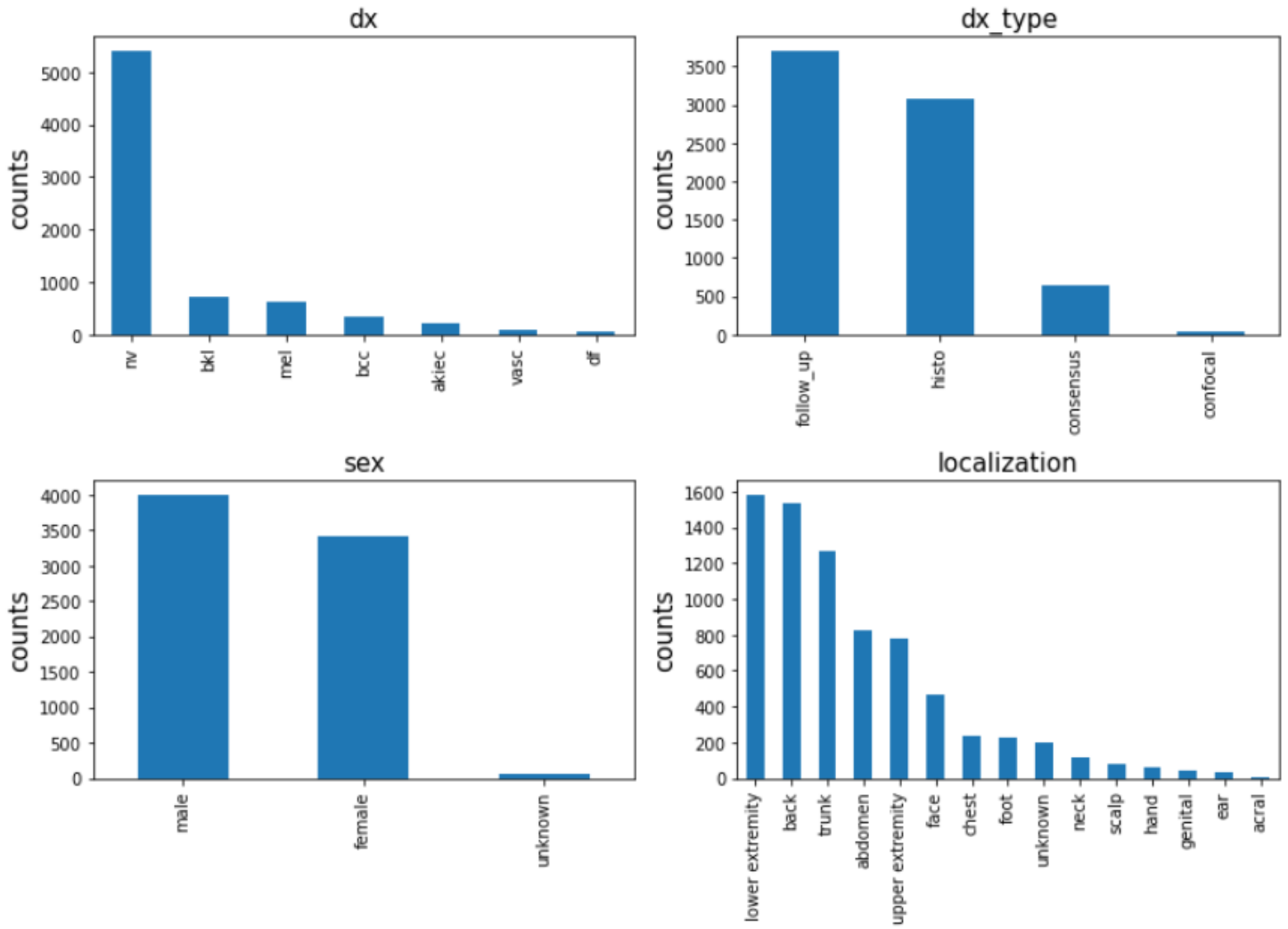
Module 5 final project

Zhiqiang Sun

- **Overview**
- The skin cancer dataset contains many medical images that show various kinds of skin cancer. In this project, we will analyze and visualize the relationship between cancer and age and the location of the body. Furthermore, we will use machine learning to train a model that can distinguish the cancer type by given images.
- **Data**
- The whole dataset were download from kaggle (https://www.kaggle.com/code/rakshitacharya/skin-cancer-data/data). The folder contains several csv files and two images folder. All the name of images were named with image id which can be found in the metadata excel file. I will try to train a model of 7 different skin cancer classes using Convolution Neural Network with Keras TensorFlow and then use it to predict the types of skin cancer with random images. Here is the plan of the project step by step:

- **Plan**

1. Import all the necessary libraries for this project
2. Make a dictionary of images and labels
3. Reading and processing the metadata
4. Process data cleaning
5. Exploring the data analysis
6. Train Test Split based on the data frame
7. Creat and transfer the images to the corresponding folders
8. Do image augmentation and generate extra images to the imbalanced skin types
9. Do data generator for training, validation, and test folders
10. Build the CNN model
11. Fitting the model
12. Model Evaluation
13. Visualize some random images with prediction



We checked the distribution of columns 'dx', 'dx_type', 'sex', 'localization' for different patients. The graphs show that:

1.In dx features, the 'nv': 'Melanocytic nevi' case take more than 70% of the total cases. The number suggests that this dataset is an unbalanced dataset.

2.In dx_type features, the histogram suggests most of the cancer were confirmed in Follow-up and histo Histopathologic diagnoses.

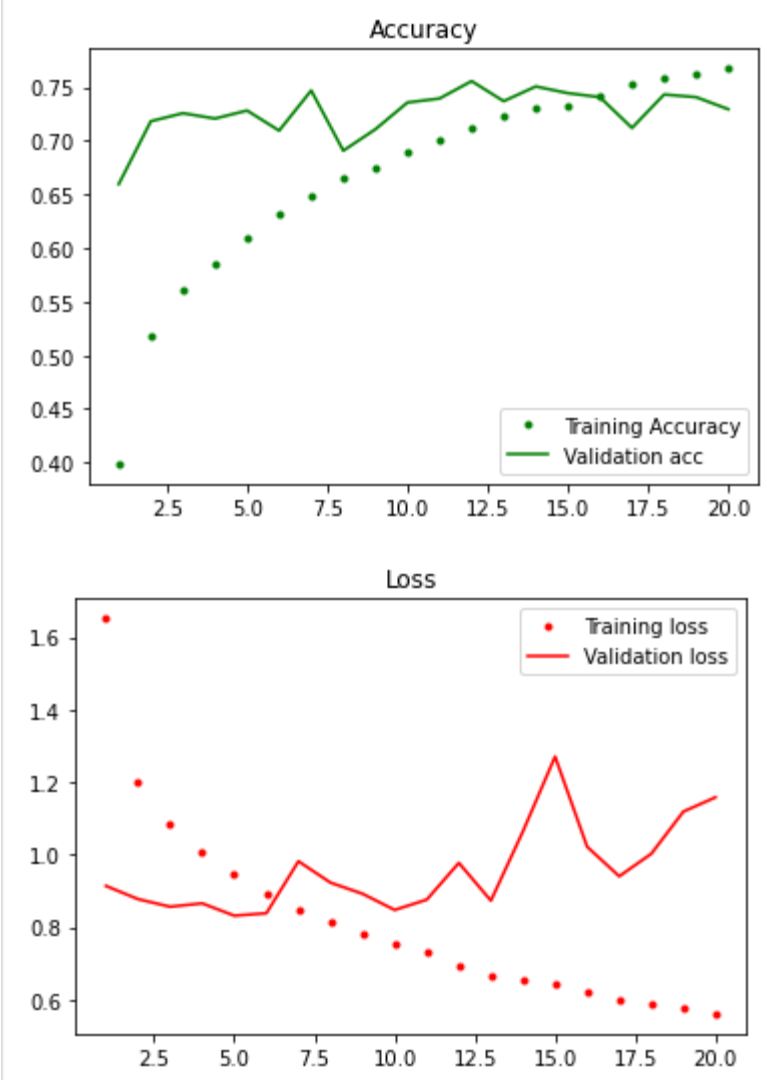
3.The sex feature shows that the amount of male who had skin cancer is slight larger than female but still similar to each other.

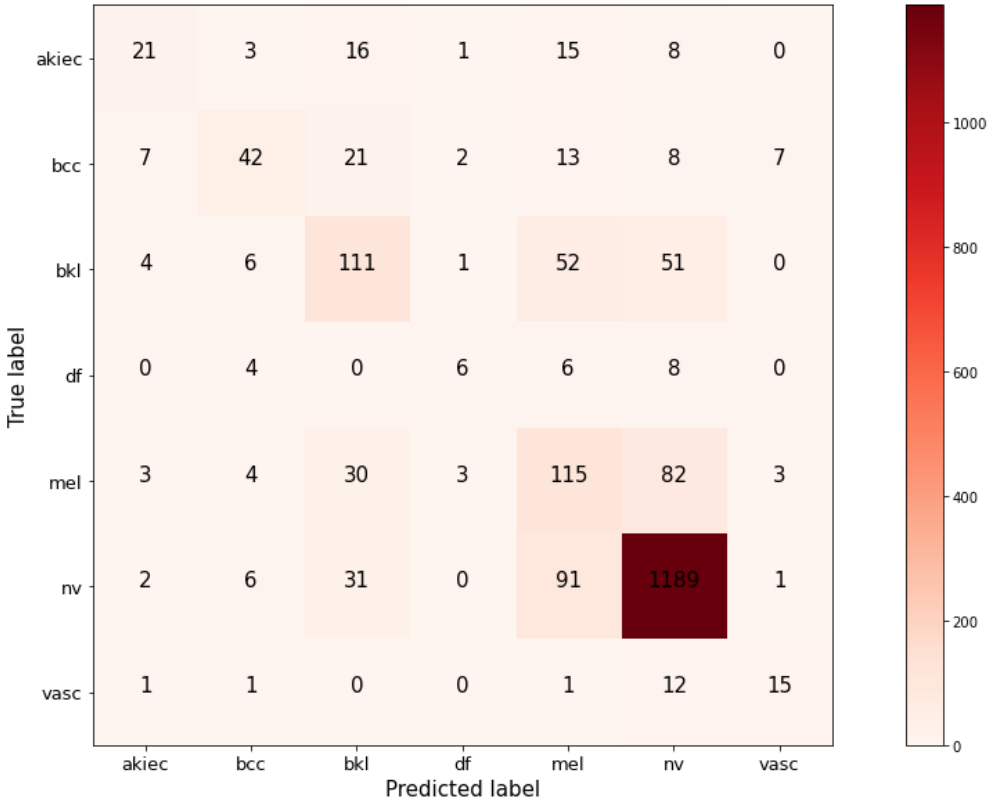
4.The localization analysis shows that lower extremity, back ,trunk abdomen and upper extremity are heavily compromised regions of skin cancer

Distribution of Ages on different skin cancer

In general, most cancers happen between 35 to 70. Age 45 is a high peak for patients to get a skin cancer. Some types of skin cancer (vasc, nv) happen to those below 20, and others occur most after 30.

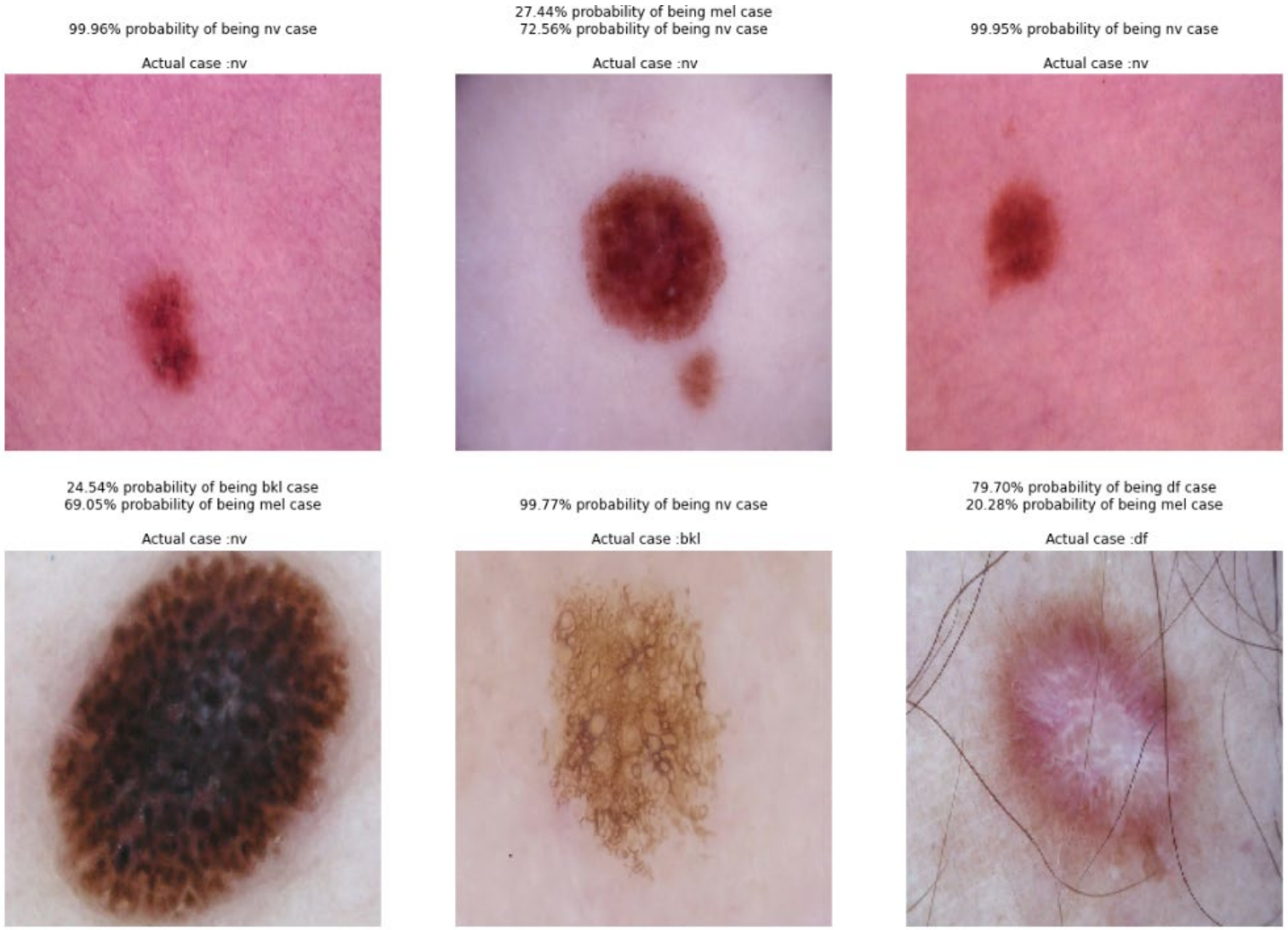
The results for the CNN model.





Accuracy is 74.84%

Several samples with predicted results



Conclusion

- We can extract the information about skin cancer from the metadata and explore the distribution of various features. For example, the most often age of skin cancer occur is around 45.
- We make one CNN model which can fit and predict the type of skin cancer well based on the images. The accuracy is 74.9% which is more efficient than detection with human eyes.

THANKS!

- Zhiqiang Sun
- sunzhiqiang04@gmail.com