**Discovering Gene Sets for Melanoma Tumor Malignancy**

**(Group 5)**

**Kaustubh Prabhu & Team**

**Abstract**—Melanoma is a under researched problem considering how commonly it happens. Detrmining malignancy is a key thing that is studied in this topic, we do the same. While trying to determine malignant versus non malignant cells the gene expression data is used. The gene expression data along with some anlysis tools like machine learning, feature extraction and GSEA, gave insights on the gene sets that were most distinguishing between malignant and non malignant

**Index Terms**—Melanoma, Malignant, Gene sets.

—————————— ◆ ——————————

# 1 Overview

ThIS project uses the **GSE72056** data, which is a data containing single-cell expression profiles of metastatic melanoma. These samples were from 19 individuals and had over 4500 samples of which around 1200 were found to be malignant cells the other cell types were T-cells, B-cells and others. The data was originally from a paper studying metastatic melanoma and it gave a list of 229 probable genes that were distinguishing factors in malignant versus non malignant cells. Hence, we as a group decided to do something similar. We using different feature extraction methods tried to get our set of 229 genes that would create a high accuracy then compare the results. We hoped that our genesets will have a significant intersection with their 229, if we get those results then we experimentally validate those results, but if we do not then we actually find new genes that are under explored or omitted by the paper.

As a team we worked on five different feature extraction methods and the selected optimal parameters from SVM for each of the individual methods. The 4 teammates used 4 methods and got sets of 229 genes, we took a union of these sets. The union gave 637 different genes, these were again passed through another feature extraction method which gave our team’s final 229 genes. The output of every feature extraction was used as an input to the SVM which was trained to classify the data into 2 groups malignant versus non malignant. This classification had an gave an accuracy of around 99.5%.

Later we studied and explored our results, this included getting information on the genes in our final subset, analyzing the most common genes in the area, looking at the genes that were found in all the 5 feature extraction methods. Then validation of our results by comparing our genes to the base paper’s genes, running a GSEA to find if any of our known genes are enriched in the known melanoma geneset. We found some interesting results these will be discussed in further sections.

# 2 Background

## 2.1 General Concept

Melanoma is a very commonly occouring cancer type and it is known to be curable and treatable, the approaches such as immuno therapy are specialy applicable to this particular cancer because it is caused by the mutations in cells triggered by the environment such as the ultraviolet light ray. And the absence of a heridatory cause helps reduce the complexity and gets us in a better position to study them.

Bulk genomic analyses and the expression profiling on the clinical specimens have created much of our understanding of cancer. However, it is known or studied that human tumors are an intricat ecosystems composed of diverse cells, including malignant, stromal subsets, and immune, whose, precise characterstics are often masked by the bulk methods. Hence, the single-cell expression techniques have been emerged as significantly powerful approach to dissect human tumors at the resolution of individual cells, providing a compelling approach to deciphering cancer biology.

This single cell data has expression of genes mapped from human genome and can add accountability to all different feartues of dufferent cell types in tumor.

Traditionally the cells are studied, looking at their neuclei, rarely or only for research purposes the expression levels are studied and the malignancy for the samples is tested.

## 2.2 The problems in the domain

In humans there exist any where from 20 to 25 thousand genes. These genes have certain effects on diffirent functions throught the life of cell and the life of body. To study any genomics problem, we need to study all the genes currently known to find any relation ship between diseased cells and normal cells. The sheer number of genes and the apparent lack of samples makes it difficult to study them individually.

Hence there exists a whole domain of genomic studies which works on getting important relationships between the genes and diseases using various bio inspired statistical tools. The problems that lie here are basically because of misinterpretation of data, misrepresentation of data and importantly under representation of the data.

Here, the data has a lot of noise because of multiple factors that are liked to extraction of genome and sequence reading. The expression data is very noisy. The more difficult part arises when the results are to be interpreted. In the general field of biology correlation does not have high weightage. Even if we find co existing patterns it is very difficult to understand and explain the relation. Thus, a causal relation is needed to validate any significant hypothesis or theory in ths domain.

## 2.3 Importance of our work

Melanoma is a high occourance cancer, but is treatable. People study melanoma to find cure using techniques of immunotherapy. The one important thing in this context is the target genes. Since the cure to melanoma is targeting particular cancer genes and then later assessing how to block or alter the effects of mutation on that gene and find ways to mask the effect of that gene by down regulating an upregulated gene or by upregulating a downregulated gene. Here if you see the key word or concept along which the whole research lies is particular genes. Many different types of gene can cause many types of melanomas and hence every sub type will have a different cure. Now what our work does is that, it finds probable genes that are the distinguishing factors in malignancy or non malignancy. Then these genes can be further explored by the biologists to determine the whether there is any causal relation in the genes and if there exists a causal relation can it be corrected to correct the malignant melanoma.

## 2.4 Existing Approaches

Researchers have been using multiple feature extraction techniques to extract featurese from the expression data. Some of them include ranking methods, Unconditional Mixture Modelling, which is a univarriatye method that assumes two different states of the gene on and off that is upregulated or down regulated and checks if the change in binary state of the gene affects the classification. It, checks using mixture overlap probability. Markov Blanket Filtering, which is a multi varriate method, finds features that seem to be independent of the class label and removes them as they will not affect the accuracy of the classification. Others are the correlation approaches such as, Error-Weighted Uncorrelated Shrunken Centroid (EWUSC), this is a multivariate method, this method is based on the uncorrelated shrunken centroid (USC) and shrunken centroid (SC). The shrunken centroid is calculated by taking the average gene expression for each gene and dividing it, in every class by the standard deviation for that gene in that particular class. In this manner higher preference is given to genes whose expression is the same among different samples in the particular class. Correlation-based feature selection (CFS) as stated by Hall [17] follows the principal that “a good feature subset is one that contains features highly correlated with the class yet uncorrelated with each other.” The above were the filter methods. There also exist wrapper methods. They can be deterministic or randomized. Gradient-based-leave-one-out gene selection (GLGS) was originally introduced for selecting parameters for the SVMs. The method starts by applying PCA on the dataset. A lower dimension vector is obtained using gradient based methods. The scaling factors for the original set of genes are calculated. Then the genes are selected based on their correlation factor. Leave-one-out calculation sequential forward selection, is a very widely used feature selection method for cancer data based on sequential forward selection (SFS). Then lastly the randomized wrappers they use genetic algorithms.

## 2.5 Possible Alternative Solutions

The above discussed methods are good but they are not all encompassing, similar to these methods we found a few other methods such as Univariate Feature Selection, Feature Importance, Pearson, Correlation, Recursive Feature Elimination, Lasso Selection. These methods are known to perform well in other fields of data so it was an obvious choice to work on these methods aince we wanted some novelty. Along with the abovementioned methods. We use the GSEA tool to validate our results against a published upregulated melanoma Geneset. On top of all the feature extyraction methods we use the SVM classifier which is the most commonly used classifier in this particular domain. Now one interesting thing we did was no one had clubbed feature extraction methods. We took the output of the first 4 feature extraction methods as an input to the 5th method which was the Lasso. We selected lasso as the final featureselection method because historically people have used it in this domain and have found decent results.

The above mentioned methods are defined here ine brief.

1. Univariate Feature Selection:

It picks top k features using chi squared test. Commonly used, and appropriate for a classification model.



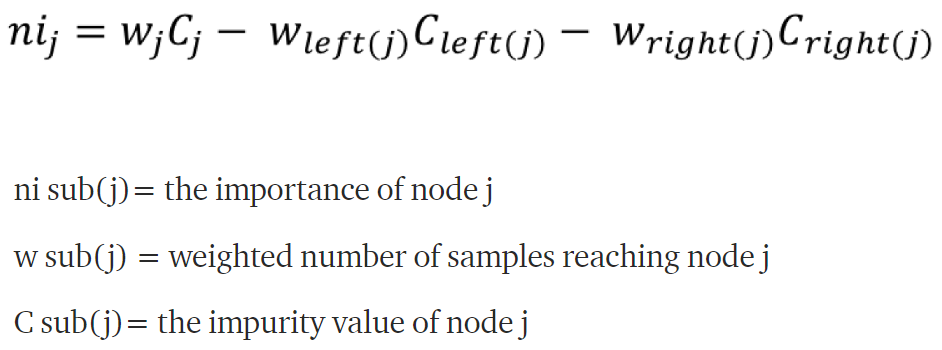
Oi is the observed result, and Ei is the expected result

Compares each feature to the label to measure correlation

1. Feature Importance:

This method fits an ExtraTree Classifier using gini index,

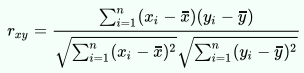
Picks top k features based on feature importance, fi



1. Pearson, Correlation:

A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related. The Pearson correlation is also known as the “product moment correlation coefficient” (PMCC) or simply “correlation”.

Calculates the features using



Where, n is sample size, xi, yi are the individual sample points indexed with I, xbar = 1/n sum(xi) (the sample mean); and analogously for yi

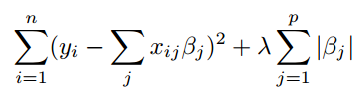
1. Recursive Feature Elimination:

Recursive feature elimination (RFE) is a feature selection method which fits a model and then eliminates the weakest features until the previously specified number of features is reached. Features are ranked by the model’s coeffitients or feature importance attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

1. Lasso Selection:

Lasso regression is a linear regression that uses shrinkage. Shrinkage is the function where data values are shrunk to a central point, just like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi co lineartiy or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

It uses a L1 norm for penalty.



Which is the same as minimizing the sum of square with constraints, Σ |Bj|< s. Some of the βs are shrunk to exactly zero, resulting in a regression model that’s easier to interpret.

A **tuning parameter**, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

* When λ = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.
* As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when λ = ∞, all coefficients are eliminated).
* As λ increases, bias increases.
* As λ decreases, variance increases.

If an intercept is included in the model, it is usually left unchanged.

1. GSEA:

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes). The Gene Set Enrichment Analysis PNAS paper fully describes the algorithm. The GSEA software makes it easy to run the analysis and review the results, allowing you to focus on interpreting the analysis result.

# 3 Our Approaches

## 3.1 Original Project Approaches

Our initial approach was pretty well defined, we were going to do feature selection based on the lables to determine if a tumor is malignant or benign and even take different genes into consideration, look at multiple classification techniques and then compare each group member’s selected features and results against each other. Here three of us were working on this data and had 3 different list of genes that we were working on, one was working on all genes, one was working on 229 genes, one was working on all but the 229 genes. Along with it we were also exploring to determine which cells come from, as there was a variation of location from where the samples were taken from, which patient which was being handled by the 4th person. Then we were going to analyze new insights from what we have observed and compare to the known genes which are frequently expressed in tumors, these frequently expressed genes were either going to be from the GSEA or the 229 from paper or some other online sources, Gene set enrichment analysis was going to be explored to validate our results and further get more insights o what our results imply.

## 3.2 Changes to the Project Approach

There were a few major changes in our final approach that we ended up with. They were as follows:

1. We changed from working on different methods of classification to just one method.
2. We reduced the importance of the role of GSEA analysis results
3. We dropped the work on exploring the mapping in samples and patients/location of extraction.
4. We also changed the approach on how we select the features.
5. Now one of us found a list of cancer genes and decided to work only on those 2305 genes. While others looked at all the genes.
6. We also changed the way we normalize the data and also considered working on removal of housekeeping genes.

## 3.3 Reason for Changes

As discussed before there were a few changes that were made along the way. There were many reasons for them, the first was when we initially ran our data for all diffiernt methods we found SVM was performing much better than most other algorithms, such as Random forest and MLP, Naieve bayes and logistic regression, and thus we dropped other methods from the initial study when doing individual feature extraction and testings. While the final analysis of the combined geneset we did use all the algorithms. This was done to reduce the time spent un necessarily on grid searches of all these methiods.

Secondly, reducing the role of GSEA was a decision we took due to the lack of time that we had. The GSEA tool did not work on my machine and hence had to use a python version of it. Setting it up for our data was a mammoth task as no one had ever used this tool before, never had heard of it and neither had any bioigical understanding of how the results are to be interpreted. So, we just used GSEA to validate the results not at all as features. The GSEA was used to check enrichment of the published melanoma geneset for our final 229 features.

Third, the idea of mapping sample to patient/location was pioneered by Dominic, and after the mid term presentations it was dropped as it was not giving good accuracy and was confused on the comments during presentation.

Fourth, since Dominic declined to work on the earlier problem he had to be added into out our distribution of genes, again taking advice from the professor we decided that every one should be working on all the genes because different methods and different data should not be compared. Soon we found out a list of 2305 genes wjich were known to be found predominantly in cancer data, thus we decided one of us can use those genes instead of a complete gene list.

Fifth, Interesting thing we learned during mid term presentations was that housekeeping genes have little to no effect on cancer and are always highly expressed. Thus we decided to remove them completely. Along with that we also changed our normalization and filtering, by using max norm, and filtering all who were only zero expressed.

## 3.4 Our Implementation

We implement the project in a very straightforward way. First we take the data, from the NCBI website **GSE72056,** this data had 4105 samples from 19 individuals of which 1200 were malignant, 3200 were non malignant and the rest were not certian or unclear. We first of all remove those uncertain samples. Then in the second step we drop all the housekeeping genes as it was decided that those genes are always highly expressed and the literature on them suggests that they generally do not play any role in cancer. For the third step we delete all the genes which have a expression value of 0 in atleast 85% samples. Then we normalize the data over the genes. This is done by dividing the expression value by the range of expression values, more like min-max normalization.

This normalized data is shared amomg everyone except me as I am using just 2305 cancer genes published in 2016. So I do not have to remove the housekeeping genes, but filter out the un expressed genes and also normalize them.

After this pre processing we visualize the data using PCA and T-SNE, the visualization helped us decide wether we should use a gene wide normalization or a samplewise.

This is the part we separate and work on our own feature extraction methods, we use the first 4 feature extraction methods discussed above, with the aim to reduce the number of features from whatever to just 229, we run the feature extraction methods. I chose to work with recursive feature elimination method did loops of ten less features every iteration.

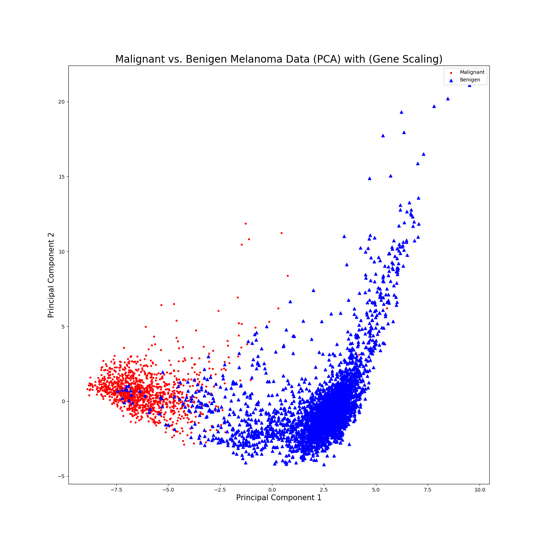
The output of these feature extractions gave us our individual genests, these genesets were used to train an SVM for all the 4 team mates separately. While training the SVM we take out samples from 4 patients so that none of the erpression levels are seen by the machine and it acts as a true validation set. Every one ran a grid search for the svm and found the best parameters for the SVM. We then compared everyones accuracy, it was similar, so then we took all the 4 genesets and looked at the intersection and union and we ended up with 679 genes in total. These 679 were again passed through another feateure exteaction method known as LASSO, this method also gave us 229 genes, this geneset was passed to all the 6 classification models SVM, KNN, Naive Bayes, Decision Tree, Random Forest, & MLP. We check for the test train accuracy and other evaluation metrics.

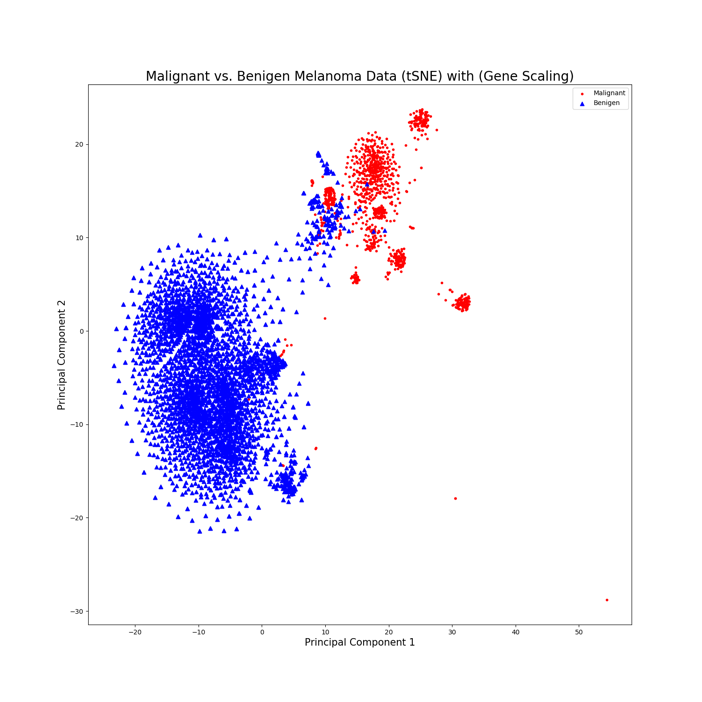
This final geneset was then passed to GSEA to check it’s enrichment profile against a known melanoma geneset released by the GSEA platform.

This gave us genes and a lot of other information. This output was further analysed and some information on this was recorded.

# 4 Accomplishments, Results and Visualisations

1. PCA and T-SNE for our data.

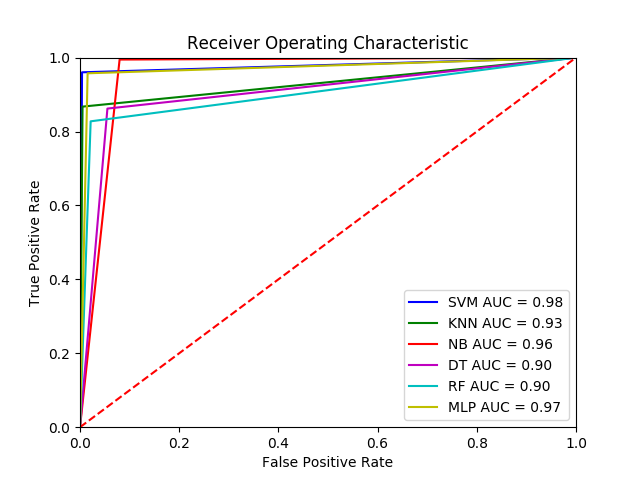




The PCA and T-SNE after preprocessing.

2. The results of various models on the original 229 Genes. Followed by ROC curves.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Acr** | **F1** | **Prisn** | **Rcl** | **Speci** |
| **SVM** | .98597 | .97443 | .98907 | .96021 | .9959 |
| **KNN** | .95938 | .92243 | .98494 | .86737 | .99488 |
| **NB** | .94092 | .90361 | .82781 | .99469 | .92016 |
| **D Tree** | .92171 | .85979 | .85752 | .86207 | .94473 |
| **R F** | .93648 | .87887 | .93694 | .82759 | .9785 |
| **MLP** | .9771 | .95883 | .96011 | .95756 | .98465 |

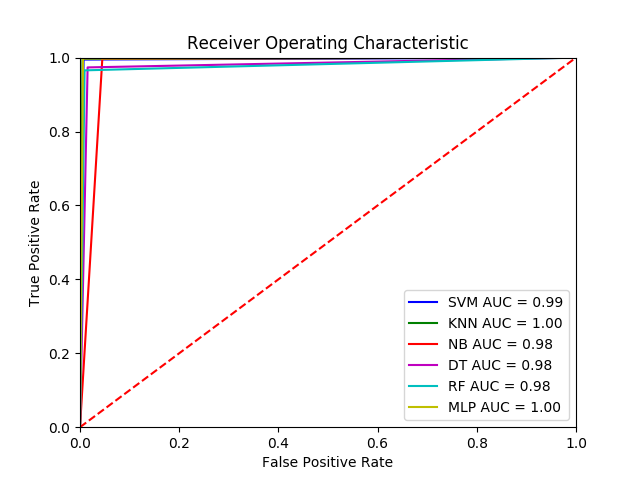


ROC for original 229

3. The results of various models on our 229 Genes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Acr** | **F1** | **Prisn** | **Rcl** | **Speci** |
| **SVM** | .99335 | .98814 | .98168 | .99469 | .99284 |
| **KNN** | .99557 | .99208 | .98688 | .99735 | .99488 |
| **NB** | .96677 | .94353 | .89523 | .99735 | .95496 |
| **D T** | .98154 | .96706 | .96073 | .97347 | .98465 |
| **R F** | .98375 | .970667 | .97587 | .96551 | .99079 |
| **MLP** | .99483 | .99077 | .98429 | .99735 | .99386 |

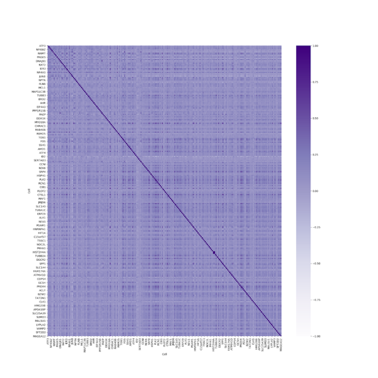
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **TN** | **FP** | **FN** | **TP** |
| SVM | 970 | 7 | 2 | 375 |
| KNN | 972 | 5 | 1 | 376 |
| NB | 933 | 44 | 1 | 376 |
| DT | 962 | 15 | 10 | 367 |
| RF | 968 | 9 | 13 | 364 |
| MLP | 971 | 6 | 1 | 376 |



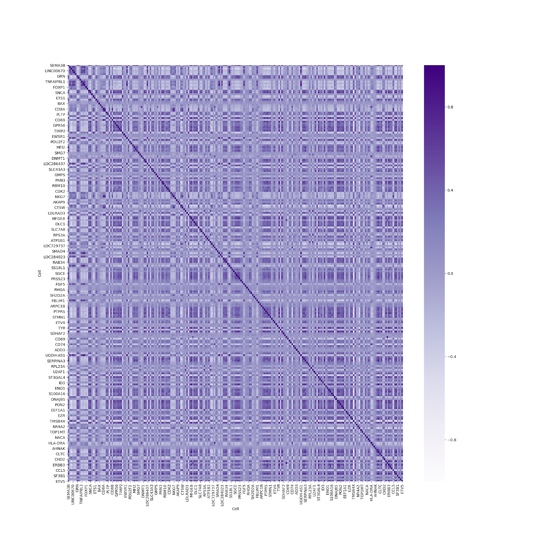
ROC for our 229

As you can see from the from the above 2 charts and ROC curve it is seen that our 229 genes perform better than the original 229 genes.

4. Heatmaps of original vs our



Original 229



Our 229

As perceived in the above heatmaps the heatmap our genes is more corelated than the paper’s 229 gene.

5. Personal Accomplishments:

Created an SVM with an accuracy of Train accuracy - 0.9961, Test accuracy - 0.9955, the parameters for the SVM are (﻿'C': 1, 'degree': 1, 'gamma': 'scale', 'kernel': 'linear'). Created a workable script and supporting files for the GSEA. Experimented on GSEA, searched a lot of biology and found functions of GENES.

# 5 What we learnt

The most important thing that I learned was that there needs to be a lot more data than what we have. It is always good to have domain knowledge and it is better to first understand what the data is about rather than just considering it as some data and fitting a model on top. Sometimes we develop models on data we get certain accuracy but later we reaslise that there are some unnecessary varriables that need to be omitted because they create certain un necessary dependencies in the model which makes it impossible for it to be reproducable or extendable or even it makes it biologically in correct.

This was My first time working on any biological or genetic data, thus this proved to be a very important experience in understanding how to handle this data, how to preprocess it and how to extract features.

An important concept I learned was that the genes can be downregulated and still have a significant impact on outcome. Hence it is not wise to drop downregulated Genes.

# 6 Redesigning & Future enhancements

Looking back at the system or the project as a whole there are certain things that I would do differently and a lot of things that I would keep. The main changes would be in the preprocessing we still are not sure about normalization wether it should be gene wise, sample wise or a hybrid one. We would like to look more into that. Second thing is it seems that There are a few Housekeeping genes that have effect on cancer and hence we would avoid removing them in the future. Next is we can try to add the GSEA data as the feature. Put in all the data and see if it enriches the genesets, if they are getting enriched check the set overlap as the features. I do not think we did anything wrong or there are any improvements that I would like in the methods part.

# 7 Impact of our work

Our work will have a decent impact to the scientific community, we found 3 genes which were able to classify the data to malignant or not upto a accuracy of 97%, On further studying those genes we found out that one of the gene was a promotor of tumors they were EDNRB, CSPG4, CD74, one of them was found to be the control factor for survivability, growth and mobility of the cells and has been identified as a possible target for immune therapy. The third was candidate target for the Immunotherapy of B-Cell Neoplasms and has high expression seems to be crucial for better patient prognosis of brain metastatic tumor cells and also is a positive prognostic marker for patient survival. We validated the findings. There were 5 other genes in our final geneset which have been very important in determining melanoma they were CD2, DNMT1, EIF1AX, CSNK1A1, HLA-DRA.

The protein encoded by this gene is a surface antigen found on all peripheral blood T-cells, it may be an independent predictor of disease recurrence and overall survival among patients with primary cutaneous melanoma. Another one encodes an enzyme for DNA Methylation overexpression of the DNMT1 gene may result in methylation and silencing of genes called tumor suppressors. Found to be an independent predictor of poor overall survival in melanoma patients. This gene encodes an essential eukaryotic translation initiation factor. EIF1AX mutations occur in approximately 17-24% of uveal melanomas. Mutations in EIF1AX have been associated with a good prognosis in uveal melanoma. Another of the gene was a tumor suppressor gene involved in cell motility, which underlies invasion and metastasis of human cancer and was considered as novel targets for anti-metastatic therapies. The last one could contribute to enhanced anti-tumor immunity, significantly associated with better overall survival mediation can increase melanoma migration and metastatic dissemination and was also 46th most closely correlated gene in cutaneous melanoma.

**In total, our research found,** 11/229 genes are not related to cancer**,** 26/229 genes have undetermined relation192/229 genes relating to cancer. So, if you consider the 29 can be explored with the perspective of probable relation to melanoma.

# 8 Contributions

Every one worked really well, we were happy and satisfied as a team. Every one tried to work the whole time, I agree that not everyone does the best but Devin Really did great. He held up the team together. Since we all had no biology background we were working on a lot of stuff. We read a lot of stuff. The team decided I should look on the biological aspects of the paper Devin looked into methods and Dominic and Dylan helped him a lot. The basic expoloring of the data looking at the genes, looking at normalization, deleting the under expresses with the threshold was explored by Devin, Dylan and Dominic tried to explore the PCA and helped Devin code the normalization.

The Idea of the column slice was of Dominic he tried working on it but due to a low accuracy and having not much further ahead he decided to drop it. Dylan helped a lot in collecting the results from us and analyzing them. All the presentations were made together in live meetings. We had regular meetings to keep up with the progress. Me and Devin helped Dylan and Dominic with concepts and doubts. The models were developed and decided by all but The GSEA, the list of housekeeping genes, the publication of the 2305 cancer, and the biological comparison with finding info was done by me. The analysis of all the 229 genes and looking their functions and importance was again a complete team task.

**Acknowledgment**

The authors wish to thank Professor. Hong. This work was supported in part by the class discussions on CSE5713.

**References**

1. Chen YY, Chang WA, Lin ES, Chen YJ, Kuo PL. Expressions of HLA Class II Genes in Cutaneous Melanoma Were Associated with Clinical Outcome: Bioinformatics Approaches and Systematic Analysis of Public Microarray and RNA-Seq Datasets. Diagnostics (Basel). 2019
2. Costantini, Barbieri. The HLA-DR mediated signalling 2017
3. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens.
4. *Repana D., Nulsen J., Dressler L., Bortolomeazzi M., Venkata S. K., Tourna A., Yakovleva A., Palmieri T., and Ciccarelli F.D.*
5. Repana, Dimitra, et al. "The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens." *Genome biology*20.1 (2019): 1.
6. Price, Wanshura, Yang et al. CSPG4, a potential therapeutic target, facilitates malignant progression of melanoma. 2011
7. Tirosh, Izar, Prakadan at al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. 2016
8. Zeiner PS, Zinke J, Kowalewski DJ, et al. CD74 regulates complexity of tumor cell HLA class II peptidome in brain metastasis and is a positive prognostic marker for patient survival. Acta Neuropathol Commun. 2018
9. Ewens, Kathryn G., et al. "Chromosome 3 status combined with BAP1 and EIF1AX mutation profiles are associated with metastasis in uveal melanoma." *Investigative ophthalmology & visual science* 55.8 (2014): 5160-5167.
10. Decatur, Christina L., et al. "Driver mutations in uveal melanoma: associations with gene expression profile and patient outcomes." *JAMA ophthalmology* 134.7 (2016): 728-733.
11. Ferrara, Gerardo, and Mirna Bradamante. "Melanocytic Skin Tumors: Does the Molecular Progression Model Fit With the Routine Clinicopathological Practice?." *Dermatology Practical & Conceptual* 10.1 (2020).
12. Sun D, Zhou M, Kowolik CM, et al. Differential expression patterns of capping protein, protein phosphatase 1, and casein kinase 1 may serve as diagnostic markers for malignant melanoma. Melanoma Res. 2011;21(4):335–343
13. Maric, Helena, et al. "DNMT1 and DNMT3B genetic polymorphisms affect the clinical course and outcome of melanoma patients." *Melanoma Research* 29.6 (2019): 596-602.
14. Harcharik, Bernardo, Moskalenko et. al Defining the role of CD2 in disease progression and overall survival among patients with completely resected stage-II to -III cutaneous melanoma. (2012)
15. Zeiner, P. S., et al. "CD74 regulates complexity of tumor cell HLA class II peptidome in brain metastasis and is a positive prognostic marker for patient survival." *Acta neuropathologica communications* 6.1 (2018): 18
16. E. Eisenberg, E. Levaynon., et al. “Human housekeeping genes revisited”, Trends in Genetics, October 2013, Vol. 29, No. 10
17. M. A. Hall, “Correlation-based feature selection for discrete and nu- meric class machine learning,” in Proceedings of the 17th International Conference on Machine Learning (ICML ’00), pp. 359–366, Morgan Kaufmann, San Francisco, Calif, USA, 2000.