Cancer Type Classification and Mortality Prediction

Jing Lin[1], Lawrence Wong[1], Maggie Wu[1], Yun Boyer[1]

***Abstract*— Abnormal gene expressions due to mutations in gene regulatory mechanisms result in the uncontrolled proliferation of cancer cells. Recent advances in machine learning as applied to gene expression profiling data have achieved great accuracy in the classification of cancer cell and have aided prediction on clinical outcomes for cancer patients. However, these prediction methods often deal with gene expression microarray datasets in high-dimensional space. Moreover, the time required to generate these datasets scales proportionally to the number of genes probed. We demonstrate using a logistic regression model and a 2D convolutional neural network that the classification accuracy will not suffer significantly when classifying with a selected subset of the gene expression data. This subset of genes is carefully selected by and verified with multiple feature selection techniques and extensive literature research on known driver genes. Subsequently, this subset is further utilized to predict the mortality of cancer patients using a ridge regression model and a convolutional neural network model. These results combined will help shed light on the mechanisms of gene regulation in cancer cells and set the stage better for future gene expression data analysis studies.**

I. INTRODUCTION

Classification of cancer cells is one of the most important works in cancer diagnosis and has been applied to multiple common cancers[2][3][4]. Traditional cancer classification methods have always been clinical-based but have limited diagnostic ability[22]. Cancer research over the past decades has improved early diagnosis and prognosis of cancers[25]. With recent technological advancement in the field of medicine, a large amount of cancer data has been collected and studied in the medical research community.

With current technologies, cancer classification using gene expression data of the entire genome (20,000 - 25,000 genes) has become more feasible and has gained importance in genetic testing procedures. Numerous studies have shown that the expression levels of cancer-related genes inform us about the cancer type and its mortality rate[5][6]. Amongst the entire human genome, 291 (approximately 1.5%) are reported as cancer-related genes. Therefore, the majority of the gene expression data may actually be unnecessary for the cancer types classification[7]. Through this study, we examined if we can, as accurately as using all genes in the human genome, classify cancer types using only a subset of genes. We used this result to validate existing cancer-related genes and discover new genes that can serve as unique biomarkers for different cancer types. Knowing the subset of important genes can drastically decrease the time it takes to create a gene expression dataset per sample. This is because the time to measure gene expression using microarray technologies increases proportionally to the number of genes probed. Finally, using this subset, we predicted the mortality of cancer patients and examined our prediction accuracy using common regression metrics.

The paper is ordered as follows: we first list out related works that are most relevant to our project and discuss the novelty of our approaches. Then, we identify the source of our dataset, explain the dataset, and visualize it with dimensional reduction techniques. Next is the method section where we discuss each approach we take to 1) preprocess our data, 2) select important features (i.e genes), 3) classify cancer types, and 4) predict cancer mortality. We then display the results we get and offer explanations for our results in the discussion section. Finally, we conclude our paper with a summary and future works.

II. RELATED WORKS

Machine learning experts have attempted to classify cancer types using various methods. Most of their approaches focus on using K-nearest neighbor(KNN) algorithms[20] and Support Vector Machines (SVM)[21]. Each method has its own advantages and drawbacks based on the problem at hand. KNN is an unsupervised learning algorithm that clusters by the distance between data points. As a result, KNN does not build a model that finds genes that are important for cancer cell type classification. On the other hand, SVM is a supervised learning algorithm that solves binary-class problems by maximizing the margin between positive and negative samples. The algorithm can fail to converge for a large dataset as the runtime of SVM scales proportionally with the size of the input data.

Models that are relevant to our project are network, logistic regression, and neural network models. Some important previous studies that used these models for gene expression data are outlined below.

One study applied network models on gene expression pathways to explain several underlying biological relationships between different genes[9]. As suggested by Pita-Juárez et al., genes do not function alone but rather co-express to carry out certain biological functions. Therefore, their research focused on building a network model using gene co-expressions on the same pathway. The model reveals correlations between pathways that bring insight into the mechanism of complex diseases.

Another study selected informative genes in co-expression networks using centrality measures. Edges in the network are estimated using Pearson's correlation coefficient, which

*This work was not supported by any organization

[1]The authors are with Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. These authors contributed equally to this project.

provides information about expression correlations between genes. Degree and betweenness centrality are then used to detect genes with important functionality since highly connected genes are typically associated with disease-related pathways. However, this study focused on the application to the zebrafish genome and has yet to be tested on the human genome[26].

In another study led by Zhou, et al., a logistic regression model was used to classify cancer types[10]. In addition, the group used Gibbs sampling and Markov chain Monte Carlo methods to select for important genes as features for their models. The models were evaluated with respect to large RNA-seq dataset and showed consistent success in identifying important genes that are biologically-supported. At the same time, their logistic regression model was reported to achieve high classification accuracy.

A different study used a logistic regression model to identify gene expressions that are predictive of mortality induced by squamous cell lung and breast carcinoma[22]. The same logistic regression model was further used to identify immune-related genes associated with breast cancers. Both studies achieved great success in determining relevant genes.

Deep learning approaches combined with dimensionality reduction techniques have been explored for cancer detection. For example, the dimension of the input data can be reduced using either an autoencoder (a type of Artificial Neural Networks trained to ignore signal noises) or Principal Component Analysis (PCA). Analysis of neural network parameters has revealed a list of potential biomarkers for different cancers. However, a deep learning system usually requires a large dataset, which makes it ill-suited for predicting rare cancers with little data[24].

Several deep learning models have been developed for cancer classification. Most of these models are convolutional neural networks (CNN) because of their ability to handle high-dimensional datasets. Mostavi, Milad, et al. proposed a hybrid CNN model with 2D mapping of the gene expression data as input matrices. These multiple binary classification models achieved an accuracy as high as 95.0% among 34 classes (33 cancer genomes and healthy human genome). One interesting aspect to note is the pre-processing step for this hybrid CNN model. It involves reshaping each 1D gene expression data into a 2D matrix, which is an idea borrowed from the typical image classification model in the field of computer vision. A traditional image classification model generates heatmaps for each input image. These heatmaps indicate the weight of each pixel as input to the classification model. This idea can be extended to identify gene significance, which is important for classifying different cancer types[27]. Another interesting aspect is the use of parallel towers, which is inspired by Resnet Model in computer vision. More specifically, the architecture uses two 1D-kernels to slide across each 2D matrix, with one kernel sliding across rows and the other sliding across columns. The authors of this study believed that this CNN architecture can capture more information in the input gene expression[12].

From the above studies, we can conclude that there exists a subset of important genes whose expression patterns can be used to predict cancer types. Upon identification of this subset of genes, we can propose classification models to classify our choice of cancers. Therefore, this project utilizes four feature selection methods to search for genes that characterize cancers. These genes are then fed into our classification and regression models.

III. DATA

A. Source

The Cancer Genome Atlas (TCGA) datasets, provided by National Cancer Institute, catalog genetic mutations responsible for cancer[1]. This project uses the gene expression RNA-seq dataset generated by the Illumina HiSeq 2500 protocol for five cancer types: Breast (BRCA), Colon (COAD), Liver (LIHC), Stomach (STAD), and Thyroid (THCA). Each dataset is a matrix with rows as gene identifiers, columns as patient sample IDs, and entries as mean-normalized gene expressions. The normalization is performed with respect to all TCGA cohorts, and, for each cancer type, the corresponding gene expression values are extracted. Patient samples consist of both solid tissue normals taken from normal tissues near the site of the tumor and primary tumor tissues taken directly from the site of the tumor. This choice of tissue enforces a direct comparison between the same individuals. At the same time, the entire dataset is derived from TCGA and each data point is normalized the same way to ensure a fair comparison between gene expressions in normal and cancer cells.

Cancer Type	# Samples	# Solid Tissue	# Recorded Dead
Breast	1097	121	148
Colon	286	43	67
Liver	371	52	129
Stomach	415	35	154
Thyroid	505	67	14
Total	2,674	318	512

TABLE I
A SUMMARY OF DATASET FOR EACH CANCER

B. Visualization

In order to visualize the structural patterns and variations within the dataset, dimensionality reduction techniques, e.g. Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (tSNE), are applied separately to transform the dataset to a 2D representation. Using such representation, we colored data points with respect to each cancer cell and normal cell. Here, blue data points correspond to normal cells and all other colors correspond to various cancer cells. The results are shown in Fig. 1, with PCA results on the top and tSNE results at the bottom.

Looking at the PCA representation, we can see that liver cells express differently from other cell types, but there is no clear distinction among the rest of the cells. On the other hand, the t-SNE representation shows clear separation among tissue types. This suggests that classification among cancers would be an easier task than classification between cancerous and normal cells.

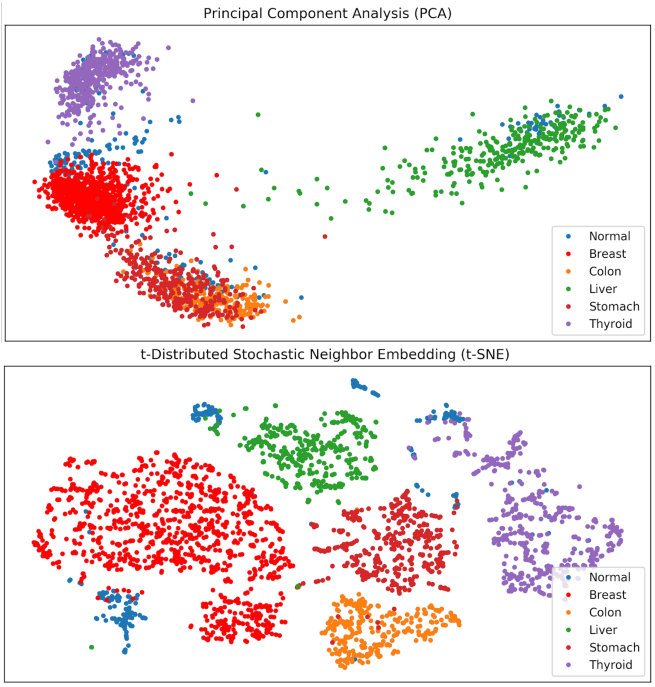

Fig. 1. PCA and t-SNE 2D Visualization of TCGA Dataset

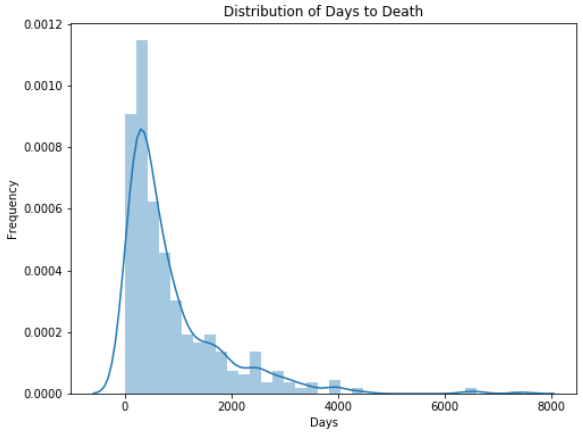

Fig. 2. Mortality Distribution

In addition to cancer type classification, we are also interested in predicting mortality. The distribution of days-to-death is plotted, as shown in Fig. 2. While the distribution is limited to patients who have since passed away, it clearly shows that the life expectancy of patients from the point of diagnosis is mostly between 3 and 5 years. The average life expectancy is ∼2.5 years for a patient diagnosed with one of the five cancers we studied. In addition, the standard deviation for days-to-death is a little more than 2.5 years.

IV. GENE SELECTION METHODS

As stated in our introduction, the time required to obtain gene expression data increases proportionally to the number of genes analyzed. Therefore, it is beneficial to find a subset of genes that is important to each cancer type. Aside from biological experiments that identify driver genes, which are genes whose mutation is responsible for tumor growth, there is no specific technology or algorithm that recognizes important genes. Given the size and the electronic format of these datasets, a computational approach may speed up this research and suggest potentially important genes to focus on in further research.

We used multiple feature selection methods to find important genes. Genes that are identified as important by multiple methods have better credibility than genes selected by one method alone. Therefore, the intersection of the genes selected by each method is used for cancer classification.

The gene selection methods are

1) identified driver genes,
2) gene co-expression network model,
3) multiple hypothesis testing with correction, and
4) multinomial logistic regression with LASSO.

Each method finds up to 100 of the most important genes, which is ∼0.5% of the original dataset. These batches of ∼100 genes are then compared to select a set of important genes for each cancer type.

A. Identified Driver Genes

As published in the article *IntOGen-mutations identifies cancer drivers across tumor types* by Gonzalez-Perez et al., the IntOGen-mutations platform summarizes driver gene information for various cancer types[11]. Using the platform, we are able to find 27-84 driver genes for the set of cancer types we are working with. These genes are good baseline genes for building our network model.

The set of driver genes we found is a subset of the genes recorded by the TCGA dataset. Therefore, correlation coefficients between each gene and each driver gene may be computed.

B. Gene Co-expression Network Model

Network models are typically built from an entire dataset in order to capture the correlations between every pair of genes. However, given that TCGA documented as many as ∼20,000 genes, building a complete correlation adjacency matrix would require us to process a $\sim 20,000 \times 20,000$ matrix. This requires computational power and time beyond the capability of a typical CPU. To work around this problem, we decided to take a slightly different approach. For each cancer type, we

1) compute an absolute Pearson's correlation matrix (P1) between each driver gene and all genes,
2) build an undirected, unweighted network (G1) where each vertex represents a gene and each edge is drawn when there is sufficient correlation (r) between each pair of genes,
3) drop isolated, non-driver genes,
4) compute a second absolute Pearson's correlation matrix (P2) for all remaining genes,

5) build a fully-connected, undirected, weighted graph (G2) between the remaining genes based on P2,
6) compute betweenness centrality for each gene in G2, and
7) select genes with the highest betweenness centrality.

Our alternative approach significantly reduces the size of the problem and generates a network within a reasonable time frame.

Because there are no more than 100 driver genes available for each of the cancer types chosen, both P1 and P2 for each cancer type have matrix sizes that are smaller than $\sim 20,000 \times 100$. In addition, we construct G1 by creating edges whose $r \geq 0.8$. Because we only care about genes that are highly correlated with driver genes, many genes are dropped due to their low correlations with driver genes.

The second Pearson's correlation matrix - P2 - only computes pairwise correlation on the remaining genes (535 for THCA, 191 for BRCA, 104 for COAD, 55 for LIHC, 107 for STAD). Therefore, the size of P2 is significantly smaller than that of P1. This reduction makes it possible for us to build fully-connected, undirected, unweighted graphs - G2s. G2s focus specifically on a subset of potentially important genes to perform additional analysis.

In order to evaluate the importance of a gene in the context of cancer classification, betweenness centrality is chosen. Betweenness centrality of a node measures the ratio of the number of shortest paths between any two other nodes that pass through the node to the total number of shortest paths between the same two nodes. For example, the betweenness centrality $g(v)$ for a node v is given by the following expression.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}} \tag{1}$$

$\sigma_{s,t} =$ # shortest paths between a node s and a node t

$$\sigma_{s,t}(v) = \sigma_{s,t} \text{ that pass through node } v$$

A gene is important if it lies in multiple pathways associated with cancer. To effectively evaluate this, we believe betweenness centrality is the most reasonable measure. This method assumes that the expression of driver genes is correlated with the expression of genes that can serve as biomarkers.

C. Hypothesis Testing with Correction

Another method we chose to identify important genes is multiple hypothesis testing with False Discovery Rate (FDR) correction. Independent t-test for two samples is chosen to compare the gene expression distribution of normal cells to that of cancerous cells. Specifically, we

1) identify each gene expression distribution across all cancerous cells,
2) identify each gene expression distribution across all normal cells, and
3) perform independent t-tests to compare these distributions

A gene is considered differentially expressed if its expression level is much greater or much less than usual. Comparisons are only drawn for expression levels of the same gene. Using the result of each t-test, we identify genes that are differentially expressed as important genes.

The t-test threshold for significance is $\alpha = 0.05$. We attempted to use both Holm-Bonferroni and Benjamini-Hochberg correction to correct for multiple hypothesis testing. Because it is important to correct for False Negatives when selecting important genes, where true significant genes are not discarded, Benjamini-Hochberg correction is a better choice. Benjamini-Hochberg trades off Type I error to correct for Type II error; as a result, more genes are considered significant.

D. Multinomial Logistic Regression with LASSO

Multinomial logistic regression with LASSO regularization is commonly used for feature selection. This regularization method drives insignificant features to have coefficients of zeros. This model takes in the entire dataset as input and fits labels that are one-hot encodings of the cancer types. For our purpose, *Saga* is the chosen solver as it supports the non-smooth $l1$ penalty, picks sparse multinomial coefficients and is suitable for very large datasets. Genes with large corresponding coefficients are deemed important.

E. Assemble Important Genes

As described above, each method identifies up to 100 of the most important genes, and the genes that lie on the intersection of these methods are chosen to be important genes. Then, the union of these important genes across cancer types is taken to form a new dataset whose rows correspond to important genes and columns correspond to patient sample IDs.

V. CLASSIFICATION & PREDICTION METHODS

After assembling a set of important genes, we perform cancer type classification and mortality prediction. A baseline model and an improved model are proposed for each objective.

For cancer type classification, we propose these models:

1) baseline: logistic regression with $l2$ regularization
2) improved: 2D hybrid convolutional neural network

For mortality prediction, we propose these models:

1) baseline: linear regression with ridge regularization
2) improved: 2D hybrid convolutional neural network

For each model proposed, comparisons are drawn between performance on the original dataset and important gene subsets. Accuracy is used to evaluate our classification models and root-mean-square-error (RMSE) is used to evaluate our prediction models.

A. Logistic Regression with $l2$ Regularization

The input data is a gene expression matrix and a list of one-hot encoding labels for corresponding cancer types. The solver for this regression model is also *Saga* and the evaluation metric is accuracy. The only difference is that this method uses $l2$ instead of $l1$ regularization.

B. Neural Network for Cancer Type Classification

A 2D hybrid convolutional neural network, adopted from Mostavi et al, is chosen as the improved model to classify cancer types. The classification, according to the paper, reached an average accuracy as high as 95.0% for 33 different cancer types. Given that our objective is to classify among 5 different cancers only, we believe we can reach an even higher classification accuracy.

Their neural network reshapes every 1D vector input, the array of gene expressions for a patient sample, to a 2D matrix. Each input array is padded with zeros at the end to make the array size divisible by the kernel size. Then two convolutional kernels are separately applied to the matrix, one for columns and the other for rows. The outputs of these individual convolutional kernels go through a pooling layer prior to being flattened. The model concatenates the two flattened 1D vectors to form one large 1D vector. The large 1D vector goes through a dense layer and then the output layer with softmax activation[12].

Given that we have fewer cancer types to classify, the original model described by the paper can potentially overfit. Therefore, we added a dropout layer to help regularize overfitting to imbalanced data. Our specific architecture is shown in Fig. 3.

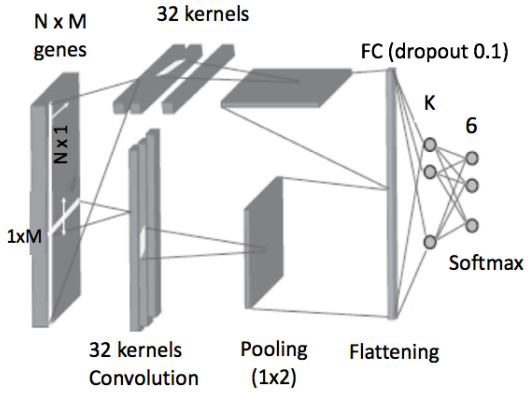

Fig. 3. 2D Hybrid Convolutional Neural Network with Dropout and the last layer with softmax activation function

Dataset	N	M	K
All genes	100	204	128
55 genes	10	61	55

TABLE II

2D HYBRID NEURAL NETWORK PARAMETERS FOR CLASSIFICATION AND PREDICTION TASKS ON ALL GENES & 55 GENES

For the important gene subset, we repeat gene expression data multiple times (in this case, 11, which results in 55x11 = 605 gene expression data). Since we believe individual gene in the important gene subset carries a great amount of information in characterizing cancer types, we do not want to lose much information from convolution filtering. Thus, repeatedly stacking genes help preserve gene information in the presence of filtering. This stacked input is then padded with zeros and goes through the same architecture as described above. The only difference between the model for all genes and the model for the selected 55 genes is at the last dense layer where we have 128 hidden nodes in one and 55 hidden nodes in the other. The architecture details for the original high-dimensional dataset and the 55 selected genes are specified in Table II.

C. Linear Regression With Ridge Regularization

To understand the correlation between mortality and gene expressions, we focus on predicting days-to-death given gene expressions of those who have passed away. Since we are outputting a numerical value, a reasonable model would be a linear regression model. We hypothesize that there is a linear relationship between mortality and gene expression levels. In this case, RMSE is used to evaluate our prediction models.

D. Neural Network for Mortality Prediction

We propose using a similar 2D hybrid convolutional neural network to predict mortality, as shown in Fig. 4. A slight modification here is to use *linear* activation function instead of *softmax* activation function at the output layer. Because the model is supposed to output only one number, using a *linear* activation function creates an output signal that is proportional to its input.

Because we would like to draw a fair comparison of the performances using all genes versus using only 55 selected genes, we use the same architecture for mortality prediction with minimal modifications to avoid the confounding effects. The exact same method of padding with zeros and reshaping into a 2D matrix is also applied to be consistent across.

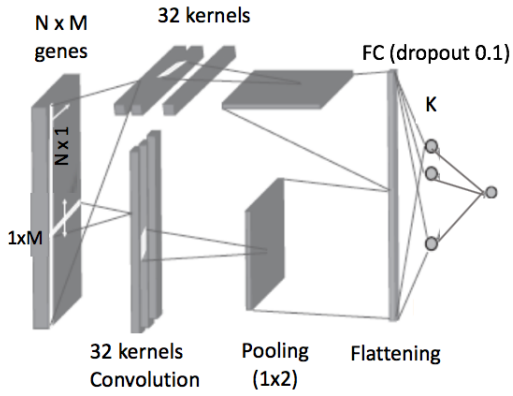

Fig. 4. 2D Hybrid Convolutional Neural Network for Mortality Prediction with dropout and last layer with linear activation function.

VI. RESULTS

A. Gene Selection

As suggested in the methods section, each gene selection method selects up to 100 significant genes. The intersection

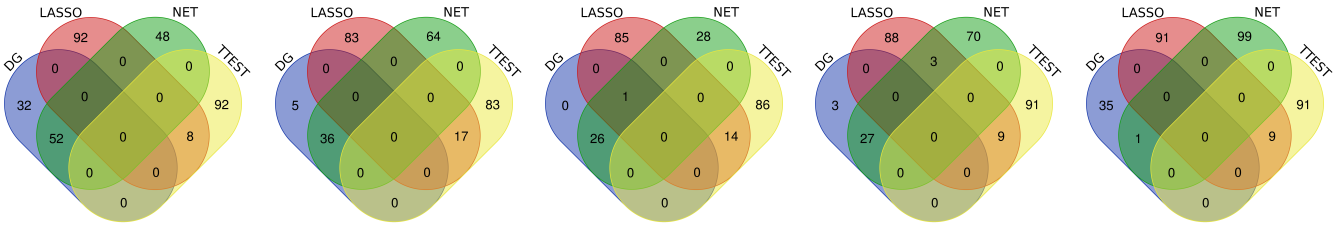

Fig. 5. Gene Intersection from each Method for Cancer Types in Order: BRCA, COAD, LIHC, STAD, THCA

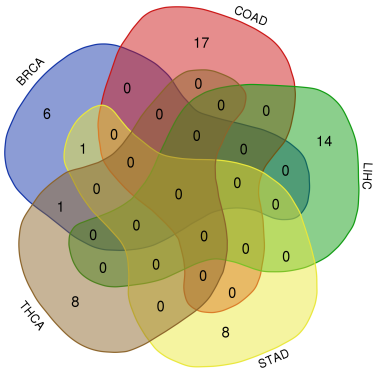

Fig. 6. Important Genes across Cancer Types

of the genes found by each method is shown in Fig. 5. The results are, from left to right, for BRCA, COAD, LIHC, STAD, and THCA cancers. The methods are: Identified Driver Genes (DG), Gene Co-expression Network Model (NET), Multiple Hypothesis Testing with Correction (TTEST), and Multinomial Logistic Regression with LASSO (LASSO).

All important genes across cancer types are plotted to look for any intersections. A total of 55 genes is plotted and the intersection result is shown in Fig. 6.

B. Gene Property Distribution

We believe a gene's property will provide us with an explanation on why, in the context of biology, a gene is important. In particular, we focused on the distribution of

1) detection,
2) protein location,
3) protein class, and
4) cancer specificity.

Detection focuses on the extent to which a gene is expressed in cancerous cells. When a gene is detected in more than one, it means the expression is present in more than one type of cancer.

Protein location corresponds to the location of the specific protein that is transcribed by the particular gene of interest. These proteins can be found inside the cell (intracellular), on the plasma membrane, or outside the cell (extracellular).

Protein class corresponds to the type of protein transcribed by the particular gene of interest. Majority of the protein types are plasma membrane proteins, disease-related proteins, or FDA approved drug targets.

Cancer specificity corresponds to how specific the elevated gene expression is to other cancer types. Not detected means that gene expression is not significant in all cancer types. Low specificity means an expression level that is slightly elevated in at least one cell type. Enrich means that the expression level of a particular cell type is at least four times as elevated. Enhanced means that the expression level of a group of cells is at least four times as elevated.

These distributions are plotted for the 55 important genes, as shown in Fig. 7. The results correspond to, from top to bottom, *detection*, *protein location*, *protein class*, and *cancer specificity*.

C. Cancer Type Classification

With the important genes identified, we can compare our classification models by their performance on the original dataset and on the important gene subset. The test accuracy for the models is shown in Fig. 8.

D. Mortality Prediction

Similarly, we compare our prediction models by their performance on the original dataset and on the important gene subset. In this case, the RMSE, instead of accuracy, is shown in Fig. 9.

VII. DISCUSSION

A. Gene Selection

The Venn diagrams shown in Fig. 5 - gene intersection from each method - suggest large overlaps between DG and NET. This makes intuitive sense as NET is built on DG. In order for genes to be potentially important as evaluated by NET, it must co-express with at least one driver gene. On the other hand, there is very little overlap between genes identified via other methods. This is because different algorithms compute significant genes differently, and therefore, find different sets of important genes.

We note that there are few overlaps between DG and LASSO or TTEST. This suggests that driver genes themselves are not necessarily too differentially expressed but can mutate and cause other genes that are correlated with driver genes to differentially express (e.g. a gene that is in the same pathway or located downstream of a driver gene). Thus, only genes that lie on the intersection of genes found by LASSO and TTEST for different cancer types are further studied.

A Venn diagram is used to demonstrate the intersection of the genes found in various cancer types, as shown in

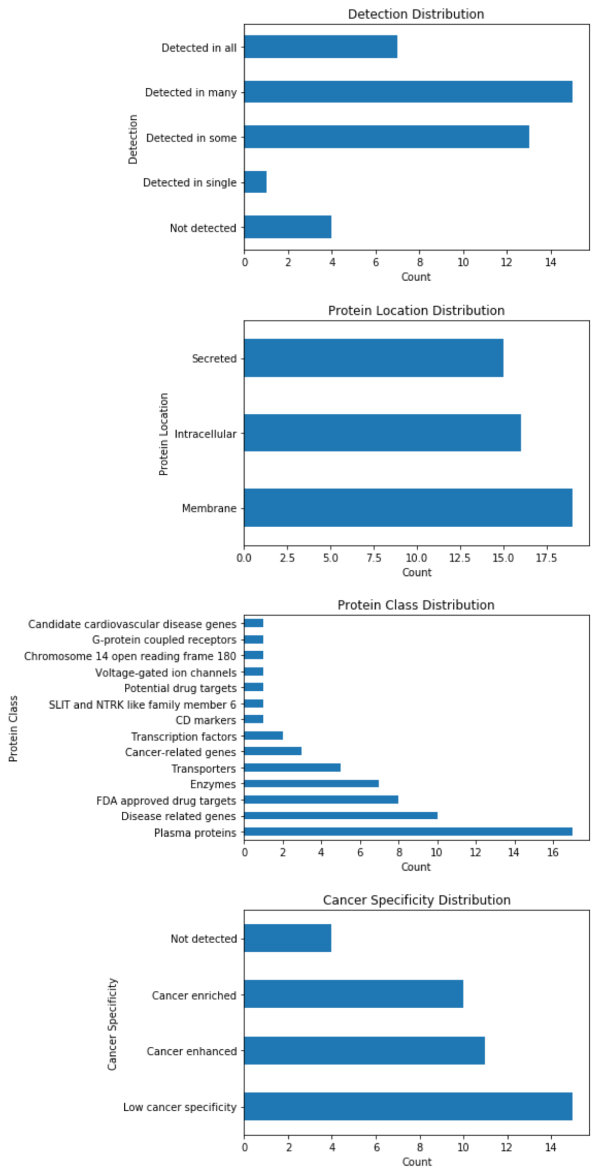

Fig. 7. Gene Property Distribution

Fig. 6. There is almost no overlap on genes found in different cancers which suggests that different cancers have significantly different genes that are important. The union of these genes (our final 55 genes) is selected to classify cancer types and predict mortality.

B. Cancer Type Classification

Based on the t-SNE result and our feature selection analysis, cancers are differentiable. Thus, Logistic regression can classify well for both conditions (97.8% with all genes and 96.5% with 55 selected genes). Our 2D hybrid neural network performs slightly better (98.0% with all genes and 97.5% with 55 selected genes). The performance with 55 selected genes only drop $\sim$1% for both models which confirms that these 55 selected genes have good predictive value in determining cancer types.

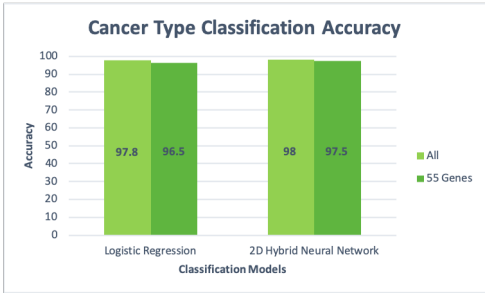

Fig. 8. Cancer Type Classification Models Accuracy Comparison

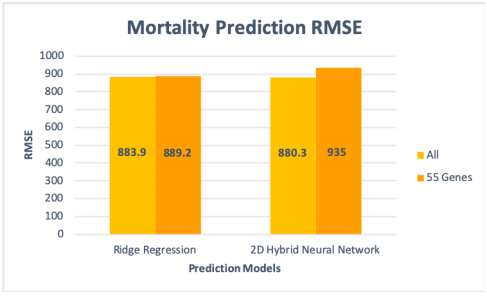

Fig. 9. Mortality Prediction Models RMSE Comparison

C. Days-to-death Prediction

Because there is a small number of patients for each cancer with mortality information, we predict days-to-death for all cancers combined (a total of 512 samples). The RMSE is high for both the linear regression model with ridge regularization and the 2D hybrid convolutional neural network model. This is due to a lack of sufficient training data and differences in death rates for different cancers.

Judging by the standard deviation - 2.5 years - of the mortality data, it is not unreasonable to get large RMSE values. Because the standard deviation is already large among patients, it is very hard for models to capture any significance. Furthermore, we use Pearson's correlation coefficient to visualize the correlation between each gene expression and mortality, as shown in Fig. 10.

The correlation coefficient distributions based on the entire dataset and our subset of important genes suggest that most genes are not correlated with mortality. Genes are correlated with mortality with a correlation coefficient of magnitude no greater than 0.5, which explains our high prediction RMSE.

D. Selected-Gene-Information

The majority of the selected genes have a differential expression detectable among cancers, which confirms that they are indeed important. The locations of the protein associated with genes are evenly distributed amongst three categories. These genes are mainly associated with three classes: FDA approved drug targets, disease-related genes and plasma protein (helps protect against apoptosis). A majority of the genes' normalized expression value is significantly elevated, enriched, or enhanced among cancer cells.

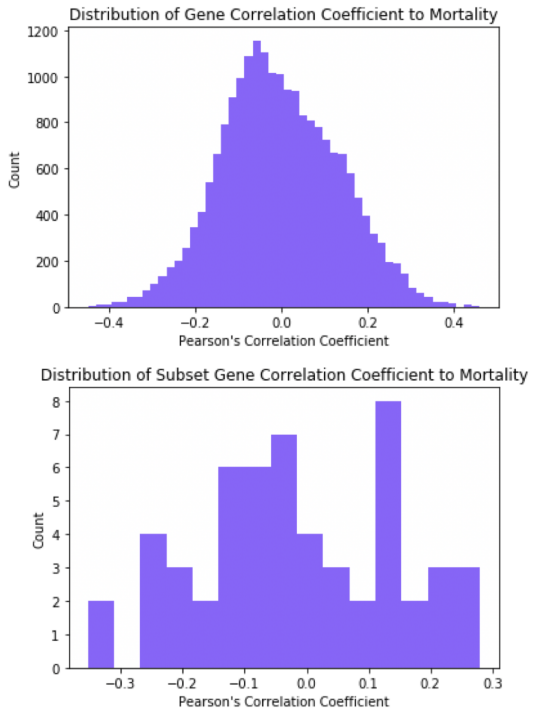

Fig. 10. Distribution of Correlation between Gene Expression and Mortality

E. Biological Explanation

In normal cells, one way that driver genes regulate cell division is through the expression of plasma membrane proteins. These proteins are responsible for cell-cell signaling, detection of environmental changes, transport of substances across the membrane, cell shape, and cell size[13]. In cancer cells, driver genes are mutated and results in the overexpression of plasma membrane proteins. This enhances a cell's ability to acquire nutrients from the surrounding so that it can divide more rapidly than usual (hyperplasia) while ignoring nearby cell's signaling for apoptosis[14]. Moreover, this causes the cell shape to change form (dysplasia) and then detach from their native environment to metastasize at distant sites[15]. This postulates that 1) driver genes in cancer cells are mutated, but not overexpressed, leading to changes in DNA methylation and 2) overexpression of plasma membrane proteins leads to the observed phenotype in cancer cells[16]. This aligns with our observation from the logistic regression and independent t-test since the majority of the selected features are proteins that are differentially expressed from the plasma membrane proteins class. Moreover, the lack of overlap between DG and both LASSO and TTEST suggests that driver genes are not significantly differentially expressed[17]. Cancer cells have driver genes that normally express but are mutated. Therefore, these mutated genes propagate incorrect information down signaling pathways, where its effect is amplified through a network of enzymes. Membrane proteins are then differentially expressed[18]. Finally, the lack of overlapping selected genes across our studied cancer types suggest that each cancer cell has unique driver genes[19].

F. Limitations

Ideally, we hoped to construct our co-expression network using the entire gene expression dataset but were not able to do so due to computational limitations. Therefore, we generated our network around driver genes assuming that genes that co-express with driver genes are critical to cancer development. If we could generate the network without this assumption, the genes selected should overlap more with those found by LASSO and TTEST.

For mortality prediction, we ignore still-alive-patients ($\sim$80% of all data), which assumes all patients with cancer will pass away within a certain time frame. A better model should consider both days-to-death and survival rates.

G. Future Works

It may be interesting to analyze the important genes for each cancer individually. This would reveal whether or not different cancers have a different distribution of important proteins.

For this project, we aim to classify samples by a certain type of cancer when the cancer has already developed. A potential next step is to gather sample data of patients before and after their cancer diagnosis. With this data, we can assess whether gene expression can forecast the onset of cancer.

Moreover, we hope to perform classification on subgroups of a particular cancer type (e.g. luminal-A, luminal-B, and HER2-enriched for breast cancer). This result will have more clinical significance because we often know the part of the body the cancer cells come from during biopsy, but not the subgroup that the cancer cell belongs to.

VIII. CONCLUSIONS

Informative genes, selected by LASSO and TTEST, vary greatly across cancer types. A subset of 55 genes is identified to be sufficient for classifying cancer types. Functional annotation of this subset is consistent with current understanding of cancer progression mechanisms, and, therefore suggests these computationally-found genes also have a biological significance.

In terms of mortality analysis, we found that gene expression data alone is not sufficient to predict the days-to-death. More information such as clinical data and other vital metrics may be more useful for this kind of analysis.

We were not able to find new driver genes since mutated driver genes, which cause cancer, do not necessarily differentially express. This means that the expression levels of these genes in cancer cells are no different than those of normal cells.

ACKNOWLEDGMENT

L. W. W. B thank the staff members of the MIT course *6.439 Statistics Computation and Applications* for their guidance. Special thanks go to Nathan Hunt for his suggestions and help on this project.

REFERENCES

[1] Chang, K., Creighton, C., Davis, C. et al. *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet 45, 1113–1120 (2013) doi:10.1038/ng.2764

[2] Zhang, L, et al. *Gene Expression Profiles in Normal and Cancer Cells.* Science (New York, N.Y.), U.S. National Library of Medicine, 23 May 1997, www.ncbi.nlm.nih.gov/pubmed/9157888.

[3] Bao, Ting, and Nancy E Davidson. *Gene Expression Profiling of Breast Cancer.* Advances in Surgery, U.S. National Library of Medicine, 2008, www.ncbi.nlm.nih.gov/pmc/articles/PMC2775529/.

[4] Ismail, Rubina S., et al. *Differential Gene Expression between Normal and Tumor-Derived Ovarian Epithelial Cells.* Cancer Research, American Association for Cancer Research, 1 Dec. 2000, cancerres.aacrjournals.org/content/60/23/6744.

[5] Narrandes, Shavira, and Wayne Xu. *Gene Expression Detection Assay for Cancer Clinical Use.* Journal of Cancer vol. 9,13 2249-2265. 5 Jun. 2018, doi:10.7150/jca.24744

[6] Sweeney, Timothy E, et al. *Mortality Prediction in Sepsis via Gene Expression Analysis: a Community Approach.* BioRxiv, Cold Spring Harbor Laboratory, 1 Jan. 2016

[7] Futreal, P Andrew et al. *A census of human cancer genes.* Nature reviews. Cancer vol. 4,3 (2004): 177-83. doi:10.1038/nrc1299

[8] Heng, Yujing Jan, et al. *Whole Blood Gene Expression Profile Associated with Spontaneous Preterm Birth in Women with Threatened Preterm Labor.* PloS One, Public Library of Science, 14 May 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4020779/.

[9] Pita-Juárez, Yered, et al. *The Pathway Coexpression Network: Revealing Pathway Relationships.* PLOS Computational Biology, Public Library of Science, 19 Mar. 2018

[10] Zhou, Xiaobo, et al. *Cancer Classification and Prediction Using Logistic Regression with Bayesian Gene Selection.* Journal of Biomedical Informatics, Academic Press, 11 Sept. 2004

[11] Gonzalez-Perez, Abel, et al. *IntOGen-Mutations Identifies Cancer Drivers across Tumor Types.* Nature News, Nature Publishing Group, 15 Sept. 2013, www.nature.com/articles/nmeth.2642.

[12] Mostavi, Milad, et al. *[PDF] Convolutional Neural Network Models for Cancer Type Prediction Based on Gene Expression: Semantic Scholar.* Undefined, 1 Jan. 1970

[13] lund-neth larsen, jensen ditzel. *Efficient isolation and quantitative proteomic analysis of cancer cell plasma membrane proteins for identification of metastasis-associated cell surface markers* Journal of proteome research, 1 Jun 2009

[14] Wang, J., Kondo, T., Yamane, T. et al. Expression of nuclear membrane proteins in normal, hyperplastic, and neoplastic thyroid epithelial cells. Virchows Arch 467, 427–436 (2015) doi:10.1007/s00428-015-1816-6

[15] Dobrzyńska, Izabela et al. "Characterization of human bladder cell membrane during cancer transformation." The Journal of membrane biology vol. 248,2 (2015): 301-7. doi:10.1007/s00232-015-9770-4

[16] Youn, A., Kim, K., Rabadan, R. et al. A pan-cancer analysis of driver gene mutations, DNA methylation and gene expressions reveals that chromatin remodeling is a major mechanism inducing global changes in cancer epigenomes. BMC Med Genomics 11, 98 (2018) doi:10.1186/s12920-018-0425-z

[17] Zhang, Junhua, and Shihua Zhang. "Discovery of cancer common and specific driver gene sets." Nucleic acids research vol. 45,10 (2017): e86. doi:10.1093/nar/gkx089

[18] Nussinov, Ruth et al. "Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers." PLoS computational biology vol. 15,3 e1006658. 28 Mar. 2019, doi:10.1371/journal.pcbi.1006658

[19] Lu, X., Lu, J., Liao, B. et al. Driver pattern identification over the gene co-expression of drug response in ovarian cancer by integrating high throughput genomics data. Sci Rep 7, 16188 (2017) doi:10.1038/s41598-017-16286-5

[20] Parry, R M et al. "k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction." The pharmacogenomics journal vol. 10,4 (2010): 292-309. doi:10.1038/tpj.2010.56

[21] Vanitha, C. Devi Arockia, et al. "Gene Expression Data Classification Using Support Vector Machine and Mutual Information-Based Gene Selection." Procedia Computer Science, Elsevier, 17 May 2015.

[22] A. Azuaje. Interpretation of genome expression patterns: computational challenges and opportunities. IEEE Engineering in Medicine and Biology, 2000.

[23] Mount, David W et al. "Using logistic regression to improve the prognostic value of microarray gene expression data sets: application to early-stage squamous cell carcinoma of the lung and triple negative breast carcinoma." BMC medical genomics vol. 7 33. 10 Jun. 2014, doi:10.1186/1755-8794-7-33

[24] Danaee, Padideh et al. "A DEEP LEARNING APPROACH FOR CANCER DETECTION AND RELEVANT GENE IDENTIFICATION." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing vol. 22 (2017): 219-229. doi:10.1142/9789813207813-0022

[25] Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of Cancer: The Next Generation." Cell, Cell Press, 3 Mar. 2011.

[26] Azuaje, Francisco J. "Selecting biologically informative genes in co-expression networks with a centrality score." Biology direct vol. 9 12. 19 Jun. 2014, doi:10.1186/1745-6150-9-12

[27] Boyu Lyu and Anamul Haque. 2018. Deep Learning Based Tumor Type Classification Using Gene Expression Data. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '18). ACM, New York, NY, USA, 89-96. DOI: https://doi.org/10.1145/3233547.3233588