

The original dermoscopic images in the dataset have a resolution of 600 x 450 pixels and are saved in the RGB format. However, it was found that processing time increases with larger image sizes. Therefore, to speed up processing time, all samples in the dataset were resized to 28 x 28 pixels.

**Proposed Model (TRNet):**