Analyzing mRNA Based Prostate Cancer Dataset

Ehsan Ul Haque
Group 6: Ehsan Ul Haque, Aditya Kulkarni, Mohammad Behzadi, Mohammad Madani
Advisor: Dr. Seung-Hyun Hong

***Abstract*—In this data mining project, we sought to look at developing detection and prediction models for different cancer types. With this in mind, we looked at four different NCBI geo datasets on various cancer types and examined different data mining techniques keeping our individual and team goals in mind. Individual goals were set to look closely on Prostate Cancer data in a serum mRNA dataset, GSE 112264. Additionally, two different combination of classifiers (mRNA vs peptide) have been trained on combined datasets as part of team goal. The results show great accuracy and detection rate for Prostate Cancer (PC) and Negative Prostate Biopsies (NPB) for a trained classifier. Further, Peptide and mRNA-based classifiers has been compared in terms of efficacy and data characteristics. Our approaches, findings, data mining results, and future directions have been discussed in the paper.**

***Index Terms*—data mining, pca, t-sne, clustering, classification, prostate cancer, cancer detection**

I. Overview

Having a set of data that is complete, easy to interpret, and provide valuable insights are the key part of any data mining task. At the data collection phase, we tried to cross validate these preconditions to get a good set of data. Specifically, we wanted to collect a set of data, that is i) compatible with each other (e.g., similar organ, tissue etc.) and ii) can be easily combined. These considerations led us to choose four datasets from the series GSE 112264, GSE 124158, GSE 52580, and GSE 52581, respectively. All of these datasets contain Homo Sapiens data for various cancer types. The first two datasets contained mRNA serum-based cancer data and the last two contained peptide-based cancer data. Our overall objective was to analyze and perform different data mining and data analysis techniques in both individual and combined datasets. I worked on the dataset GSE 112264 which contains mRNA-based cancer data. Following sections include description, of the dataset, individual and combined goals, and the approaches used to tackle these goals.

A. Dataset Description

The dataset that I worked on has been collected from the series GSE 112264 and has been used in the paper by Urabe et al. [1]. The platform is $GPL212633D - GeneHumanmiRNAV211.0.0$. Overall, the dataset contains serum mRNA profiles for 1591 male samples, including –

- 809 Prostate Cancer (PC) patients
- 241 Negative Prostate Biopsies (NPB)
- 41 non-cancer controls (CTL)
- 500 other cancer data

The dataset contains a good number of prostate cancer and negative biopsy data, which has potential to look closely at the differences between these two groups of patients. Additionally, this dataset also includes non-cancer control data that can be used as a baseline for observing variations in prostate cancer mRNA. Overall, the dataset showed potential for good analyzability on prostate cancer patients.

B. Motivation: Why Important?

Prostate cancer is one of the common cancer types. While some types of prostate cancer grow slowly need only minimal treatment or often gets better without any treatment, however, some types of this cancer can get very aggressive and are known to grow very quickly. Moreover, the high false rate of Prostate Specific Antigens (PSA) often leads to negative prostate biopsies (NPB). This phase does not definitively exclude the chances for prostate cancer and often require further investigation and close observation for the development of prostate cancer. Thus, accurate detection of prostate cancer and negative biopsies and modelling divergences between these two phases are crucial for the best possible treatment of prostate cancer. In this work we hypothesize that serum mRNA combination could be a strong determinant for differentiating these two groups and sought to validate this assumption throughout different phases of the project.

C. Individual and Team Goals

The dataset GSE 112264 gives a strong base for the objective of this project because it includes a good amount of prostate cancer and negative biopsy mRNA data. Additionally, healthy control mRNA data is also present for comparative reasons. With the motivation of early detections of prostate cancer and negative biopsies in mind, I set following individual goals –

- Comparing prostate Cancer mRNA and healthy control mRNA, which might help determining the divergences among the groups
- Comparing negative biopsy miRNA to healthy control miRNA, to understand the deviation from normal mRNA in the negative biopsy mRNA
- Work on training classifiers on detecting prostate cancer and negative prostate biopsies, possibly with high accuracy to give confidence on the detection of prostate cancers and negative biopsies

Alongside our individual goals, as a part of group, we came up with a team goal to compare mRNA and peptide-based classifiers. As mentioned earlier, among the four datasets, two of them contains mRNA-based cancer data and other two contains peptide-based cancer data. Thus, we found it

interesting to look at how these different tissues compare with each other and which gives the best result and why.

In the next section we will discuss the methodologies that was used to tackle the individual and team goals.

II. Methodologies and Change Accommodations

Over different iterations and phases of the project we performed different data mining techniques that aligns with achieving the goals. We did not perform any modifications to our initial objectives and goals due to achieving satisfactory results in the preliminary analysis. Here we discuss our approaches towards achieving individual and team goals.

A. Towards Individual Goals

In phase one, some preliminary analysis was performed on the data. Before doing the analysis, data-preprocessing has been performed. In the preprocessing step, we checked and removed null values. Next, normalization and centering were performed. Finally, the original GSE 112264 dataset was split into three reduced data sets, each with two classes. These are Prostate Cancer vs Control, Prostate Cancer vs Negative Biopsy, and Negative Biopsy vs Control. Prostate Cancer vs Control dataset contains a total of 850 samples, with 809 prostate cancer and 41 control data. Prostate Cancer vs Negative Biopsy dataset contains a total of 1050 samples, with 809 prostate cancer and 241 control data. And finally, Negative Biopsy vs Control dataset has a sample size of 282, with 241 negative biopsies and 41 control data.

In the preliminary phase, two different dimension reduction techniques PCA and t-SNE was performed on the reduced datasets to see the distinction of feature distribution in lower dimensions. In phase two, data clustering was done with k-means to look for possible clusters on the data. Finally, classification was performed on the dataset to see the detection and prediction ability in the data.

As the classification algorithm, KNN and Random Forest was chosen and applied. Accuracy has been used for the model evaluation. We have used two different metrics to compute accuracy of the models.

- Repeated cross validation - 10-fold, 10 repetitions each
- train/test set with a split ration of .75/.25 to measure prediction accuracy of the trained model

Furthermore, all three of the datasets had imbalanced samples across classes. To tackle this issue, up-sampling or down-sampling of data has been performed before running the classifiers. The sampling has been done using R's Caret package [2], which has built in support for both up and down sampling of the data before training. Also, R was used for coding tasks in all phases of the project.

B. Towards Team Goals

Our team goal was to build and compare classification models on different cancer types. Towards this, the overall process was divided into two parts. Two team members worked on each part separately before comparing the results.

Part one of the team goal that I worked on was to combine two miRNA-based datasets (GSE 112264, GSE 124158) together and run different classification models on the combined data. Once the data was combined into a single miRNA cancer dataset, the first thing I did was to perform necessary preprocessing on the data. There were some unique features on each of the mRNA datasets alongside the overlapping features. Thus, after combination, null was introduced in the data. Further, two datasets had some slightly different names in the labels for some cancer types. Trailing spaces was also preventing the same cancer types to go into the same factors. Thus, the preprocessing was needed to address those issues before building the classification models.

After data cleaning, the cleaned combined data had 17 labels, with sixteen different cancer types and control data. The final combined dataset had a total of 3003 samples, with 2540 mRNA features. Class sample size was varied from 30 to 800, causing imbalance between classes.

Random Forest and SVM with RBF kernel was chosen for the classification on the combined data. Same as before, accuracy was used for evaluation and was measured using 10-fold cross validation. We did not do 10 repetitions as before due to the GPU constrains of the machine where training was performed.

Up sampling of data was used to resolve imbalance issue and R was used for coding tasks.

III. Results and Accomplishments

In each stages of the project, we set our methods based on our results from previous phase. Here I describe the results and accomplishments I got from different phases of the project.

A. Individual Accomplishments

1) Dimension Reduction: In the initial phase of the project, after choosing the dataset, I performed some exploratory analysis on the data. Dimension reduction techniques PCA and tSNE was performed to understand variability and separation in lower dimensions. PCA and tSNE was done on Prostate Cancer (PC) vs Negative Biopsies (NPB) vs control (CTL) data. The result is shown in Fig. 1. Results show that, while CTL shows clear segregation, PC and NPB overlapped a little bit.

Different tSNE models was run with hyperparameter tuning. The best results have been achieved for $perplexity = 40$ and $max - iter = 1000$. Compared to PCA, tSNE showed much better segregation of the classes and predicted success for good clustering and classification. tSNE result is shown in Fig. 1.

2) Data Clustering: In the next phase clustering algorithm K-means was performed in the three different binary datasets, PC vs NPB, PC vs CTL, and NPB vs CTL.

For the PC vs CTL data, K-means generated best results with max silhouette value < 0.1 for 2 clusters. The clusters overlapped a little bit, as shown in Fig. 3.

For the PC vs NPB data, similar to the first one, K-means generated best results with max silhouette value < 0.1 for 2

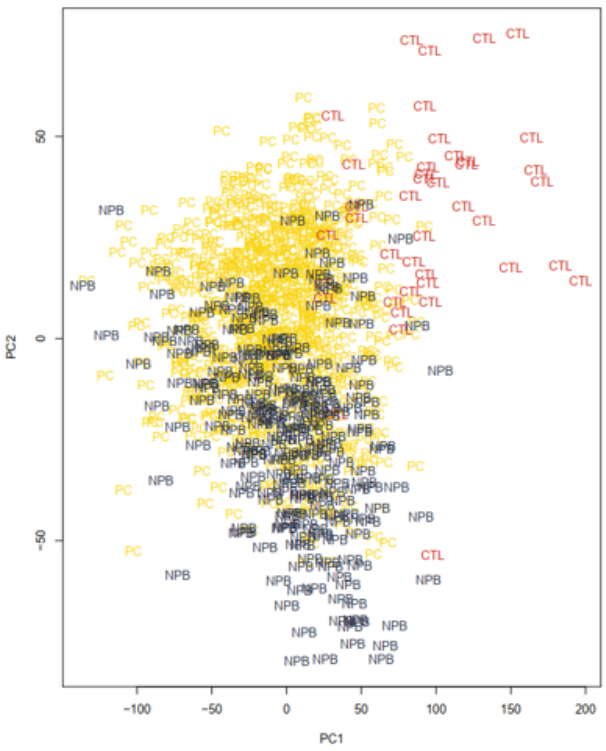

Fig. 1. PCA graph

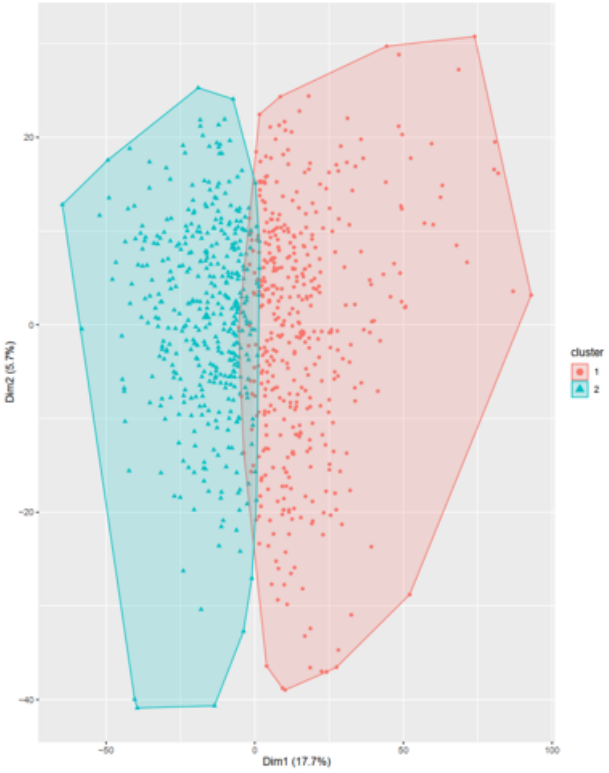

Fig. 3. k-Means: PC vs CTL

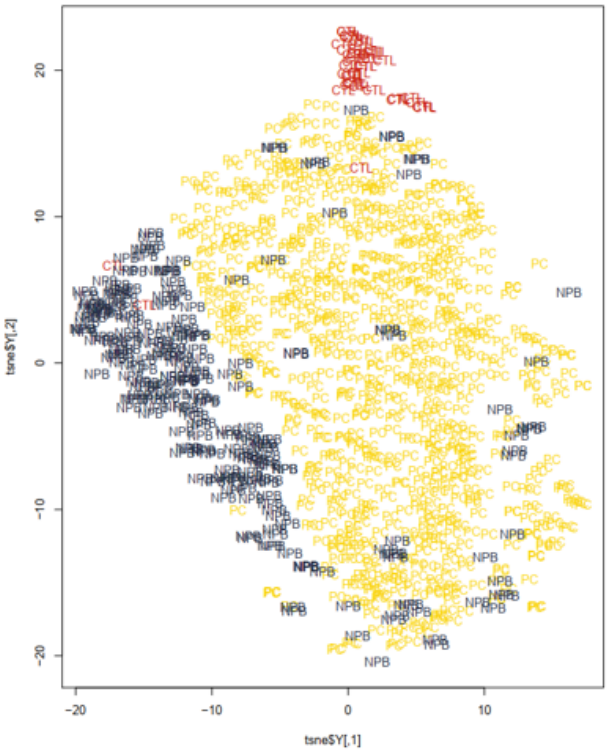

Fig. 2. t-SNE graph

clusters. Overlapping of the clusters were again noticed, as shown in Fig. 4.

However, for the NPB vs CTL data, max silhouette was found with value > 0.25 for two clusters. Unlike the first two, result showed clear separation of clusters, shown in Fig. 5.

Overall, for all three of the binary datasets, best results were found for two clusters. Even though some overlapping was noticed in two of the datasets with low silhouette value, nonetheless, results supported viability of good classification model accuracy in the data, which we validated in the final phase.

3) Classification: Based on the clustering results, we expected good accuracy from the trained classifiers in the datasets. Overall, we build multiple classifiers, one KNN and one or two Random Forest in each of the three binary datasets mentioned above. Here we describe our findings of the classifier accuracies.

The first dataset contained 809 PC and only 41 CTL data. As mentioned before, we used up-sampling of the data to generate random samples for the minority class to match the sample size of the majority class. We ran KNN for different k values as hyperparameter tuning. The best result has been obtained for k = 4. The train accuracy was found over 99% with the cross-validation method, which is surprisingly high. We then run the prediction model on the test dataset (.75/.25 split) and found a prediction accuracy of around 99.5%, which matches the training accuracy. We then ran Random Forest for the same setup with tuning different mtry values. Best model

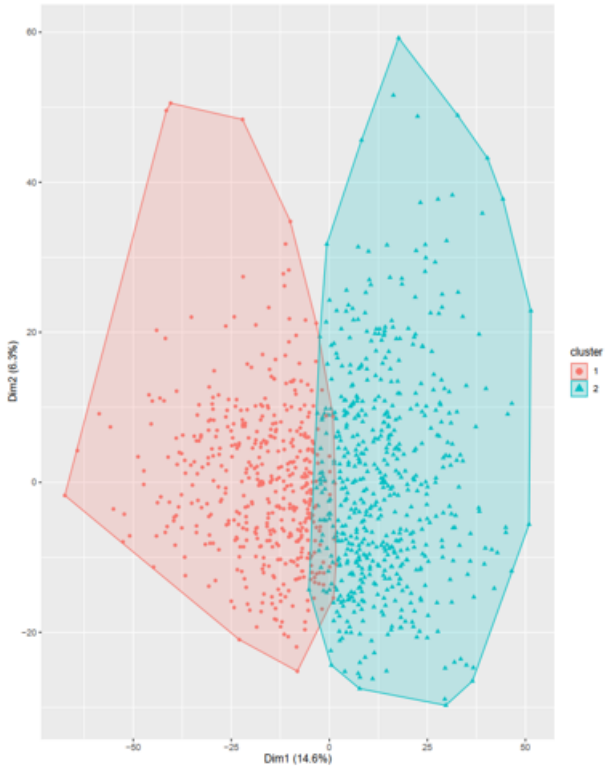

Fig. 4. k-Means: PC vs NPB

Fig. 5. k-Means: NPB vs CTL

was generated for mtry = 7, again with an accuracy of more than 99% with cross validation. Prediction accuracy in the test data was around 98.5%. Both classifiers work similarly and was efficient in detecting prostate cancer.

In the next dataset, there were a total of 1150 samples, with 809 PC and 241 NPB samples. We ran KNN and Random forest on both the up-sampled and down-sampled data (for Random Forest) in exact same setup as before. For the KNN, best model was generated for k = 1. The accuracy in this case was around 70%, with cross validation. Prediction accuracy was around 73%, which is not very good. However, Random Forest worked better for this dataset, with a training accuracy of about 95% in the up-sampled data, and 98% in the down-sampled data. The reason we performed down-sampling only for this dataset because the minority class has reasonable data compared to the last dataset (241 NPB data compared to only 41 CTL data). Prediction accuracy for Random Forest was similarly high, around 94% and 98% respectively. For both cases mtry = 10 was generated the best model.

The final dataset had 241 NPB and 41 CTL data. Similar to the first dataset, we achieved very good results for both KNN and Random Forest classifiers. KNN generated the best model for k = 2 with a training and prediction accuracy of around 99%. Random forest generated the best model with mtry = 2 and got the similar 99% accuracy in both cases. Thus, like the first dataset, both classifiers worked very well for this dataset. The comparison between classifier results are showed in Table I

B. Team Accomplishments

For the team comparison purposes, I trained two classifiers in the combined dataset with 17 classes (different cancer types and control data). As mentioned before, the dataset was highly imbalanced and thus, up-sampling was performed before classification tasks. Initially we wanted to run Random Forest, SVM different kernels, and MLP on the combined dataset to see which one worked better. But, due to high amount of time taken in each classifications and GPU constraints, we only ran Random Forest and SVM with RBF classifiers in the combined dataset.

For random Forest, we found an accuracy of about 69% for the best model, which was generated for mtry = 10. SVM with RBF kernel generated an accuracy of around 72% for cost parameter, C = 8, which was the best model. Both classifiers worked well, but not good enough compared to the results I got in the binary datasets. Compared to the peptide-based classifiers (accuracy around 90%), these results were poor. Thus, we concluded that, peptide-based classification works better for cross cancer detection. However, these results must be interpreted alongside considering various factor that may have impact on the findings, which is presented in the limitations section. Further, the data mining findings of the results will be presented in later sections.

TABLE I
COMPARISON OF CLASSIFIERS IN DIFFERENT DATASETS

Dataset	Algorithm	Sampling	Best Tuning	Accuracy (cross-fold)	Accuracy (train/test)
PC vs CTL	KNN	Up	K = 4	99.8%	99.53%
	RF	Up	mtry = 7	99.8%	98.58%
PC vs NPB	KNN	Up	K = 1	69.5%	73.2%
	RF	Up	mtry = 10	95.3%	94.27%
	RF	Down	mtry = 10	98.3%	98%
NPB vs CTL	KNN	Up	K = 2	98.9%	98.6%
	RF	Up	mtry = 2	99%	98.5%

IV. GATHERED KNOWLEDGE AND EXPERIENCES (WHAT I LEARNED?)

With the experience of my individual work alongside being a team member, I have learned a lot from this data mining project. The knowledge gathering process can be viewed from two different angles.

A. As a Team

For the project, we worked as a team. I was in Group 6 with three of my fellow team members. As part of the team, I learned how a team project can be broken down into individual parts and then merged back together for the outcome. A team works with collaborations from the team members. As data mining can be a tedious and time-consuming task from the beginning to end, there are a lot of scope for collaboration, from the beginning, where brainstorming on the data collection and goals are set, to the very end, when data mining results have been extracted and evaluated. As a team, we met occasionally to discuss about our progress and tried to solve problems as a group when someone was stuck. Team discussions helped us a lot in understanding the sequential process of data mining and how they can be applied in real life.

B. As a Team Member

Alongside working as part of a team, I set up and worked on my individual data mining goals based on the chosen dataset. While working on these goals, I learned the overall process of datamining, from the preprocessing, data cleaning to the extraction of results and visually representing the findings. The lecture videos helped me a lot in understanding the underlying data mining techniques in different steps and how they can be combined for best results. Overall, I found the course and the project very helpful in understanding the concepts of data mining and how they can be applied in real world scenarios. As a newbie in the field of data mining, the course and the project set a basis for me to walk along the path in the field of data mining.

V. RESULT DISCUSSION AND FUTURE WORK (REDESIGNING THE SYSTEM)

A. Data Mining Findings (Individual)

In the project, I worked on a Prostate Cancer dataset collected from GSE 112264 and set my individual goals to look at divergences on prostate cancer and negative biopsies, which is important because of difficulties related to high false rate of Prostate Specific Antigens (PSA). Here are my key takeaways based on the results –

- mRNA based classification works fairly well for separating prostate cancer patients from negative biopsies which can improve early detection of prostate cancer, which is crucial for successful treatment.
- Accuracy of detecting negative biopsy patients from healthy individuals was also found very good, which may enable early detection and chance for close observation for further development of prostate cancer.

The results aligned with my expectations and the goals I set for my individual work.

B. Data Mining Findings (As a Team)

As a team our target was to compare mRNA vs peptide-based classification for cancer detection. After combining datasets and training classifiers on the combined datasets, we found better accuracy from the peptide-based classifiers (around 90%), which can be leveraged in designing effective cancer detection and prediction models. Further, we tried to rationalize the reasoning why mRNA-based models did not work as similarly as peptide-based models which is important to consider for interpretation and generalizability of our results.

C. Limitations And Failures

For the individual work, the dataset that I worked on was imbalanced in nature. Even though up-sampling was performed in the data before training, the resolution may not be the best to accurate explanations of the findings. Additional sampling techniques like ROSE, SMOTE etc. needs to be considered and compared to see if the findings stay the same. Also, efficacy of our model on different prostate cancer datasets may be necessary to validate the results that we found.

I ran classification on the overall feature set presented in the dataset. However, it might be possible that certain features have high significance towards the divergences compared to others. Thus, selecting important features, which also helps resolving ”curse of dimensionality” might help understanding the results better.

For the team tasks, we set up two different sub tasks (mRNA classification vs peptide classification) and different team members worked on the sub-tasks in different set up. So, keeping a same unbiased environment for building and running the model was not possible. Also, the protein-based

datasets were more balanced across groups compared to the mRNA-based datasets, which may have impact on the high accuracy of the peptide-based classifiers. Further, due to GPU constraints on the machines, not exactly same classification algorithms were run on the two subtasks. MLP or different kernel of SVM may have worked differently on mRNA-based datasets. Rigorous hyperparameter tuning can also attribute to a different result from the mRNA models. Thus, the comparison of the findings between the two different classifiers need to incorporate these considerations.

VI. Future Work

Overall, the result we got are very interesting, not only in terms of efficacy, but also the underlying takeaways and data mining findings that we gathered. These findings also led to possible future direction and expansion of our current work. From the individual work, I found great efficacy of classification model for both prostate cancer and negative prostate biopsies. However, the model did not incorporate any feature selection strategy prior to the classification. Extraction of significant features and running the model on those features can get us more valuable insight on the negative biopsy complicacy. For the new model, my expectations will be to improve the performance of detecting Prostate Specific Antigens (PSA), so that negative biopsy patients can be categorized into groups based on their probabilistic risk of getting the disease in future.

Also, in this work we looked considered limited instruments. For example, I only used k-means fro clustering purposes and only KNN and Random Forest for classification. However, other algorithms and classifiers,including different sets of hyperparameter tuning may be better suited for this dataset. Future work can look at different models using other popular and relevant classification algorithms to see which and works better and why.

Resolving the limitations, that we had as per our team goal, could also lead us to a better understanding of mRNA and peptide models and confidently determine the best one among them. Specifically, we want to consider a future model with following adjustments in mind -

- Choosing datasets that have close distribution of cancer types for both mRNA and peptide types
- Datasets having lower differences in terms of imbalance across cancer types
- Choosing a set of instruments and tools (e.g., same set of classifiers) for all datasets, to make the results more comparable and relevant to each other
- Controlling external environment related factors and machine restriction for providing a unified baseline for all the classifiers
- integrating relevant pre-processing, feature selection with the classification tasks

Future models incorporating above considerations would give us a better understanding of mRNA vs peptide based classification and provide recommendations based on the findings.

VII. Impact of the New System to the Community

The new system that we discussed in the future works, can work for the improvement in medical science. Early detection of cancer is one of the key parts for best treatment and chance for success. Technological advancement in a incremental process that advances and gets better with time. There is always room for improvements and optimization.

Current technology for prostate cancer detection is not accurate due to prevalence of negative biopsies. This often leads to delay in cancer detection, which can significantly reduce the chance for treatment. Prostate Specific Antigen (PSA) detection contains high false positive rate. Our new model sought to reduce this issue by looking into a close set of features that can best describe the divergences between negative biopsies and actual detection of cancer.

Furthermore, Looking into mRNA vs peptide differences will help understand the best architecture for cancer detection in general. The proposed model considers various control variables to find and discuss mRNA vs peptide-based architectures with confident. Choosing one over the other may be beneficial for different reasons. It can further reduce the cost for detection and increase the accuracy of the current models that is being used in practice.

Considering all these factors, we believe that our proposed model and future targets can be impactful for the advancement of medical science.

VIII. Summary

In this project, our team worked over a collection of cancer datasets, that is collected from the series GSE 112264, GSE 124158, GSE 52580, and GSE 52581, respectively. All of these datasets contain data for various cancer types for humans. The first two datasets contained mRNA serum-based cancer data and the last two contained peptide-based cancer data. Overall objective was to perform different data mining in both individual and combined datasets.

Individually I looked at prostate cancer dataset and found great accuracy on the prostate cancer and negative biopsy generative models. Dimension reduction techniques showed differences among different types of cancers and suggested possibility of good clustering and classification.

Generated clusters showed separation among prostate cancer, negative biopsy, and control mRNAs and predicted good classification accuracy. I further worked on two different classification tasks on each of the datasets and validated that classification models are highly accurate in detecting prostate cancer and negative prostate biopsies, which is very important because negative biopsies need close observation for future development of prostate cancer.

Further, as part of the team goal, we combined different datasets, and trained classifiers on combined datasets. We then compared the accuracy of different classifiers for mRNA and peptide-based architectures and found better accuracy from a peptide-based cancer detection approach.

Overall, though there were several limitations of our work, this was a first step for us in the field of data mining and the

knowledge that we gathered throughout the course and project was very helpful in understanding and applying techniques for effective data mining.

REFERENCES

[1] Urabe, Fumihiko, et al. “Large-Scale Circulating MicroRNA Profiling for the Liquid Biopsy of Prostate Cancer.” Clinical Cancer Research, vol. 25, no. 10, 2019, pp. 3016–3025., doi:10.1158/1078-0432.ccr-18-2849.

[2] Kuhn, Max. ”Building predictive models in R using the caret package.” J Stat Softw 28.5 (2008): 1-26.