# SKIN CANCER IMAGE CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORKS

**Group Members**
-Kothari Krunal -Mathkar Aishwarya -Patel Krutarth -Vora Darshit -Zunjarrao Tanvi

**Problem Statement:** Skin Cancer is one of the most deathful of all the cancers. It is bound to spread to different parts of the body on the off chance that it is not analyzed and treated at the beginning time. It is mostly because of the abnormal growth of skin cells, often develops when the body is exposed to sunlight. Considering the limited availability of the resources, early detection of skin cancer is highly important. Accurate diagnosis and feasibility of detection are vital in general for skin cancer prevention policy. Skin cancer detection in early phases is a challenge for even the dermatologist.

**Objective:** The objective is to propose a system that detects skin cancer and classifies it in different classes by using the Convolution Neural Network and to analyze the result to see how the model can be useful in practical scenario.

**Audience:** The target audience of our project are the healthcare sectors, researchers, etc.

**Dataset Description:** We used the MNIST HAM-10000 dataset for Skin Cancer which is available on Kaggle. It consists of 10,015 images of skin pigments and 7 features. The dermatoscopic images are divided amongst seven classes. The number of images present in the dataset is enough to be used for different tasks including image retrieval, segmentation, feature extraction, deep learning, and transfer learning, etc.
The various features of our dataset are:

1. Lesion id
2. Image id
3. Dx: This column includes the types of cancers. Our dataset focuses on 7 major types of skin cancers:
    a. Melanocytic nevi (nv)
    b. Melanoma (mel)
    c. Benign keratosis (bkl)
    d. Basal cell carcinoma (bcc)
    e. Actinic keratoses (akiec)
    f. Vascular lesions (vasc)
    g. Dermatofibroma (df)
4. Dx_type: This column basically tells us how the cancer was validated. It includes:
    a.Histopathologic (hist)
    b. Follow-up
    c. Confocal
    d. Consensus:
5. Age: This column basically tells us the age of the patients.
6. Sex: Tells us whether the patient was male or female.
7. Localization: It is basically the part of the body where the cancer was present like back, lower extremity, trunk, upper extremity, abdomen, face, chest, foot, unknown, neck, scalp, hand, ear, genital and acral.

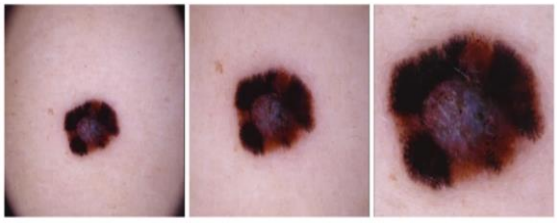
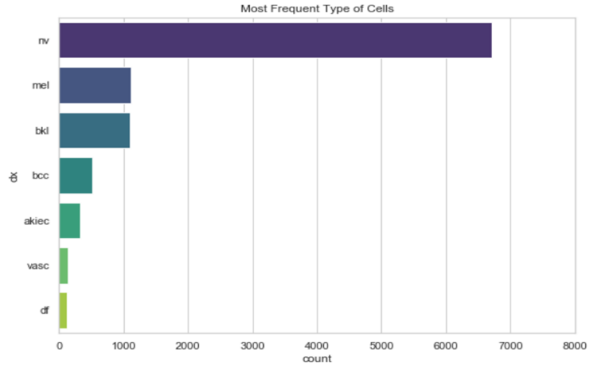
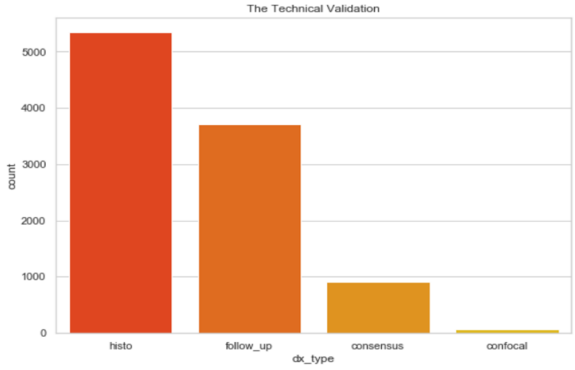


Figure1: Example of same lesion

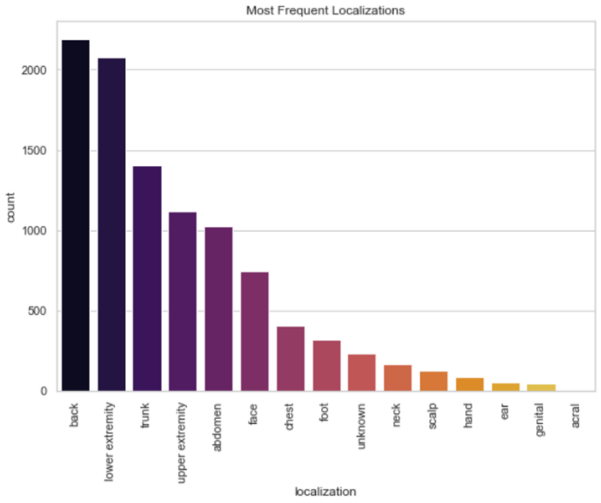
## Exploratory Data Analysis:

**1.Bargraph -** Type of skin cancer, Technical validation, Localization and Sex

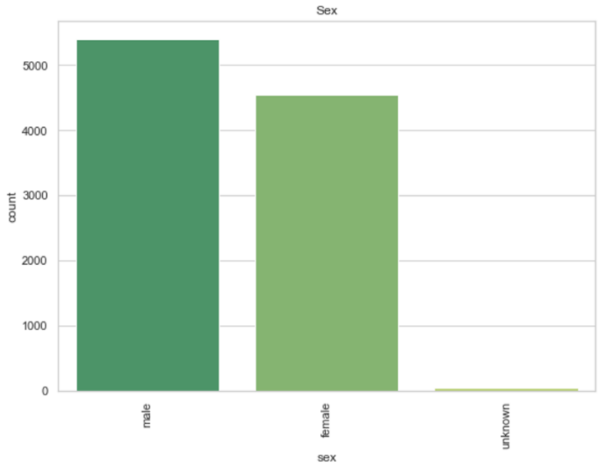


Plot 1




Plot 2



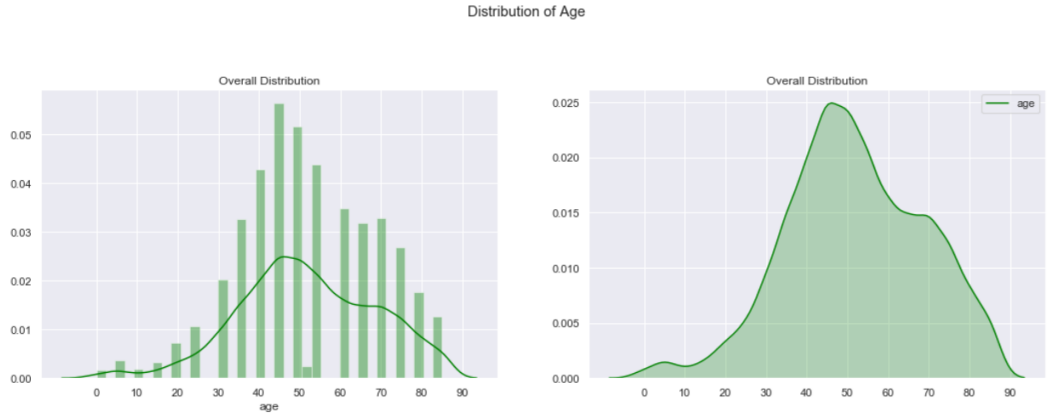

Plot 3

Plot 4

**Observations:**

- Plot 1: It can be seen that Melanocytic Nevi(nv) is around twice more than all of the other types combined with the count more than 6000.
- Plot 2: We can observe that histo(Histopathologic) and follow up are the most common ways by which the cancer was validated.
- Plot 3: Back, lower extremity and trunk are heavily compromised regions of skin cancer.
- Plot 4: There are slightly more males detected with cancer than the females.

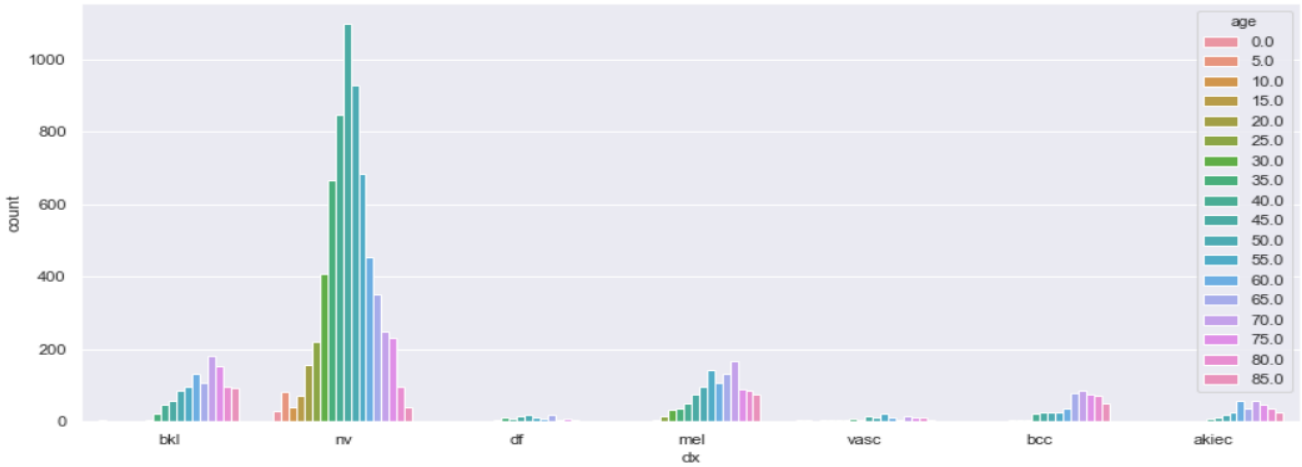
**2. Univariate Distribution**: Age




**Observation:**

- It can be observed that most patients are in the age group of 35 to 65 with most in the range of 40-55.

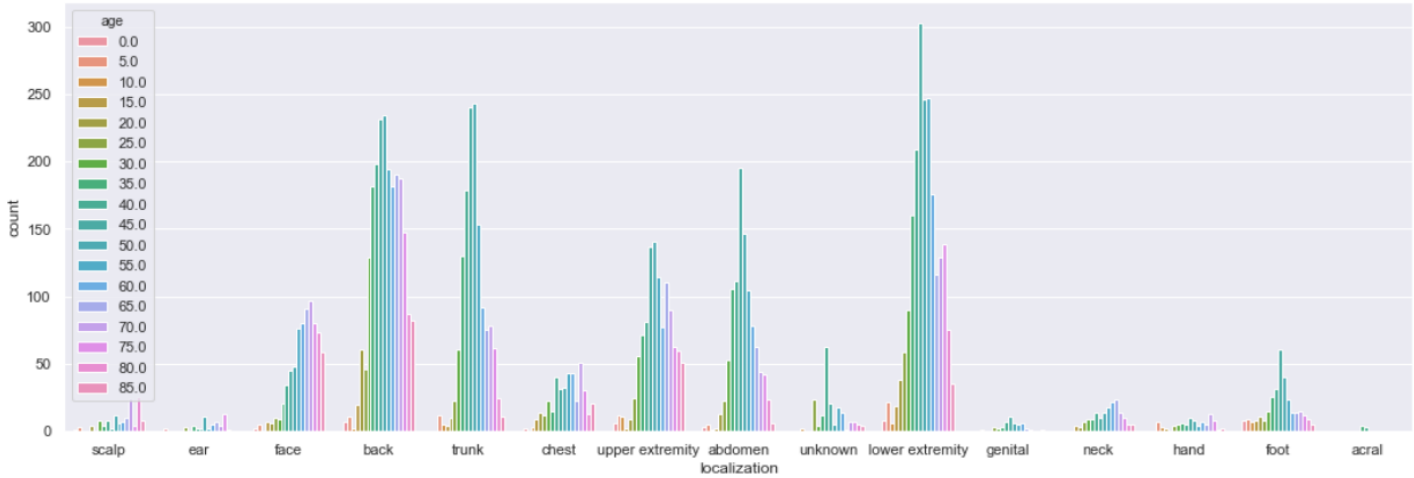
**3.Countplot**: Correlation between Types of Cancer and Age




**Observation:**

- It can be seen that Melanocytic Nevi is more prevalent with the age ranging from 35 to 45.
- Dermatofibroma is lowest seen skin cancer.

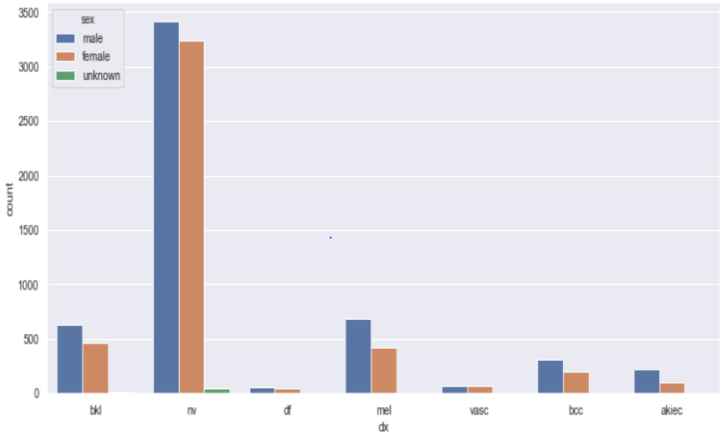
**4.Countplot**: Correlation between Localization and Age

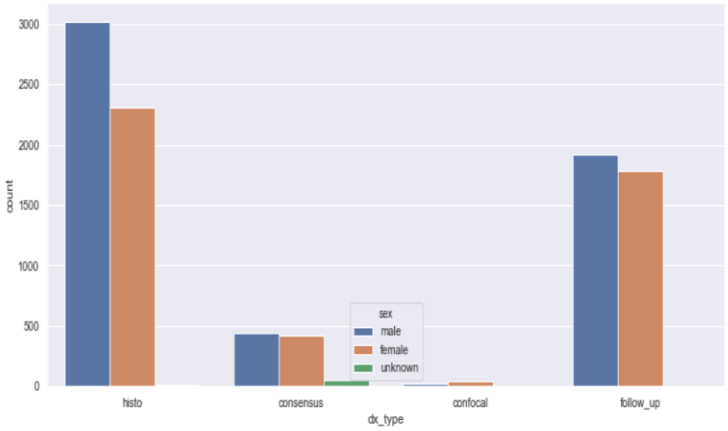



**Observations:**

- It is seen that cancer is found more in lower extremity then in back followed by trunk.
- Mostly observed in age from 40 to 55.

**5.Countplot**: Correlation between Types of Cancer with Sex & Correlation between Technical Validation with Age

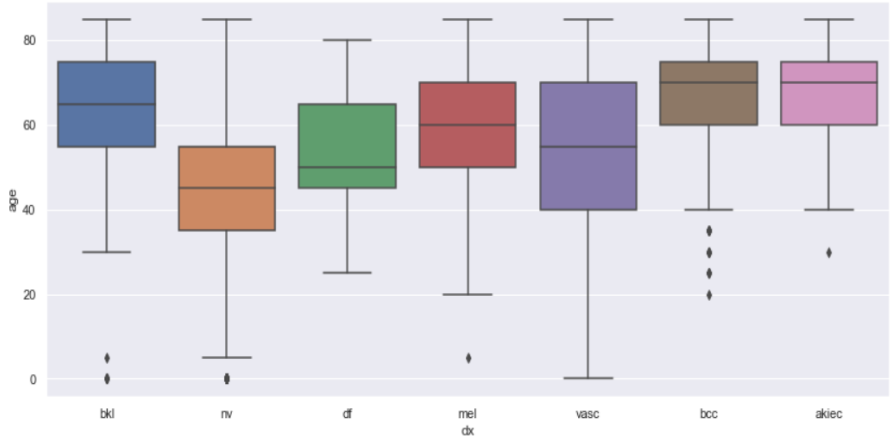



**Observations:**

- In dx vs sex, for both males and females Melanocytic Nevi (nv) has highest count.
- In technical validation vs age, we can observe that histo and follow up are the most used methods of validation and number is greater in males than females.

**6. Boxplot** - Types of Cancer and Age




**Observation:**

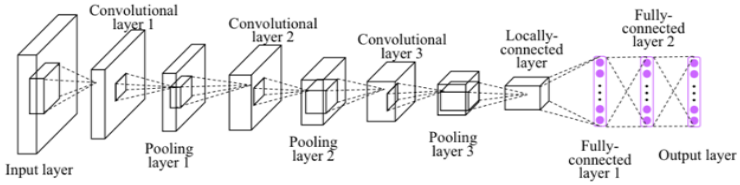
- It can be seen that skin cancers are not much prevalent below the age of 20 and there are some outlier's inn bkl, nv, mel, bcc and akie types of cancers.

**Data Preprocessing:**

- Remove the Null Values
- Remove the Duplicates
- Splitting the dataset into Training and Testing sets in ratio of 80: 20, respectively.
- Convert categorical columns into Numerical columns: One Hot Encoding and Label Encoding
- Resizing the Images: Resize the images as the original dimension of images is large, and the processing them takes very long.
- Normalization

**CNN Architecture**

1. Convolution layer -Conv2D
2. Pooling layer MaxPooling2D
3. Flatten layer
4. Fully connected layer -Dense




**Model Building:** We have used the Keras Sequential API, where you have just to add one layer at a time, starting from the input and OpenCV for Image Processing.

```
model = Sequential()
model.add(Conv2D(32, kernel_size=(3, 3),activation='relu',input_shape=(100,100,3),padding='same'))
model.add(MaxPooling2D((2, 2),padding='same'))
model.add(Dropout(0.20))

model.add(Conv2D(64, (3, 3), activation='relu',padding='same'))
model.add(MaxPooling2D(pool_size=(2, 2),padding='same'))
model.add(Dropout(0.40))

model.add(Conv2D(128, (3, 3), activation='relu',padding='same'))
model.add(LeakyReLU(alpha=0.1))
model.add(MaxPooling2D(pool_size=(2, 2),padding='same'))
model.add(Dropout(0.20))

model.add(Flatten())

model.add(Dense(64, activation='linear'))
model.add(LeakyReLU(alpha=0.1))
model.add(Dense(128, activation='linear'))
model.add(Dense(256, activation='linear'))
model.add(Dense(7, activation='softmax'))
model.summary()
```

1) Convolution Layer

- It is like a set of learnable filters. We choose to set 32 filters for the firsts conv2D layers, 64 and 128 filters for the two last ones.
- Each filter transforms a part of the image (defined by the kernel size) using the kernel filter. The kernel filter matrix is applied on the whole image. Filters can be seen as a transformation of the image.
- The CNN can isolate features that are useful everywhere from these transformed images (feature maps).

2) MaxPool2D layer

- This layer simply acts as a down sampling filter.
- It picks the maximal value.
- These are used to reduce computational cost, and to some extent also reduce overfitting.
- Combining convolutional and pooling layers, CNN can combine local features and learn more global features of the image.

3) Dropout

- It is a regularization method, where a proportion of nodes in the layer are randomly ignored (setting their weights to zero) for each training sample.
- This drops randomly a proportion of the network and forces the network to learn features in a distributed way.
- This technique also improves generalization and reduces the overfitting.

5) Flatten

- This layer is used to convert the final feature maps into a one single 1D vector.
- This flattening step is needed so that you can make use of fully connected layers after some convolutional/maxpool layers.

6) In the end we used the features in the Fully Connected (FC-Dense) layers which is just artificial an neural networks (ANN) classifier. In the last layer (Dense (10, activation="softmax")) the net outputs distribution of probability of each class

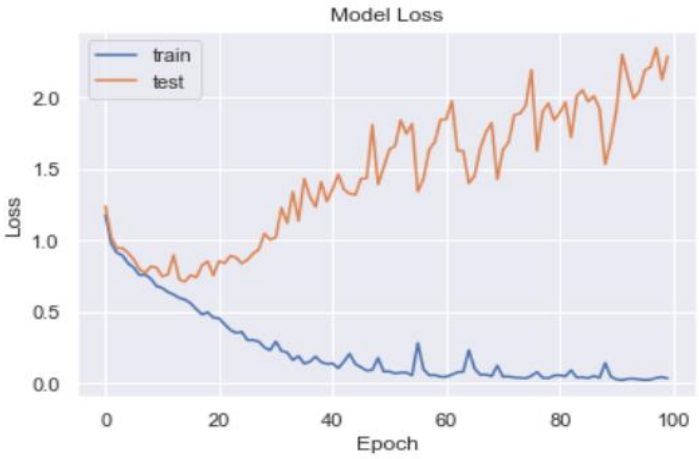
### Setting Optimizer and Annealer

```
model.compile(optimizer='adam',loss='categorical_crossentropy',metrics=['accuracy'])
history=model.fit(train_X,one_hot_train,batch_size=128,epochs=10,validation_split=0.2)
```

- Once our layers are added to the model, we need to set up a score function, a loss function and an optimization algorithm.
- We define the loss function to measure how poorly our model performs on images with known labels.
- It is the error rate between the observed labels and the predicted ones. I used a specific form for categorical classifications (>2 classes) called the "categorical_crossentropy".
- The optimizer function will iteratively improve parameters like filters kernel values, weights and bias of neurons. It helps to minimise the loss.
- We have choose Adam optimizer because it combines the advantages of Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp).

### Model Evaluation

As the number of epochs increases, there is decrease in loss and increase in accuracy.





| No of Epochs | Accuracy rate |
|---|---|
| 10 | 76 |
| 50 | 87 |
| 100 | 98 |

**Conclusion:** Accuracy is higher if model is trained on more samples of lower resolution than small samples of high resolutions. We have achieved the accuracy of 98% with 100 epochs.

**Future scope:** Going forward, we can continue to refine the model to achieve a stable decrease in loss function with every epoch, build an interface such that given an image of a skin lesion within the two classes, the output will give a % probability of which of the seven classes it belongs to.