Melanoma Skin Cancer Detection

Madhu Babu Sikha   Zeming Zhang   Ajith Kumar Ethirajulu   Deepak Raj Mohan Raj

NEED FOR MELANOMA DETECTION

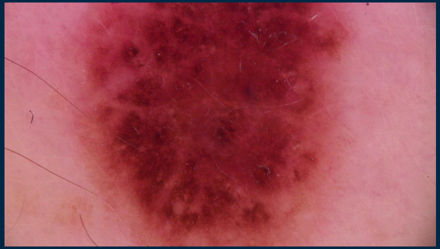
- Melanoma is the least common skin cancer, but responsible for 75% of skin cancer deaths.
- Estimated new cases in 2022: 99,780
- Estimated deaths in 2022: 7,650
- Detection in early stages helps in effective treatment and can save lives.

OBJECTIVE

- The objective of this project is to predict whether a patient has Melanoma, given a lesion image.
- A binary classification problem to classify a given image as Benign (non-cancerous) or Malignant (cancerous).

DATASET

- Referred from the official dataset of the SIIM-ISIC Melanoma Classification Challenge.
- Total training images: 33126
- Features:
  1. image_name - unique identifier, points to filename of related image
  2. patient_id - unique patient identifier
  3. sex - the sex of the patient
  4. age_approx - approximate patient age at time of imaging
  5. anatom_site_general_challenge - location of imaged site
  6. diagnosis - detailed diagnosis information
  7. benign_malignant - indicator of malignancy of imaged lesion
  8. target - binarized version of the target variable



Model design flow

01

02

03

04

DATA LOADING

Skin lesions images
as training data

CLEANING & EDA

Removing columns that are
not required for our model,

MODEL

Compiling and training

ACCURACY

Hyper parameter fine
tuning

CHALLENGES OF THE DATASET

HIGHLY IMBALANCED DATASET

- 32524 Benign images occupies 98.23% and Malignant-584 images are of 1.763%
- Since the percentage of malignant images are very less, we were unable to proceed with the conventional methods.
- For this project, we have considered the images as the input and the target column as the output to be predicted.
- Hence, the integrity of the other columns were not validated
- The image and the target column have no missing data or NA values.

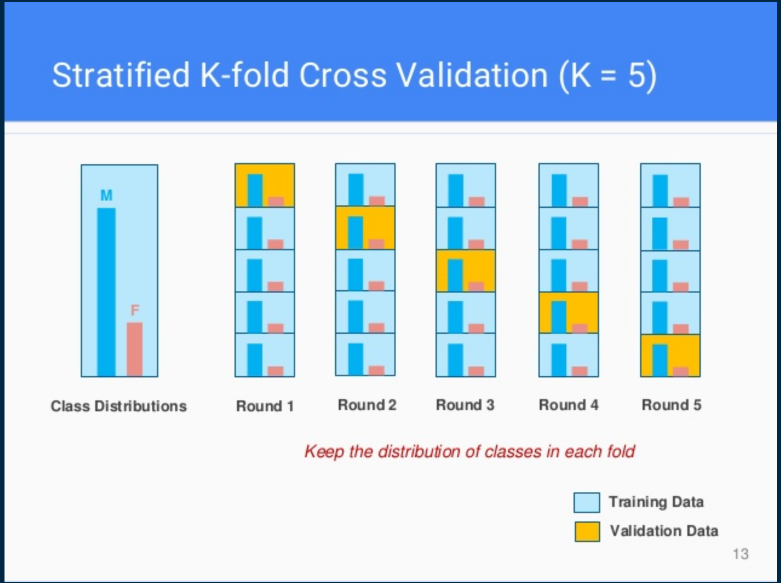
OUR SOLUTIONS FOR DATA IMBALANCE

STRATIFIED SAMPLING

IMAGE AUGMENTATION

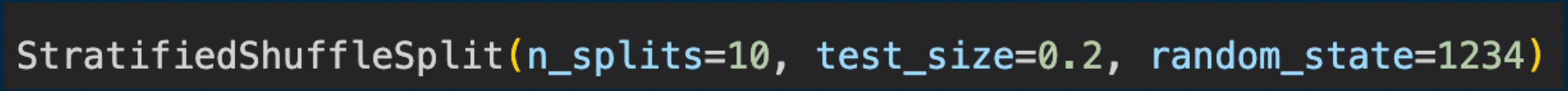
STRATIFIED SAMPLING

- In regular train-test split, the split will usually by completely random, which is not good for imbalance datasets
- Stratified sampling splits the data such that the ratio between the target classes is the same as it is in the full dataset.
- Using stratified k-fold has shown improvement in accuracy.



Stratified K-fold Cross Validation (K = 5)

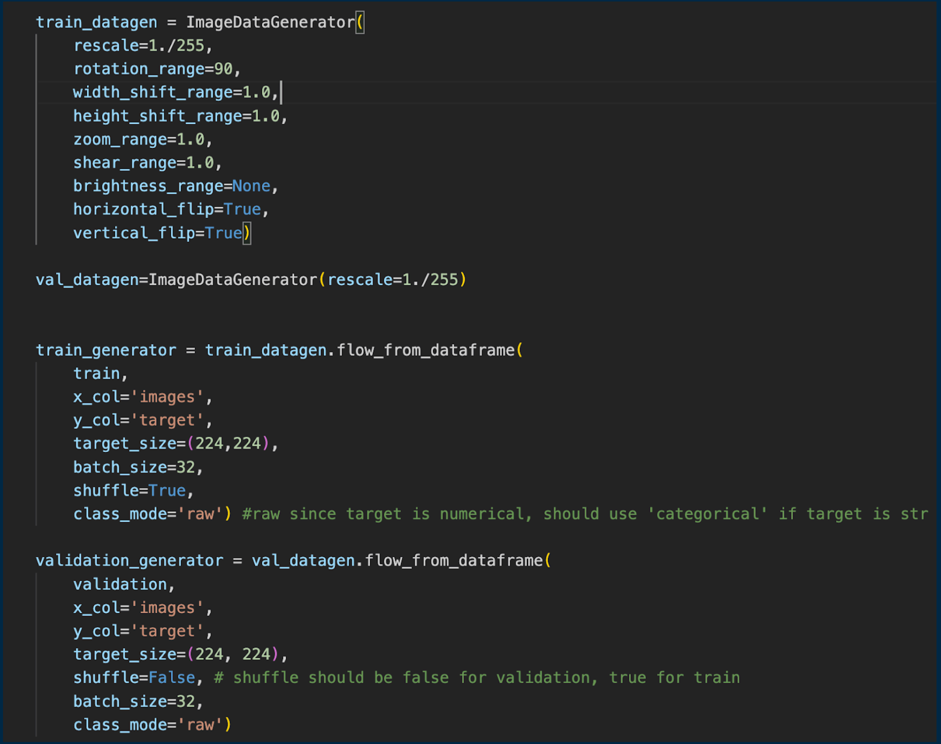
Keep the distribution of classes in each fold



```
StratifiedShuffleSplit(n_splits=10, test_size=0.2, random_state=1234)
```

IMAGE AUGMENTATION

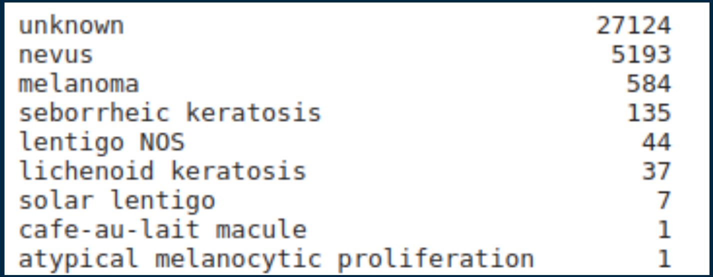
- Image augmentation is the artificial way of creating new images from existing images by applying geometrical transformations(horizontal, vertical) and also applying noise like gaussian blur, brightness and scaling.



```
train_datagen = ImageDataGenerator(
    rescale=1./255,
    rotation_range=90,
    width_shift_range=1.0,
    height_shift_range=1.0,
    zoom_range=1.0,
    shear_range=1.0,
    brightness_range=None,
    horizontal_flip=True,
    vertical_flip=True)

val_datagen=ImageDataGenerator(rescale=1./255)


train_generator = train_datagen.flow_from_dataframe(
    train,
    x_col='images',
    y_col='target',
    target_size=(224,224),
    batch_size=32,
    shuffle=True,
    class_mode='raw') #raw since target is numerical, should use 'categorical' if target is str

validation_generator = val_datagen.flow_from_dataframe(
    validation,
    x_col='images',
    y_col='target',
    target_size=(224, 224),
    shuffle=False, # shuffle should be false for validation, true for train
    batch_size=32,
    class_mode='raw')
```

DATA CLEANING

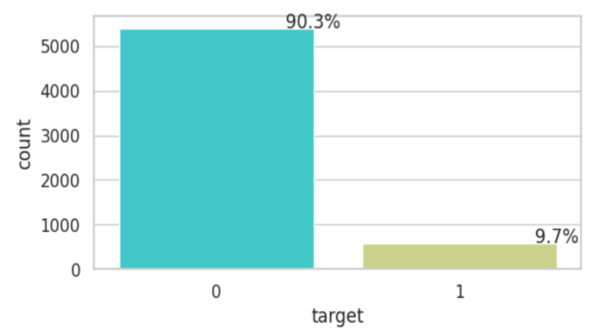
- In diagnosis feature, there are different types of skin diseases, along with Melanoma.
- One of the values in ‘diagnosis’
- feature is unknown and all the
- images corresponding to the ‘unknown’ belongs to benign.
- Hence we have deleted all images corresponding to unknown category, along with last three diagnosis types as they are less in count and are of unknown category.
- Now, Total Images in training set: 5993



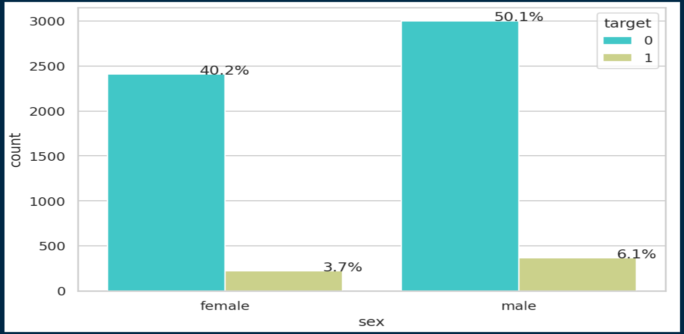
```
unknown                               27124
nevus                                  5193
melanoma                                584
seborrheic keratosis                    135
lentigo NOS                              44
lichenoid keratosis                      37
solar lentigo                             7
cafe-au-lait macule                       1
atypical melanocytic proliferation        1
```

EXPLORATORY DATA ANALYSIS

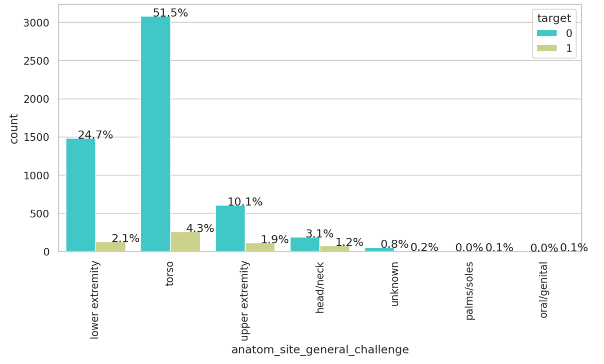
Target count Distribution



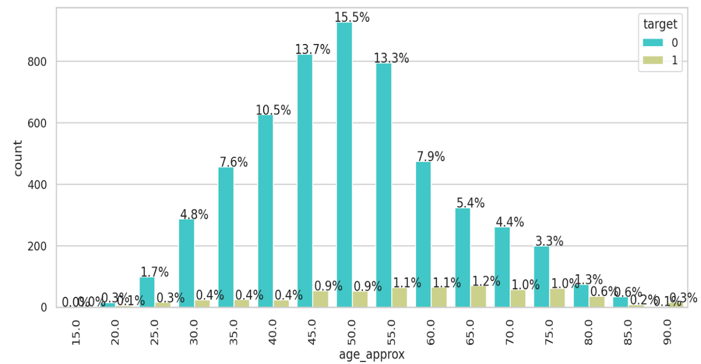
Target Distribution by Gender



Target Distribution by Image location



Target Distribution by Age

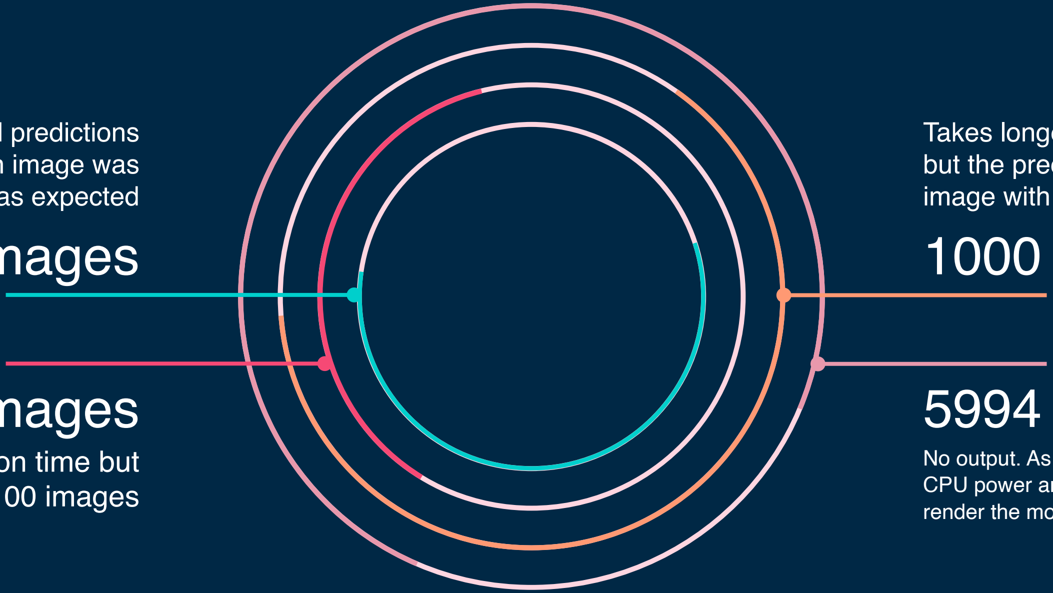


Modeling: SVC

Building the model using support vector classification

01

SVC MODEL RUNS



Output with good predictions accuracy but lesion image was not predicted as expected

100 images

500 images

More execution time but same result as 100 images

Takes longer time to render but the predicts the lesion image with 92.53% accuracy

1000 images

5994 images

No output. As the model uses all CPU power and is impossible to render the model

DISADVANTAGE OF USING SVC

SVC IS NOT SUITABLE FOR LARGE DATASETS

- The original SVM implementation is known to have a concrete theoretical foundation, but it is not suitable for classifying in large datasets for one straightforward reason — the complexity of the algorithm’s training is highly dependent on the size of the dataset.
- In other words, training time grows with the dataset to a point where it becomes infeasible to train and use due to compute constraints missing data or NA values.

SVC PERFORM POORLY IN IMBALANCED DATASETS

- The reason arises from the issue of an imbalanced support vector ratio, i.e. the ratio between the positive and negative support vectors becoming imbalanced and as a result, datapoints at the decision boundaries of the hyperplanes have a higher chance of being classified as negative.

Modeling: CNN

Building the model using convolutional neural network

02

MODEL COMPILING

OPTIMIZATION FUNCTION

- ADAM is the best stochastic gradient optimization algorithm that is used in deep learning applications.
- **Learning rate** for ADAM: 0.001 (default value).

LOSS FUNCTION

binary_cross_entropy

- Binary cross entropy used widely for binary classification
- It compares the predicted probabilities with the target variable and penalizes the misclassifications heavily.

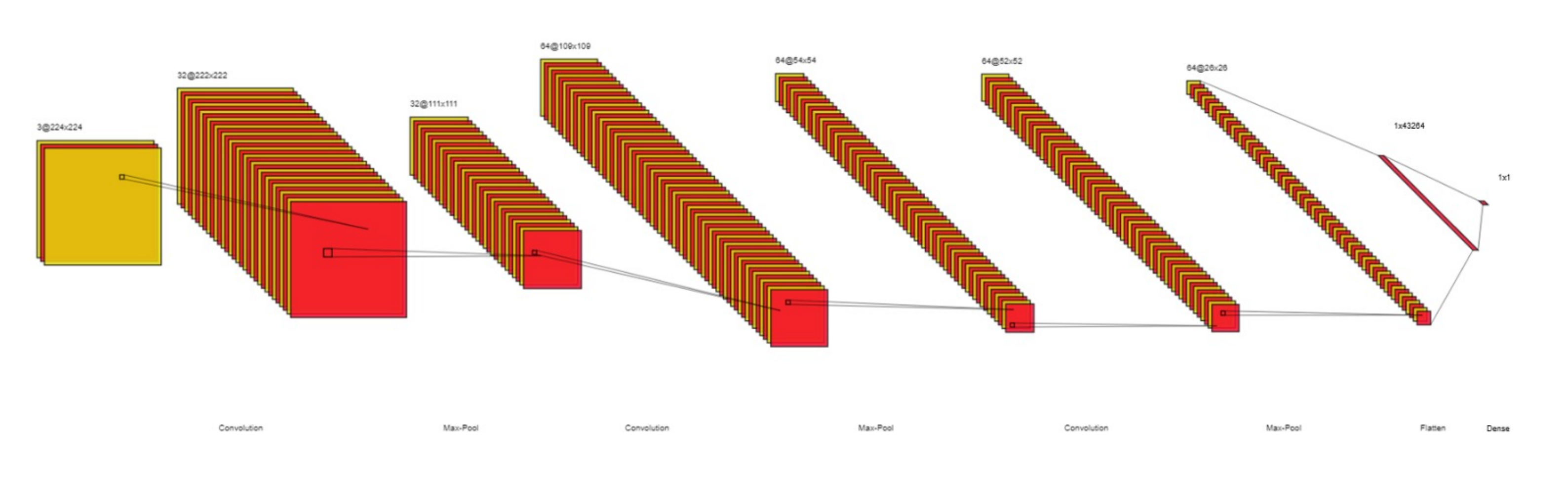
PERFORMANCE METRICS

AUC-ROC, Accuracy

- ROC curve plots TPR (True Positive Rate) against FPR (False Positive Rate). Ideally, in any binary classification problem, we need TPR to be 1 and FPR to be 0, hence AUC (Area Under the Curve) to be 1.
- AUC-ROC metric is well suited for highly imbalanced datasets.

CNN - BASE MODEL

- A simple CNN with 5 layers.
- Activation Function: ReLU for hidden layers and Sigmoid for the last layer.
- The accuracy while using this model was 92%
- Training Dataset: accuracy - 92.07%, AUC ROC - 0.9011
- Validation Dataset: accuracy - 92.99%,AUC ROC - 0.8855



Modeling:

Building the model using Transfer Learning and EfficientNetB4

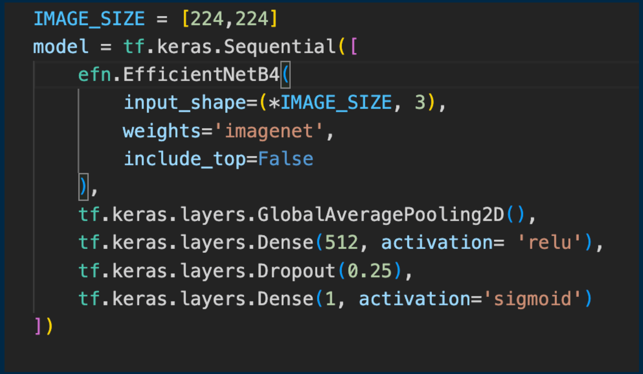
03

MODELING: TRANSFER LEARNING

- Transfer learning makes use of the pre-trained model weights for handling a new problem.
- We used Transfer Learning because of lack of resources and dataset is not large
- We used ImageNet (a widely used database, especially for image classification) weights with top_layer=false

MODELING: EFFICIENTNETB4

- EfficientNet model by Google is the best CNN in literature for image classification



```
IMAGE_SIZE = [224,224]
model = tf.keras.Sequential([
    efn.EfficientNetB4(
        input_shape=(*IMAGE_SIZE, 3),
        weights='imagenet',
        include_top=False
    ),
    tf.keras.layers.GlobalAveragePooling2D(),
    tf.keras.layers.Dense(512, activation= 'relu'),
    tf.keras.layers.Dropout(0.25),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
```

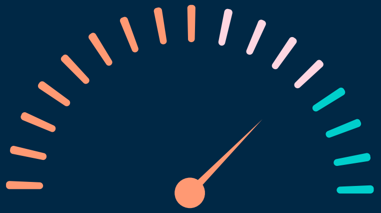
HOW TO DEAL WITH OVERFITTING?

1. dropout regularization: During model training, this approach randomly removes a number of neurons from a neural network. The performance of the model is unaffected by the lost neurons because their contribution is temporally erased.

2. Image Augmentation: Image augmentation artificially creates training images through different ways of processing or combination of multiple processing, such as random rotation, shifts, shear and flips, etc.,

TRAINING VALIDATION AND RESULTS

	Training	Validation
Loss (Binary cross entropy)	0.1471	0.1619
Accuracy	0.9420	0.9349
AUC	0.9545	0.9419



CONCLUSION

- In this project, we detected Melanoma using image classification and CNNs with an accuracy & AUC-ROC of around 95%.
- Stratified sampling and data augmentation is used to deal with imbalance in the datasets
- We could increase the performance by almost 2% with the help of these techniques

FURTHER CONSIDERATIONS: FUTURE WORK

We believe that the performance can further be improved by,

- Image preprocessing by manipulating various image artifacts.
- Other input features like Sex, Age and location of the disease is not considered for classification. We can train CNN with these features besides image data.
- Implementation using advanced CNN models like EfficientNetB7 on TPUs

FURTHER CONSIDERATIONS: PREVENT MELANOMA?

- Wearing Hats, goggles, long sleeve shirts (protective cloths), using sunscreen lotion and regular doctor/self checkups.

REFERENCES

1. https://seer.cancer.gov/statfacts/html/melan.html
2. https://www.wcrf.org/cancer-trends/skin-cancer-statistics/
3. https://arxiv.org/abs/1412.6980
4. https://challenge2020.isic-archive.com/
5. https://arxiv.org/abs/1905.11946