Melanoma Cancer Program

1st Angela Gaudio
Graduate Computer Science Student
Stevens Institute of Technology
Hoboken, USA
anggaudio@gmail.com

2nd Priscilla Lin
Graduate Computer Science Student
Stevens Institute of Technology
Hoboken, USA
priscillalin1628@gmail.com

3rd Xi Yang
Graduate Computer Science Student
Stevens Institute of Technology
Hoboken, USA
cathy0902a@gmail.com

***Abstract*—Melanoma is the most invasive skin cancer with the highest risk of death. It is a disease in which cancer cells form in melanocytes, which color the cells in the skin. In our project, we will be developing a model that classifies melanoma via the lesion in the images to analyze the melanoma gene. For the current Machine Learning algorithms we plan to use Random Forest (RF) and Support Vector Machines (SVM). The second part of our project utilizes K best and X algorithms to train our model on the RNA sequencing of our data.**

I. Introduction

Melanoma is an uncommon form of skin cancer, however, it is one of the most dangerous if not found early on. Melanoma, unlike other skin cancers, forms in one of the deepest layers of our epidermis, the melanocytes. Melanoma cells are used to create melanin suggesting most early signs of cancer would be seen as brown or black, however, sometimes they do not produce melanin making them appear pink, tan, or white. Melanoma can appear anywhere on the human body such as hands, face, legs, feet, back, neck, etc. Thus our goal is to attempt to take an image given to us by the user and analyze it and have our program return what classification of melanoma skin cancer the user may have. From 1973 to 2000, the melanoma diagnosis rate increased. Even today, melanoma is still one of the top life-threatening cancers in the U.S, as seen in Figure 1. Melanoma can be found in most parts of the human body, and early diagnosis will increase the survival rate, as depicted in Figure 2.

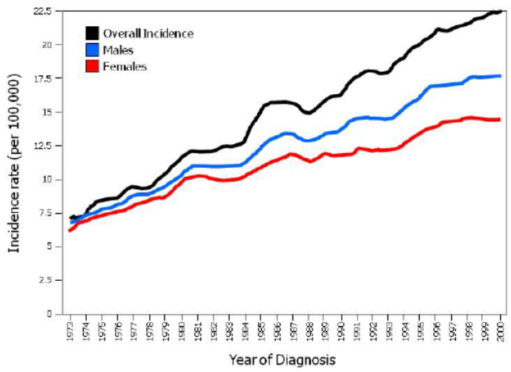


Fig. 1: 1973 to 2000 melanoma diagnosis rate in the U.S.
From: *Research Gate

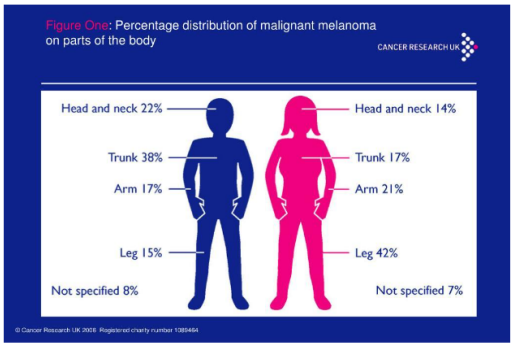


Fig. 2: Melanoma Distribution
From: *Slide Serve

For our data set, our team will be utilizing data extracted from *Kaggle's(1) data set for melanoma classification. This data set contains images for the various classifications of melanoma skin cancer, including both the training and testing sets. We will also be utilizing csv files that contain different patients' information in relation to the expression of melanoma based on images. To develop a model for the gene expression our team will apply the data sets from *Kaggle(2) as well as *HAGR. These data sets contain information on various patients with varying degrees of melanoma and what their RNA gene expression was.

Our team is utilizing the algorithms Random Forest (RF), Support Vector Machines (SVM), K-Mean Clustering, Principal Component Analysis (PCA), and Select K Best. RSF is used to obtain predictions when there are either complex or unknown factors present in a data set. Thus, we can utilize this algorithm to predict whether or not a patient's image is either benign or malignant in relation to melanoma. SVM in order to classify whether or not the image taken is likely to be melanoma cancer or just a normal mole. SVM can do this as it takes subsets of training data and linearly separates and classifies the data.

Based on what we have learned in class I believe our solution to predicting Skin Melanoma in patients is ideal. We were able to employ the techniques learned in this class in a real-life solution to a major problem that plagues our society. However, this course just scratches the surface of what Machine Learning can do, given an in-depth knowledge of the subject our team could have implemented a better solution and has many areas it could grow.

II. Related Work

Traditionally, researchers have used handcrafted features to describe skin lesions, such as color, texture, and shape-based features. These features are then fed into machine learning algorithms like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forest for classification. However, handcrafted features have limitations in capturing complex patterns in the data.

III. Our Solution

A. Description of Dataset

The data set we will apply for predicting the classification of melanoma will be the *Kaggle(1). This data contains a plethora of images for each classification of melanoma. It also contains both the training as well as the training set of data for our algorithms to employ. We also have the csv files that contain patients' information related to the expression of skin cancer based on images. The information provided from the images includes patient ID, sex, age, anatomical area of mole, and if it was benign/malignant.

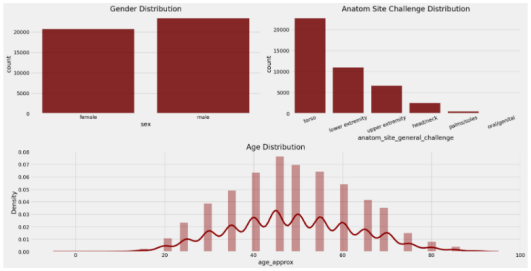


Fig. 3.1: Overview of Dataset from Kaggle(1)

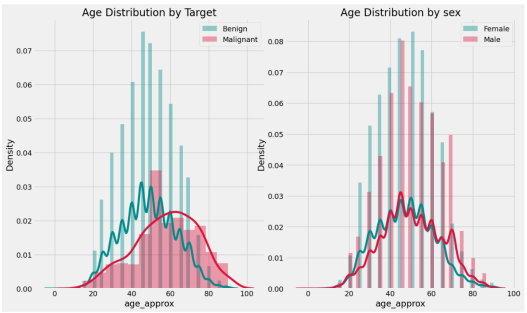


Fig. 3.2: Dataset from Kaggle(1) categorized by Age

The dataset of images from *Kaggle(3) aims to provide a resource for the development and evaluation of machine learning algorithms for the detection and classification of melanoma skin cancer. The dataset is divided into two main categories:

1) Benign lesions: These are non-cancerous skin growths that do not pose a risk to health. Examples include moles, seborrheic keratoses, and hemangiomas. In this dataset, there are 5,000 benign lesion images.

2) Malignant lesions: These are cancerous skin growths that can be life-threatening if not diagnosed and treated early. In this dataset, there are 5,000 malignant lesion images, specifically melanoma.

Analysis of gene expression in melanoma disease is complicated in terms of matching the RNA sequence data set with the human gene list. RNA sequence is widely used in gene expression analysis in biology due to its wide coverage area. However, we want to explore the relation to the actual human gene. Therefore, we will read both RNA gene expression and the human DNA gene list. The human DNA gene list will be the target gene list and find the target genes expression value from the RNA gene expression dataset and proceed with further analysis.

B. Machine Learning Algorithms

Prediction of Melanoma:

We implemented the analysis technique Random Forests to identify the levels of expression based on images that are associated with a melanoma patient's survival. From there we were able to train the model to give the training and test data sets provided. The model was able to pin how likely a patient was to have melanoma given the sex, age, and where on the patient the mole is.

The other technique used to analyze the Kaggle(1) data set was SVM. Worked ideally for this data set as it is able to classify the varying degrees of how likely to unlikely a mole is benign or malignant. Melanoma can be classified into 0-4 varying stages of cancer; 0 being the least dangerous to 4 which is the most dangerous. In our SVM we can have so that we classify the stages from 0-5 so that 0 will be benign and all the higher numbers can be signified as malignant.

For melanoma prediction, we use the Convolutional neural network(CNN) model to train the dataset. CNN is well-suited for image recognition and processing tasks. It automatically detects important features without any human supervision. A CNN convolves learned features with input data and uses 2D convolutional layers, making this architecture well-suited to processing 2D data, such as images. Automated feature extraction makes CNNs highly suited for and accurate for computer vision tasks such as object/image classification.

Analysis of Gene Expression in Melanoma:

In this portion, we will be using K-Means Clustering, Principal Component Analysis (PCA), selecting K Best, and finding frequent high expressions.

The K-Means clustering algorithm is used to group samples based on their expression profiles. We first use the elbow rule. This is a technique used to determine the appropriate number of clusters by plotting the explained variation as a function of the number of clusters and then selecting the elbow point of the curve as the number of clusters to use. Then select the optimal number of clusters and plot the K-means clustering to interpret the results of clustering.

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction, where high-dimensional datasets are transformed into lower-dimensional spaces by identifying the most important features or patterns in the data. PCA is commonly used in bioinformatics to analyze gene expression datasets.

The Select K-Best method is useful for reducing the dimensionality of the gene expression dataset and identifying

the most informative genes that are relevant to the phenotype of interest.

Find frequent high expression is a general comparison and counting technique used to identify the most frequent high-expression genes in the dataset. This technique is useful to identify the most significant genes for further disease studies.

C. Implementation Details

Prediction of Melanoma Implementation:

1) Random Forest (RF) Algorithm: Random Forest Algorithms are supervised machine learning algorithm that builds decision trees on various data points. By doing this the algorithm is then able to take the majority vote from this data for classification and average purposes. In the context of this project, we took this algorithm to classify the area of the mole, the age, sex, and whether it was benign or malignant. From this data, we were able to train our model on how likely or unlikely an image could be determined as a mole or possible melanoma. This can be seen in the figure below, Figure 4, in a confusion matrix format.

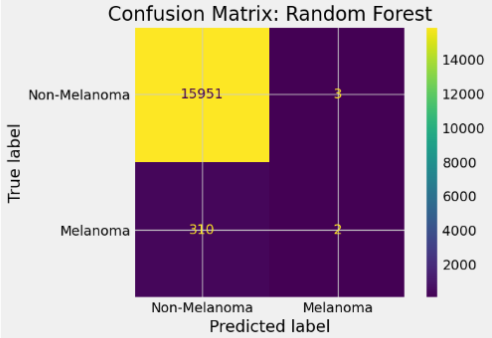
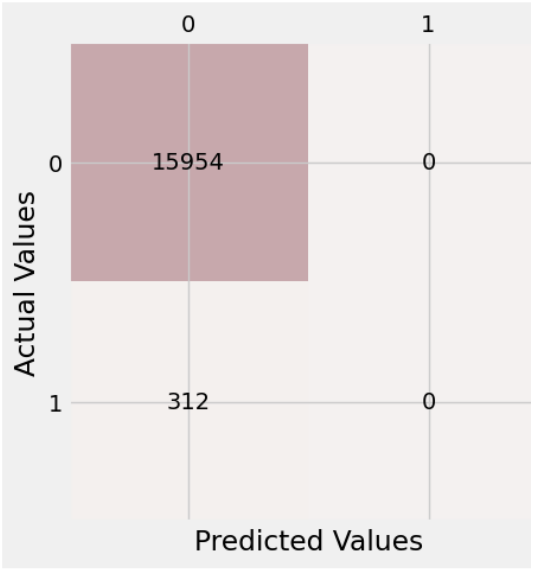


Fig. 4: Random Forest algorithm displayed in Confusion Matrix format

2) Support Vector Machine (SVM): Similar to (RF) algorithms, Support Vector Machines (SVM) are supervised machine learning algorithms. The SVM training algorithm is able to build a model that is able to assign new data points to different categories, making it a non-probabilistic binary linear classifier. For our project, this algorithm becomes very enticing to use as it will be able to classify our data into malignant vs benign categories. This will be helpful if we are able to connect the image analysis to this as our machine will better be able to determine how likely a patient's mole is to be melanoma or benign.



	precision	recall	f1-score	support
0	0.98	1.00	0.99	15954
1	0.00	0.00	0.00	312
accuracy			0.98	16266
macro avg	0.49	0.50	0.50	16266
weighted avg	0.96	0.98	0.97	16266

Fig. 5

3) Convolutional Neural Network (CNN): For melanoma prediction, we build a CNN model with 1 layer of 2D convolutional layers to get the input and process the data. Then, 1 layer of dropout to prevent the data from overfitting, and 1 layer of max pooling to reduce the dimensions of images and do the convolutional layer, dropout, and max pooling layer again. After this, a flatten function is used to flatten the data into a 1D array. Then 2 fully connected layers are added. A softmax function normalizes output to produce probabilities more suitable for multiclass classification tasks. Last, we compile the model by specifying the optimizer, loss function, and metric, then, summarize it.

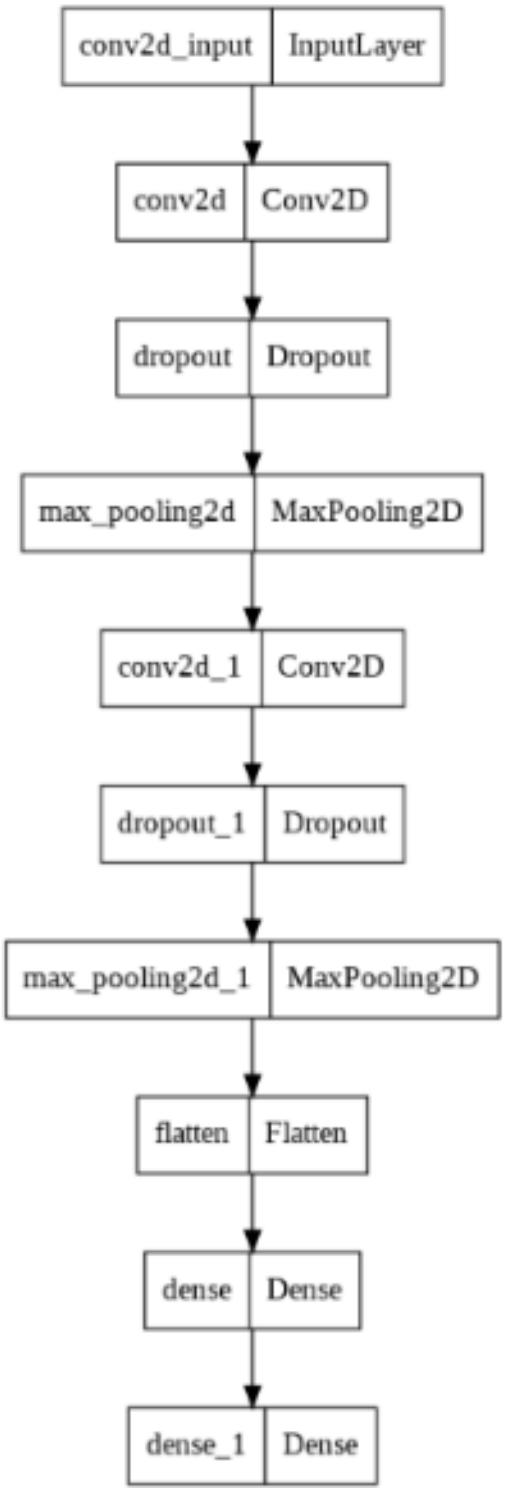


Fig. 6: CNN Model

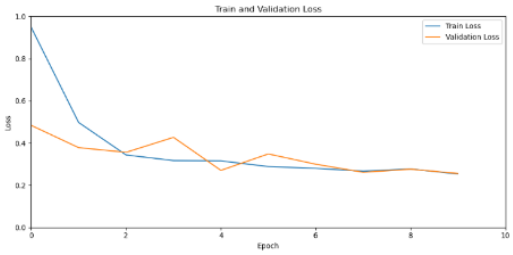


Fig. 7.1: train loss, validation loss vs epoch times

Fig. 7.2: train accuracy, validation accuracy vs epoch times

Gene Expression Implementation:

4) K-Means Clustering Algorithm: In the K-means clustering algorithm, we first perform the elbow method to determine the number of clustering, and the graph is shown in Figure 8.1. K-means clustering is using the K-Means function from scikit-learn and finding the number 3 is most optimal. Finally, we plot the clustering results which are shown in Figure 8.2. In the plot, each point represents a sample in the gene expression data, and the color of the points indicates its cluster assignment. Due to the high sample volume, the clustering is poor and needs to have a downsized sampling.

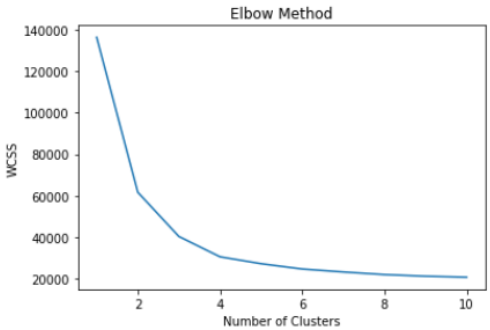


Fig. 8.1: Elbow curve method to determine an optimal number of clusters.

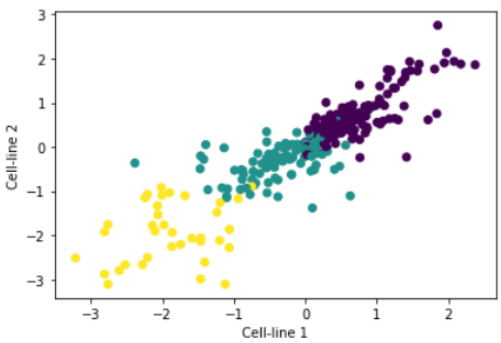


Fig. 8.2: Randomly selected cell lines to visualize the clustering of genes

5) Principal Component Analysis (PCA): Principal component analysis (PCA) is performed on a normalized gene expression dataset. PCA analysis is performed using the "PCA" function from sklearn.decomposition module. The components are set to 3 which indicates that the PCA transformation should produce three principal components. Then the "fit" method is called to calculate the principal components and store them in the object. The "transform" method is called to apply the principal components to the original dataset to produce a transformed dataset. Finally, the transformed data is plotted, and the x-axis and y-axis represent the first two principles components that are shown in Figure 9.

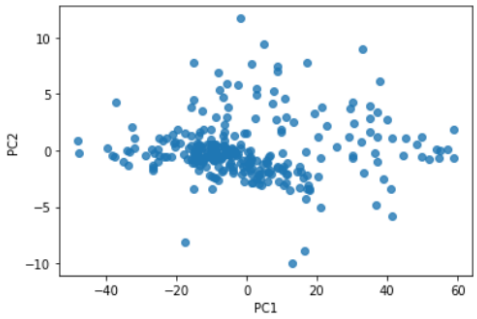


Fig. 9: Principal component analysis for gene expression measurements.

6) Select K Best: For this algorithm, we transposed the original dataset before analysis. The original dataset has the first rows for cell lines ID and the first column for gene list. We transposed the dataset so we can focus on the targeted genes. Then split the dataset into input features and target variables and selected the k-best values to find the list of the genes. The initial K-best value was selected at 50 but it did not produce the optimally selected list, so we increase it to 100 and it is the most optimal value for K-best.

```
# Split into input features (X) and target variable (y)
X = transposed_df.iloc[:, 1:]
y = transposed_df.index

k_best = SelectKBest(score_func=f_classif, k=100)
k_best.fit(X, y)

selected_indices = k_best.get_support(indices=True)

selected_genes = transposed_df.columns[selected_indices]

selected_df = transposed_df.iloc[:, selected_indices]
```

```
/opt/anaconda3/lib/python3.7/site-packages/sklearn/feature_selection/_univariate_selection.py:110: RuntimeWarning: invalid value encountered in true_divide
  msw = sswn / float(dfwn)
```

```
selected_list = list(selected_df.columns)
print(selected_list)
```

```
['BLM', 'LEPR', 'INSR', 'PDGFRA', 'PTK2B', 'BUB3', 'HRAS', 'UCHL1', 'SOD1', 'TERT', 'AKT1', 'MYC', 'HESX1', 'S100B',
'EFEMP1', 'TXN', 'WRN', 'SERPINE1', 'POLG', 'POLB', 'C1QA', 'CLOCK', 'PSEN1', 'XRCC6', 'ERCC8', 'ERCC1', 'ERCC2', 'ER
CC3', 'ERCC4', 'ERCC5', 'ERCC6', 'UBE2I', 'NGFB', 'PARP1', 'PCMT1', 'VCP', 'ABL1', 'HSP90AA1', 'TOP1', 'HSN', 'HDAC
2', 'MED1', 'MLH1', 'PRDX1', 'APOE', 'SIRT1', 'PPARGC1A', 'SIRT7', 'POLA1', 'FEN1', 'TPP2', 'JUND', 'PTPN1', 'JAK2',
'PPARG', 'PPARA', 'PRKDC', 'CCNA2', 'FLT1', 'CLU', 'MAPK5', 'TP53', 'TOP2B', 'ELN', 'LEP', 'HSPA1A', 'MXI1', 'HSPA
9', 'VEGFA', 'HTRA2', 'TP73', 'TNP', 'EGFR', 'PMCH', 'RB1', 'TAF1', 'GSR', 'BDNF', 'SOD2', 'BSCL2', 'GCLM', 'GCLC',
'ARNTL', 'GPX1', 'GPX4', 'HIC1', 'PLCG2', 'NFE2L2', 'NFE2L1', 'HDAC3', 'KL', 'RAD52', 'RAD51', 'CEBPB', 'CEBPA', 'XRC
C5', 'TFDP1', 'IKBKB', 'AGTR1', 'PDGFB']
```

Fig. 10

7) Find Frequent High Expression: This algorithm uses basic comparison and counting techniques. The high expression value is set to be 8.0 because the dataset is range from 0 to 18, and 8.0 is the most optimal value. Then we count the expression values greater than 8.0 for each gene. Since we have 473 cell lines, we want to find out the most frequent genes with high expression values, and we set the count to 470. Finally, the list of the most frequent genes is generated for further studies of disease.

```
# Filter the data to get only the high expression values
high_expression = transposed_df[transposed_df > 8.0]

# Count the number of times each gene has a high expression value
gene_counts = high_expression.count()

# Get the list of genes with frequent high expression
frequent_genes = gene_counts[gene_counts > 470].index.tolist()

print(frequent_genes)
```

```
['SP1', 'CREBBP', 'PTPN11', 'RELA', 'GRN', 'SUN1', 'SIN3A', 'BAX', 'PTK2', 'LMNA', 'PCNA', 'HDAC1', 'RPA1', 'STAT3',
'APP', 'EEF2', 'GSS', 'HRP1', 'TRAP1', 'FEN5', 'ATP5O', 'PTT', 'MTOR', 'HSF1', 'MDM2', 'CSNK1E', 'PRKCD', 'MAX', 'EEF
1A1', 'HIF1A', 'PPP1CA', 'ARRGAP1', 'GRB2', 'MAPK14', 'SUND1', 'HSPD1', 'END', 'POPK1', 'CAT', 'JUN', 'A2M', 'GSK3A',
'UBB', 'CTNNB1', 'RAE1', 'EP300', 'APEX1', 'YWHAZ', 'PML', 'HSPA8', 'DGAT1', 'CDC42', 'EMPSTE24', 'FOXO3', 'SDHC', 'G
SK3B', 'AIFM1', 'NCOR2', 'NCOR1', 'NFKB1', 'NFKB2', 'GSTP1', 'IGF1R', 'STUB1', 'HMGB2', 'HMGB1', 'SNC1', 'FOS', 'SQST
M1', 'MAPK1', 'STAT5B', 'TCF3', 'BUB3', 'SOD1', 'AKT1', 'TXN', 'POLG', 'PSEN1', 'XRCC6', 'ERCC1', 'ERCC3', 'ERCC5',
'UBE2I', 'PARP1', 'PCMT1', 'VCP', 'ABL1', 'HSP90AA1', 'TOP1', 'HDAC2', 'MED1', 'MLH1', 'PRDX1', 'APOE', 'FEN1', 'JUN
D', 'PTPN1', 'PRKDC', 'TOP2B', 'HSPA1A', 'MXI1', 'HSPA9', 'HTRA2', 'SOD2', 'BSCL2', 'GPX1', 'GPX4', 'NFE2L2', 'NFE2L
1', 'XRCC5', 'TFDP1', 'IKBKB']
```

Fig. 11

D. Comparison

Prediction of Melanoma: In this section, we will compare the RF and SVM methods in which we created models to predict whether or not a patient was likely to have melanoma. RF and SVM are equal in the sense that they both use classification in order to sort the data as well as neither does well with larger datasets. Random Forest generates an uncorrelated decision tree by choosing random sets of features in order to build a tree. It is also no greatly influenced by outliers within the data set. While looking at SVM we found that it works best when there is an obvious margin of separation between each classification level. As well as it is more effective in cases where there is a great depth in dimensions rather than the number of samples given. Also unlike Random Forest SVM is a relatively memory-efficient algorithm.

IV. Future Directions

Had we been given more time we may have been better able to develop an actual app that could take images and analyze the images and categorize the information. That could then be fed to our model to better train it to predict how likely a patient's mole was to be benign or malignant of melanoma. We also could have utilized possibly more complex algorithms and models that would better suit our data set to give a more conclusive answer.

Ideally, we also would have been able to connect the data from the Prediction of Melanoma section with the Gene

Expression of Melanoma. This would have allowed the application if it believed a mole to be malignant to send information to a doctor to inform the doctor of what to expect in the RNA expression. This could lead to a more medical analysis of how RNA gene expression can change or vary between different stages of melanoma.

V. Conclusion

Given the current knowledge and expertise of our group, this implementation has exceeded our expectations. There may be other algorithms that would have been better utilized as CNN that are new to our team as a whole. This makes them harder to utilize in our work and it may be smart to switch them out with something we are better familiar with. SVM may need to be switched out as while it's highly effective it isn't suited for large data sets due to the training time needed for this algorithm. However, beyond that, our implementation gave us multiple looks at how to analyze Skin Melanoma in a variety of patients via the physical and RNA views of melanoma cancer. We were also able to compare different algorithms for each point to test which would have the best aptitude for our purposes of predicting melanoma in a patient as well as the gene expression of the disease.

References

[1] "Siim-ISIC Melanoma Classification." Kaggle, https://www.kaggle.com/competitions/siim-isic-melanoma-classification/data.

[2] Miñoza, Jose Marie Antonio. "Melanoma Gene Expressions." Kaggle, 8 Dec. 2020, https://www.kaggle.com/datasets/jomaminoza/melanoma-gene-expressions.

[3] Javid, Muhammad Hasnain. "Melanoma Skin Cancer Dataset of 10000 Images." Kaggle, 29 Mar. 2022, https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images.

[4] "Human Ageing Genomic Resources." Human Ageing Genomic Resources (Senescence.info), https://genomics.senescence.info/download.html.

[5] Amos-Duran. "Numbers and Rates of New Cases for Malignant Melanoma, UK by Country, 2006." SlideServe, 30 Nov. 2014, https://www.slideserve.com/amos-duran/table-one-numbers-and-rates-of-new-cases-for-malignant-melanoma-uk-by-country-2006.

[6] Skin Pigmentation and Melanoma Risk - Researchgate. https://www.researchgate.net/profile/John-Dorazio/publication/221916871$_Skin_Pigmentation_and_Melanoma_Risk/links/541ffb2e0cf241a65a1aeef8/Skin-Pigmentation-and-Melanoma-Risk.pdf$.