Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Shalini Sreedhar April 9, 2019

Predicting anticancer drug sensitivity from high dimensional genomic data

by

Shalini Sreedhar

Lee Cooper, PhD, MS
Adviser

Emory Biology Department

Lee Cooper, PhD, MS
Adviser

Gordon Berman, PhD

Committee Member

Shun Cheung, PhD

Committee Member

2019

Predicting anticancer drug sensitivity from high dimensional genomic data

By

Shalini Sreedhar

Lee Cooper

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Biology

2019

Abstract

Predicting anticancer drug sensitivity from high dimensional genomic data
By Shalini Sreedhar

Acute Myeloid Leukemia (AML) is a heterogeneous cancer with at least 11 genetic classes and more than 20 subsets. Due to the highly variable nature of the disease, there is a strong need for treatment based on individual's genetic composure. This type of precision medicine for AML is relatively new due to the recent decrease in cost and increase in efficiency of genetic sequencing. In this study, the primary dataset used in making these predictions is the BeatAML dataset which provides RNA sequencing, gene mutation, and drug sensitivity information for 451 cell line samples and 122 small molecule drugs. This dataset was preprocessed through standard scaling and dimensionality reduction through principle component analysis. A deep neural network model was created to make drug sensitivity predictions on the gene sequencing data. The problem was first formed as a regression problem in order to predict specific sensitivity values for each drug. The problem was then simplified to binary classification in order to attempt to improve the accuracy of the predictions. Five drugs were chosen as the focus and the sensitivity values were discretized into 2 categories (levels) of sensitivity. This resulted in a high training accuracy (average = 0.98) and a lower testing accuracy (average = 0.62). The importance of generalization, dimensionality reduction, and equal testing and training sets was emphasized as methods that are most important when dealing with datasets with small sample sizes and large feature sizes. Future studies regarding anticancer drug sensitivity predictions should focus on regularization techniques in order to improve test set prediction performance. Feature importance was evaluated as a method of determining the biological significance found in these models. Pathway analysis was performed for each drug on the genes having the most importance in predicting drug sensitivity. The strongest correlations between the most important features and the pathway targeted by the drug were found for the drugs trametinib and selumetinib. Further work needs to be done to interpret these networks in order to improve understanding on how predictions are being made and increase the likelihood of their adoption in industry.

Predicting anticancer drug sensitivity from high dimensional genomic data

By

Shalini Sreedhar

Lee Cooper

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Biology

2019

Table of Contents

1. Introduction *1*

1.1 Acute myeloid leukemia 1

1.2 Deep learning in genomics 2

1.3 Dataset 3

2 Methods *6*

2.1 Initial model 6

2.2 Generalization methods 7

2.3 Creating equal training and testing samples 9

2.4 Dealing with sparse dataset 10

2.5 Choosing specific drugs 10

2.6 Classification problem 11

3 Results *15*

3.0 Drug sensitivity correlations 15

3.0 Drug selection and pathway analysis 16

3.1 Regression neural network model 18

3.2 Classification neural network model 20

4 Discussion *26*

4.1 Conclusions 26

References *28*

List of Figures

Figure 1: Model complexity vs error (Fortmann-Roe, 2015) 8

Figure 2: Confusion matrix outline 13

Figure 3: Log loss (cross entropy) function 14

Figure 4: Drug sensitivity correlations for the most sensitive and least sensitive drugs 15

Figure 5: Training (left) and testing (right) predictions for model trained on 5 drugs 20

Figure 6: Confusion matrix plots for each drug tested, left shows training data results and right shows testing data results 23

Figure 7: Confusion matrix plots for each drug tested, left shows training data results and right shows testing data results 25

List of Tables

Table 1: Classification metrics calculations 13

Table 2: Variation and number of samples per top drugs, selected drugs are highlighted .. 16

Table 4: Mean squared error per drug (top 10) for regression model trained on 122 drugs 19

Table 5: Mean squared error per drug for regression model trained on 5 drugs 19

Table 6: Optimal classification parameters found through randomized grid search..... 21

Table 7: Training and testing accuracy from classification problem 21

1. Introduction

1.1 Acute myeloid leukemia

Acute myeloid leukemia (AML) is a type of cancer distinguished by the infiltration of bone marrow, blood, and other tissues by differentiated cells from the hematopoietic system (Dohner et al, 2015). The cure rate of the disease is highly variable depending on the age at which it arises and can be anywhere from 15 to 40% of patients. There are currently over 21,000 diagnoses reported annually in the United States, costing a total of $0.5 billion for patients <65 and $1.5 billion for patients >65 (Mahmoud et al, 2012).

AML is a largely heterogeneous disease, with at least 11 genetic classes and more than 20 subsets. Cytogenetic analysis of the genomes of annotated cases of de novo AML show that many of the mutations contributing to the pathogenesis of AML are undefined (Ley et al, 2013). Furthermore, there were over 2,000 somatically mutated genes observed in 200 cases of AML. These analyses show that the pathogenesis is largely caused by complex interactions of genetic events. Many of the most prominent mutations are also seen in myelodysplastic syndromes which are commonly precursors to secondary AML (Lundberg et al, 2014). In this study, many types of AML are analyzed. The classification of AML is disputed since, as discussed above, there are large numbers of complex mutations causing the disease. Further research into the genetic and epigenetic changes causing AML are required to develop a fully accurate classification system for the disease.

While there is largely a standard and unchanging practice for AML treatment, some therapies have been developed to target mutational events (Tyner 2018). One current predictive therapy targets patients with retinoic acid receptor rearrangement. Another therapy targets FLT3 mutational events for which tyrosine kinase inhibitors are used.

Other therapies inhibit epigenetic modifiers such as histone lysine methyltransferase through the mutation of these factors when combined with drugs. There has been evidence in trials of some of the above therapies that suggest that the response rate of certain agents depends on genetic factors in the patient. In one example, decitabine, a hypomethylating agent, was shown to have a >30% higher response rate in patients with an unfavorable-risk cytogenetic profile than in patients with a more favorable- risk profile (Welch et. al, 2016). Furthermore, patients with TP53 mutations had a 59% higher response rate. This evidence shows the importance of prescribing therapies based on the genetic composition of the patient.

1.2 Deep learning in genomics

This increase in predictive therapies is largely due to the massive amount of data produced by high through-put genetic sequencing. This allows for complex statistical and computational models to be applied to that data (Yue and Wang, 2018). There are many ways to process this data, from simplistic linear regression models to more complex tree and neural networks. Due to the flexibility and accuracy of deep learning models, they are emerging as a preferred method of analyzing genomic data. Some of the most common deep learning algorithms that are used in genomics are convolutional neural networks, recurrent neural networks, autoencoders, with each of them being suited for specific learning tasks. While deep learning models are exciting in their ability to process and make predictions from large amounts of data, there is a cost of increasing the complexity of the models.

The algorithms used in this study will be chosen through a literature review and an analysis of the dataset. Biological considerations of the data will be used to preprocess the

data in a way that allows the models to only take in significant data. After preprocessing, CNNs are be used in feature selection, since there are small differences in variance between the genomic features. After the initial model is created, it is optimized through hyperparameter tuning and other optimization techniques. Transfer learning and multitask learning techniques are also considered in order to predict the different drug responses concurrently.

Since the interpretability of the model is also important, biological implications of the results are examined. This is done by choosing some drugs that are shown to accurately predict drug sensitivity and looking deeply into common themes between the most useful predictors. In doing so, cell lines and other pathways that connect the genes and markers that are indicated as accurate predictors of sensitivity are investigated.

1.3 Dataset

The primary dataset, *Beat AML*, is specifically concerned with acute myeloid leukemia (Tyner et al, 2018). The dataset was recently released and is composed from 672 tumor specimens from 562 patients. These data were collected through whole-genome sequencing, RNA sequencing, and analyses of ex-vivo drug sensitivity. This data is most relevant to the posed problem since it gives detailed genetic and mutational data along with drug sensitivity measures for each sample.

From the large amount of data gathered in the study, three datasets were selected to be used in the analysis. The first was the exome sequencing data which came from 622 specimens collected from 531 patients. This included information on the exact locus, type of mutation, and mutation effect. Next, we examined the RNA sequencing data, which came

from sequencing performed on 411 patients, giving data concerning 451 specimens. These data were given in counts per million (CPM) and reads per kilo base million (RPKM). The original CPM was normalized to remove technical biases in sequenced data such as the gene length and depth of sequencing. RPKM specifically considers the gene length for normalization, which is important for single end RNA sequencing experiments.

The final data set gives drug sensitivity information. There were 122 small molecule inhibitors tested against the samples. However, not every sample was tested on each of these drugs. The data was gathered through an ex-vivo drug sensitivity assay. The values given were measured with IC50 and AUC. The IC50 gives a measure of the concentration of the inhibitor where the inhibition is reduced in half, while the AUC gives the area between the drug response curve and a certain fixed reference value. The AUC was chosen to be used in future analysis since it provided more complete information. The sensitivity of each drug to each sample was not measured - this led to problems later on in attempting to predict drug sensitivity for each drug.

The other dataset that will be examined is the Cancer Cell Line Encyclopedia (CCLE). This dataset contains large-scale genomic data for 947 human cancer cell lines coupled with pharmacologic profiling of 24 compounds across approximately half the observations. The data concerns 36 tumor types. While there are many datasets describing various genomic properties of the cell lines in the CCLE, specific characteristics will be chosen to be included in the analysis in order to get the best prediction. This includes DNA copy number, mutation data, protein expression data, and mRNA expression data. Drug sensitivity will be predicted using various measures corresponding to drug response curves. These specifically are the concentration at half- maximal activity (EC50), concentration of

inhibitor where inhibition is reduced by half (IC50), maximal level effect (Amax), and the area between the drug response curve and a fixed reference value (ActArea).

2 Methods

2.1 Initial model

The gene expression data was combined with the drug sensitivity data to align the samples with their corresponding expression and sensitivity. The expression data was standardized using the sklearn StandardScaler to transform the data to have a zero mean and unit standard deviation. This was necessary for the features not to be overrepresented from differing scales of expression for the various genes in the dataset.

This problem was initially framed as a regression problem in order to predict drug sensitivity values from gene expression data. A dense neural network parametrized by the number of layers, the number of nodes, the activation function, and the dropout fraction was created in order to determine the optimal model that could predict these data. A dense, fully connected, neural network is one in which each node is connected to each node of the previous and the next layer. Due to the complex nature of this model, the relationships learned may simply be due to sampling noise, which could lead to overfitting to the training data. A dropout layer parametrized by the dropout fraction was added in order to lessen the chance of overfitting. The dropout technique involves randomly dropping out several units and their connected units during each training pass (Srivastava, 2014). Dropout is not performed when computing predictions on test data, since we do not want to ignore any information in that situation.

When training the model, hyperparameter selection was performed through a randomized grid search. This was achieved by first creating a grid of all the possible parameter combinations using the sklearn ParameterGrid function. Next, the model was

run on the training and testing data with a set number of hyperparameter combinations, chosen randomly from the ParameterGrid. Finally, the hyperparameters with the greatest test score was chosen. This randomized search methodology was chosen since the parameter space was very large and it was infeasible to exhaustively search through every combination of possible parameters in order to determine the optimal model.

2.2 Generalization methods

Overfitting is a common problem in deep learning, since there are a larger number of layers, nodes, and, in this case, a large number of features. This is when the model produces highly accurate predictions on the training data but has much greater error when predicting on data that it has not previously seen. The initial results gave several indications that the model was being overfit to the training data.

The primary methods used to deal with overfitting in neural networks are dropout, regularization, early stopping, model complexity, and dimensionality reduction. Regularization is a technique that increases the affinity for creating a simpler model by reducing model coefficients. In L2, ridge, regularization, weights are penalized in proportion to square of value of the value of weight coefficients. This drives outlier weights to smaller values. In L1, lasso, regularization, weights are penalized in proportion to the absolute value of the magnitude of the coefficient. The primary difference between the two methods is that L1 regularization drives coefficient weights to 0, eliminating them entirely, while L2 regression simply decreases the impact of these features that are not contributing new information. Early stopping involves stopping the training at an iteration at which the model has not had time to overfit to the training data and has just finished learning most of

what it could from the training set. Decreasing the complexity of machine learning models is important due to the bias variance tradeoff. As shown in Figure 1 below, as model complexity increases, the total error increases with the variance of the model. Therefore, by decreasing the complexity of the model, we can get to the sweet spot where the bias and variance are both minimized.

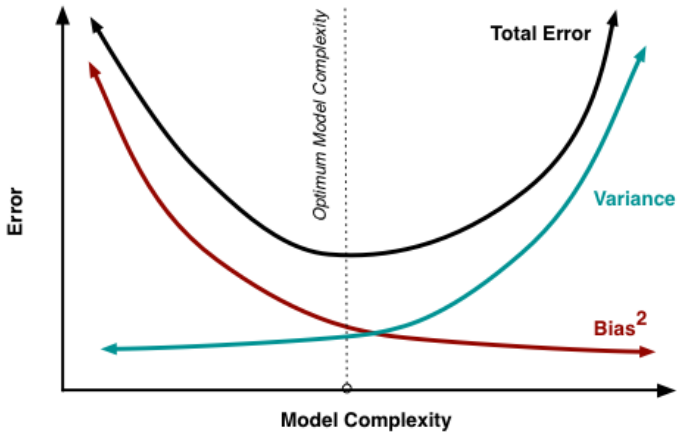


Figure 1: Model complexity vs error (Fortmann-Roe, 2015)

Dimensionality reduction is another method of reducing model complexity. Since the number of features in the dataset was very large (n=22853), with a small number of samples (n=451), the curse of dimensionality was a strong consideration. One way to decrease the impact of this factor was to perform dimensionality reduction on the feature set prior to feeding it into the deep neural network. Principal component analysis (PCA) was performed on the dataset using the sklearn PCA implementation. This implementation was a linear dimensionality reduction using singular value decomposition to project the data onto a lower dimension space. The number of components was chosen to be n=400. This number was based on experimentation looking for the greatest explained variance of the selected components.

2.3 Creating equal training and testing samples

Due to the extremely small number of samples (n=451) in the dataset, it was very important to choose training data that accurately represented the entire sample. This is because the model would learn the information presented and attempt to use this on the testing data. The model would therefore not be able to make accurate predictions on the types of gene expression combinations it had not previously seen. In order to confirm that the training data was accurately representing the entire population, two methods were used. First, the type of cancer was considered.

Furthermore, a weighted gene correlation network analysis (WGCNA) was used to cluster the genes. WGCNAs are commonly used to determine clusters of genes that are strongly co-expressed (Langfelder et al, 2008). These networks are commonly used in expression sets obtained from RNA-seq expression analysis, which are typically high dimensional datasets. The Python-wrapper iterativeWGCNA was used in this study since it is a package that minimizes information loss in determining gene clusters (Greenfest-Allen, 2017). This package works by first constructing a gene correlation network (GCN) from a weighted adjacency matrix describing the correlation between gene pairs. The adjacency adjusted for the proportion of shared connections is then calculated and then used in the hierarchical clustering. The clusters are determined by creating sections that maximize intra-connectedness among cluster genes. Since the dataset was extremely large, with 22,853 genes analyzed, the parameters put into this model were tuned to maximize the efficiency of the algorithm for the large set. The parameters of interest were the *minKME* and the *reassignThreshold*. The kME is a measure of module membership, determined by the correlation of the gene expression profile with the module eigengene of a given module.

The minKME determines the cut off at which a given gene is retained in a given module. The reassignThreshold determined the p-value ratio threshold at which genes would be reassigned between modules.

2.4 Dealing with sparse dataset

In exploring the drug sensitivity data, it was recognized that many of the samples were not tested for sensitivity on each of the 122 drugs in the study. Furthermore, there were no drugs with full responses. This introduced a significant problem since neural networks cannot deal with null values. In order to deal with this issue, two methods were used. First was filling the null values with mean sensitivity values for each drug. While this solution allowed the network to be trained, it introduced a class imbalance-like problem in which too many of the responses needed to be predicted to be the same value. The next solution was to selectively train the network on individual drugs and to create new datasets based on removing the null values from each drug's data. The second methodology was favored in order to reduce the bias from filling in new values.

2.5 Choosing specific drugs

Due to the difficulty in making and analyzing predictions on each of the 122 drugs presented in the original study, five drugs were selected to focus on. Two metrics were used to select the five drugs to focus on. These were the standard deviation and the number of samples. The standard deviation of the drug sensitivity of each of the samples was chosen as a metric as it would be important to use the data where there are large enough differences present in the sensitivity for the various drugs depending on the samples (i.e. genomic composition). The number of samples was also important since there were large

amounts of data missing, since not every sample was tested with each of the 122 drugs. The drugs chosen are listed in Table **2**, in the results section.

A smaller drug population to analyze was also helpful in determining any biological significance behind the sensitivity predictions. After selecting the five drugs to focus the proceeding analysis on, analysis on the mechanism of action and targeted pathway of each drug was performed. To do so, a literature search on any pathways targeted by each drug was first noted. Next, the gene ontology database was used to determine the specific genes affected by each noted pathway. Any genes that were in the pathway and in the feature set examined were noted and the feature rank of these genes was examined.

To calculate scores of how much each gene expression feature had on the final prediction, feature scores were calculated for each gene for each drug tested. This was calculated as the gradient of the training loss with respect to the training features. This gave feature scores for each cell line and gene combination. In order to get the feature score per gene, gene scores were averaged across each cell line. The top gene features were ranked and compared to the genes in each targeted pathway for each drug.

2.6 Classification problem

After drug sensitivity predictions were made as a regression problem, the results were analyzed and the accuracy was measured. While the training accuracy was very high, the testing accuracy was comparably lower. A variety of possible explanations for these results were considered. First was overfitting, which was addressed in earlier sections. Second, the training set was not representative of the overall population, leading to different populations being given in the test set. This issue was also addressed above.

Finally, it was considered that there were not enough training samples to make sufficient predictions on the exact drug sensitivity values. After going back to the original mission of this project, to aid in the creation of precision treatment for AML patients based on their specific genomic composition, it was determined that it may be possible to simplify the problem by considering relative sensitivity instead of exact sensitivity values.

To do so, the drug sensitivity values were discretized into 10 measures of sensitivity, from 0-9. Zero would represent the lowest level of sensitivity, while 10 would be the highest level of sensitivity. Two methods of discretization were considered. First was creating equally sized bins, resulting in normally distributed buckets. Second was creating buckets based on the percentile of each value, resulting in buckets each containing around the same number of samples. The second method was preferred since this would eliminate the issue of class imbalance while making predictions. Furthermore, this reflects the fact that we will be predicting the relative sensitivity to each drug. It also meant that the predictions would be on the same scale for each drug prediction but would still be accurate measures of how different gene compositions influenced the level of drug sensitivity for each drug. After discretizing the data, the neural network from above was adapted to a classification problem. An argmax TensorFlow layer was added to the graph to determine the class prediction from the logits (raw unnormalized probabilities) outputted by the final dense layer of the network. Next, the loss and accuracy calculations were considered.

In order to evaluate the accuracy of the model, multiple methods of accuracy calculation were considered. The most commonly used metrics for multi-class accuracy are

F1 score, average accuracy, and log loss. These metrics are based off of the confusion matrix (Figure 2), which summarizes the performance of a classification algorithm.

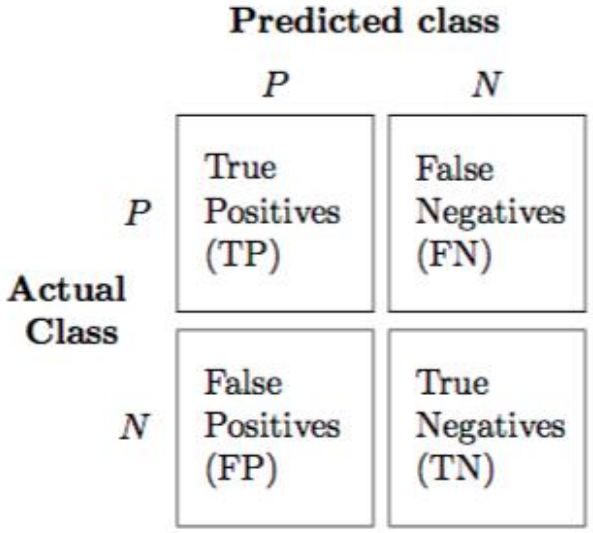


Figure 2: Confusion matrix outline

The metrics used are described in detail below (Sokolova et al, 2009).

Metric	Description	Calculation
Precision	Number of items correctly identified as positive out of total items identified as positive	$\frac{TP}{TP + FP}$
Recall	Number of items correctly identified as positive out of total true positives	$\frac{TP}{TP + FN}$
F1 Score	Harmonic mean of precision and recall	$\frac{2 * Precision * Recall}{Precision + Recall}$
Average Accuracy	Average per-class effectiveness	$\frac{TP + TN}{TP + FP + TN + FN}$
Log Loss	Loss that incorporates probabilistic confidence, based on the distance between the prediction and actual value	$CE = -\sum_{i=1}^{C'=2} t_i log(f(s_i))$

Table 1: Classification metrics calculations

The cross entropy, or log loss, function, is typically used in classification problems to estimate the distance between the actual and predicted values. Figure 3 shows that as the predicted probability of the true class gets closer to zero, the loss increases exponentially

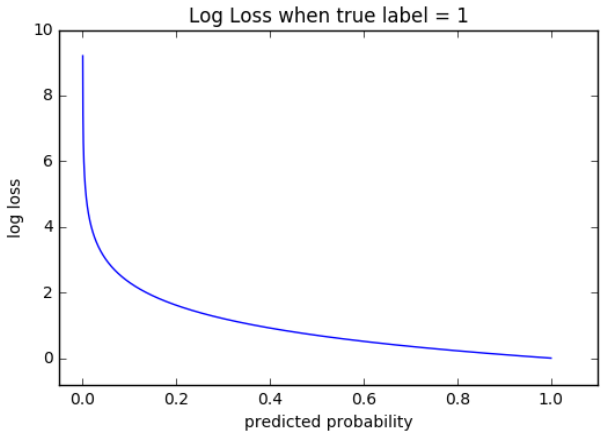


Figure 3: Log loss (cross entropy) function

When training the model and finding the optimal hyperparameters, the F1 score, average accuracy, and log loss were all calculated in order to score each model. The F1 score and the average accuracy were computed using the sklearn metrics package f1_score and accuracy functions, respectively. The log loss was computed with the tensorflow sparse_softmax_cross_entropy so that it could be used in the optimization of the model. The sparse loss function was chosen because the labels were encoded as integers and were not one hot encoded. Softmax was used because this is a multiclass prediction problem (n_classes = 10).

3 Results

3.0 Drug sensitivity correlations

Figure 4 shows the drug correlations among top and bottom 20th percentile drugs in terms of sensitivity. It is important to note the interactions between drugs that may cause similar responses. It can be seen that there are some blocks of similar colors, which show that drugs of similar sensitivities may interact with similar drugs.

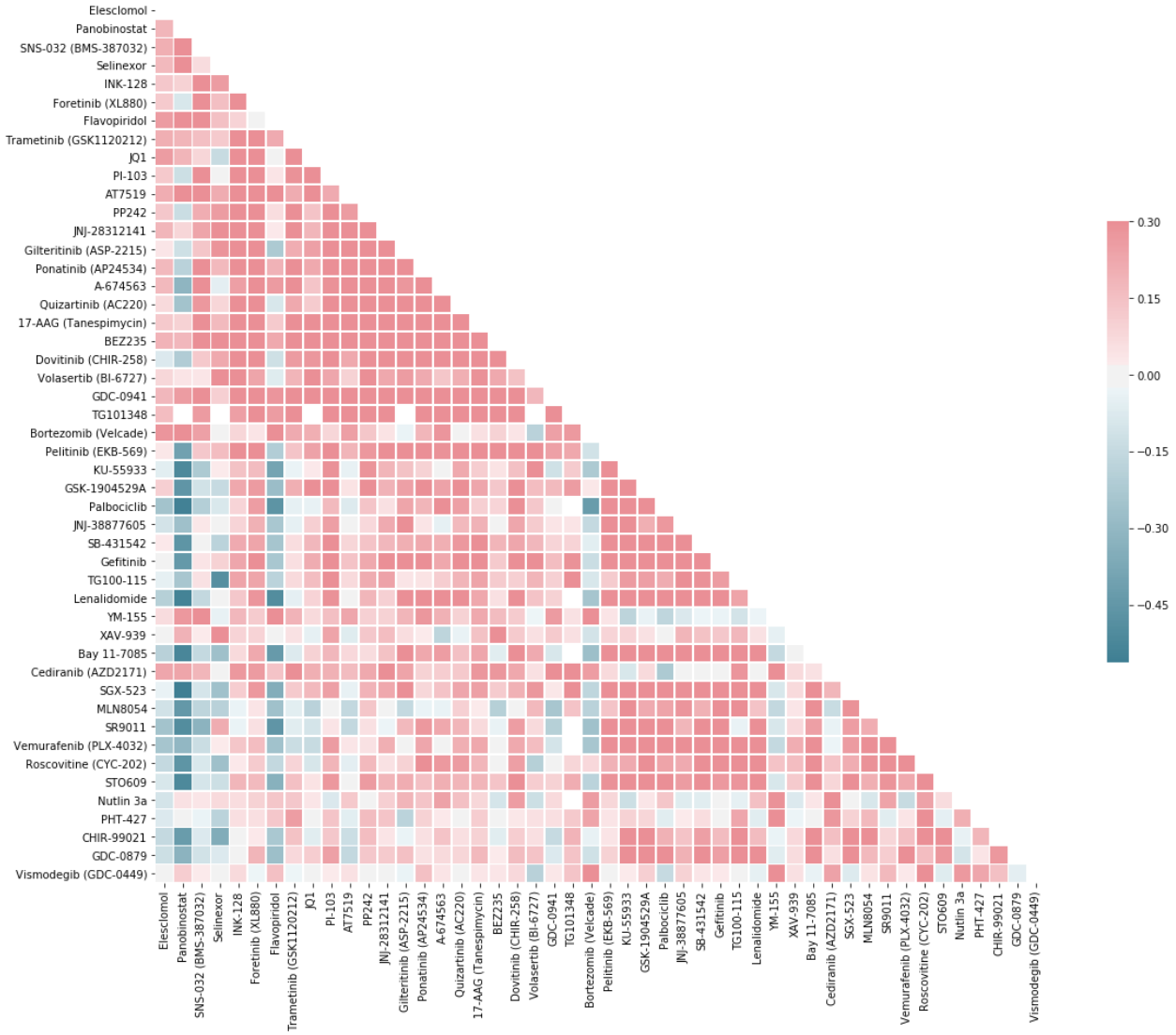


Figure 4: Drug sensitivity correlations for the most sensitive and least sensitive drugs

3.0 Drug selection and pathway analysis

The drugs selected are shown below, selected per the methodology described in section 2.5 above. The balance between high standard deviation and number of samples was considered. Since there was not much variance in the standard deviations, the number of samples was considered more strongly. This is because of the importance of having sufficient training data to derive accurate predictions.

Drug	Standard Deviation	Number Samples
Venetoclax	71.091694	186
Panobinostat	69.893501	128
Trametinib (GSK1120212)	64.854224	299
Selumetinib (AZD6244)	63.830559	287
Tivozanib (AV-951)	61.188227	284
JNJ-28312141	59.888794	279
KI20227	58.654778	284
Dasatinib	58.592302	311
Flavopiridol	58.503910	288
PD173955	56.788491	292
Staurosporine	56.458437	86
Bortezomib (Velcade)	56.075906	294
Doramapimod (BIRB 796)	55.667784	287
Elesclomol	55.655424	272
Selinexor	55.147435	76
MK-2206	54.213600	284

Table 2: Variation and number of samples per top drugs, selected drugs are highlighted

The specific method of action for each of the chosen drugs is detailed below.

Trametinib

Trametinib is a highly specific allosteric MEK1/2 inhibitor (Salama 2013). It works by inhibiting the catalytic activity of MEK through selective phosphorylation inhibition. This results in a monophosphorylated protein, rather than the dual phosphorylation required for normal MEK activity. MEK is an important part of the mitogen-activated protein kinase (MAPK) pathway. This pathway transmits signals from activated cell surface

receptors to intracellular effectors. Since it regulates many growth factor signaling processes, the pathway plays a role in many times of cancer. Inhibition of MEK causes disfunction in the MAPK pathway, which has been linked to various forms of cancer.

Through the Gene Ontology database, genes affecting the MAPK pathway were determined. In total, there were 624 MAPK genes that had measured expression values in the BeatAML study. These values were compared to the feature scores computed in the drug prediction regression problem to determine if the neural network was learning any biological pathways. There were 64 genes in the MAPK pathway that were also in the top 300 features.

Selumetinib

Selumetinib, similar to trametinib, is a small molecule inhibitor of MEK1/2. It works by inhibiting phosphorylation of ERK1 and ERK2. It works particularly well when cells contain BRAF or KRAS mutations. It targets the RAS-mediated signal transduction part of the MAPK pathway. Through the gene ontology database, genes that matched the selumetinib pathway were determined. There were 51 genes in the pathway that were also in the top 300 features in selumetinib sensitivity prediction.

Tivozanib

Tivozanib is part of the receptor type tyrosine kinase (RTK) supergene family and vascular endothelial growth factor receptor (VEGFR) family (Shibuya 2013). It is an ABC transporter that is a potent inhibitor of VEGF 1/2/3 receptors (Yang 2014). This drug works by acting on the ATP-binding cassette transporters, which causes multidrug

resistance by pumping out a variety of drugs out of the cells at the expense of ATP hydrolysis.

In the gene ontology pathway, 111 genes were found in the ABC transporter pathway. 84 of these genes were matched to those in the feature set, and 16 of these were in the top 500 features.

3.1 Regression neural network model

The optimal parameters found from a randomized grid search are shown in Table 3. The parameters for the model were chosen by examining the best performance based on the training and the testing data. It was noted that the learning rate chosen was usually quite fast for passes that had a higher test accuracy, while the learning rate for passes with a higher training accuracy had a slower learning rate. Furthermore, as per the curse of dimensionality, as less complex models with an increased dropout rate, fewer nodes, and fewer layers tended to perform better on the testing data as well.

Parameter	Value
Dropout Rate	0.6
Number of Nodes	100
Learning Rate	0.01
Activation Function	Relu
Number Layers	7
Regularization Rate	0.01

Table 3: Optimal regression hyperparameters found through randomized grid search

The network was trained using the optimal hyperparameters and the mean squared error was computed per each drug tested. The network was optimized and trained on a model of 122 drugs and a model of 5 drugs, as explained above. The training mean squared errors per drug are shown for each of the models in Table 4 and Table 5. Only the best 10

mean squared errors are shown for the initial model, for brevity. This shows the improvement in the mean squared error when training and optimizing on a smaller subset of the drugs presented in the study. The corresponding plots of predictions vs true labels are shown in Figure 5 for both the training and the testing datasets. This shows the difference between the training and testing accuracies.

Drug	MSE
TG101348	766.182556
Lestaurtinib (CEP-701)	824.458008
BI-2536	843.002625
Entrectinib	1060.949707
Selinexor	1086.696655
Volasertib (BI-6727)	1093.005859
AT7519	1126.944336
Go6976	1174.020508
Afatinib (BIBW-2992)	1182.321899
SNS-032 (BMS-387032)	1190.853760

Table 4: Mean squared error per drug (top 10) for regression model trained on 122 drugs

Drug	mse_train
Trametinib	681.903320
Selumetinib	731.857910
Tivozanib (AV-951)	578.105469
KI20227	942.601807
Dasatinib	747.132568

Table 5: Mean squared error per drug for regression model trained on 5 drugs

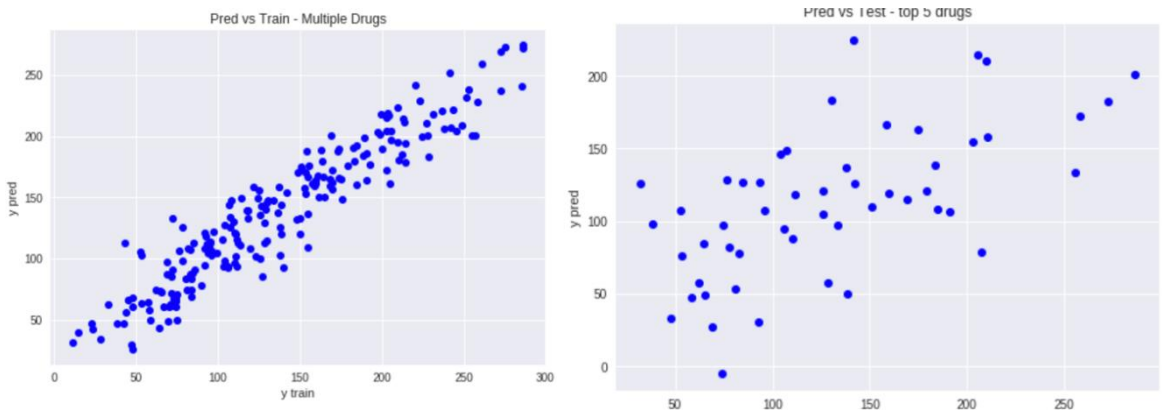


Figure 5: Training (left) and testing (right) predictions for model trained on 5 drugs

3.2 Classification neural network model

The optimal parameters found from a randomized grid search are shown in Table 6. These parameters were chosen by examining the confusion matrix from the results of training on each of the parameters with the highest training accuracy. The parameters with the best confusion matrix on the testing data were chosen. The matrices are shown in Figure 1.

Parameter	Value
Dropout Rate	0.4
Number of Nodes	100
Learning Rate	0.5
Activation Function	Tanh
Number Layers	5
Regularization Rate	0.01

Table 6: Optimal classification parameters found through randomized grid search

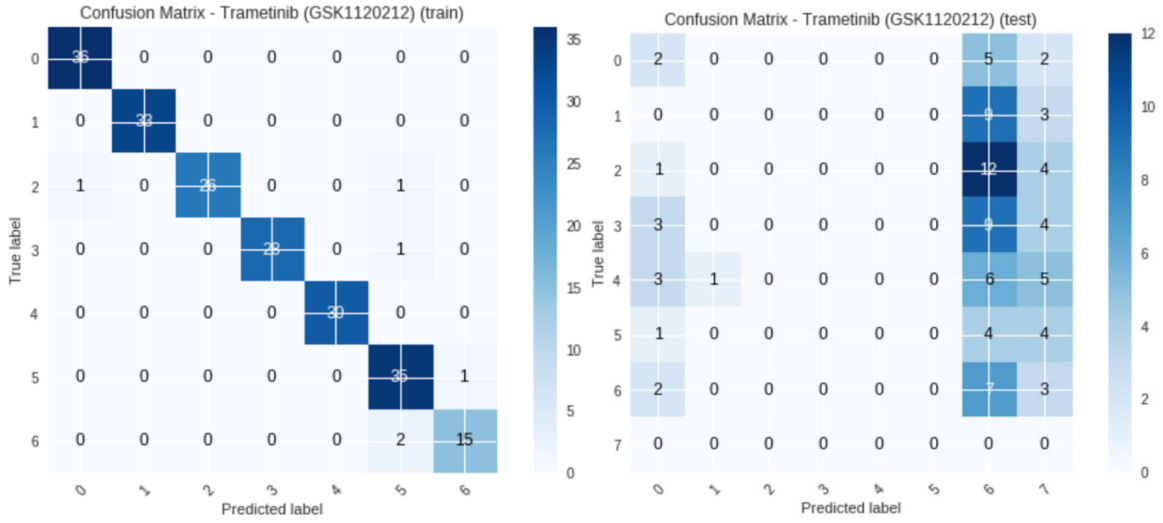
The accuracy of the classification model after being trained on the optimal parameters is shown in Table 7. The training accuracy is high, with the lowest accuracy being 0.952. The testing accuracy, however, is still low, with the highest being 0.14. Since there are 10 classes, the testing accuracy is just slightly higher than random on average.

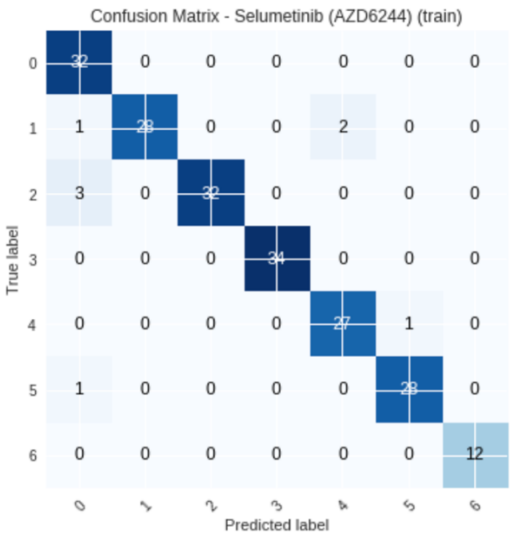
The accuracies are similar for each of the five drugs tested, although the hyperparameter optimization was only performed on a single drug.

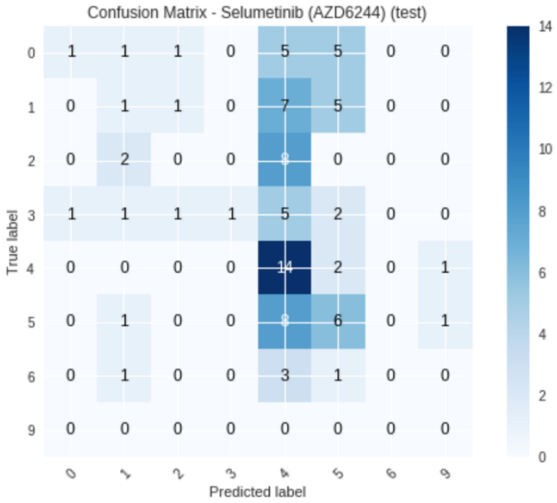
Drug	Avg Train Accuracy	Avg Test Accuracy
Trametinib	0.9904	0.0778
Selumetinib	0.9901	0.1279
Tivozanib (AV-951)	0.99	0.0941
KI20227	0.9849	0.1059
Dasatinib	0.9954	0.172

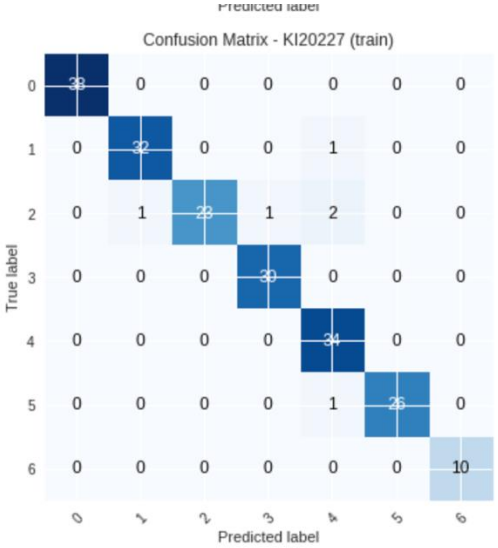
Table 7: Training and testing accuracy from classification problem

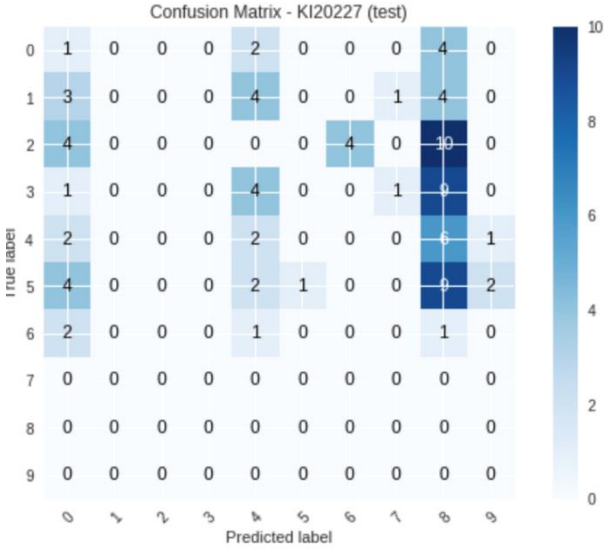
These results are also shown as confusion matrices in Figure 6 below. These matrices show the accuracy of each label prediction. As shown above, the training accuracies are much higher.




Confusion Matrix - Selumetinib (AZD6244) (train)
True label
Predicted label


Confusion Matrix - Selumetinib (AZD6244) (test)
True label
Predicted label


Confusion Matrix - KI20227 (train)
True label
Predicted label


Confusion Matrix - KI20227 (test)
True label
Predicted label

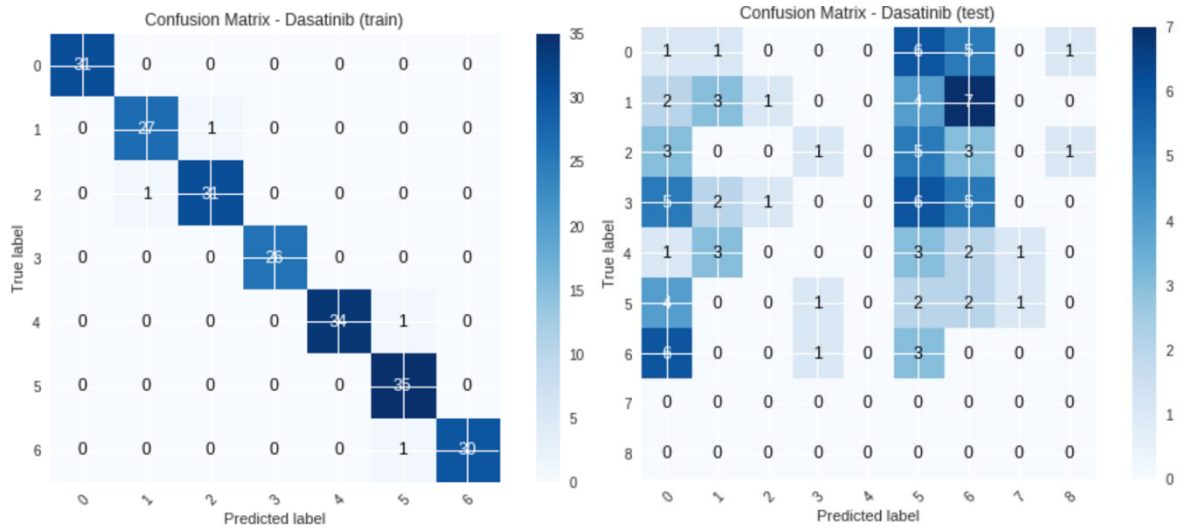
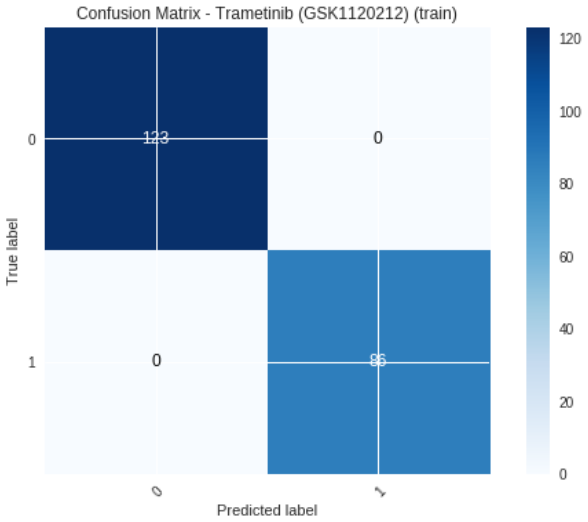
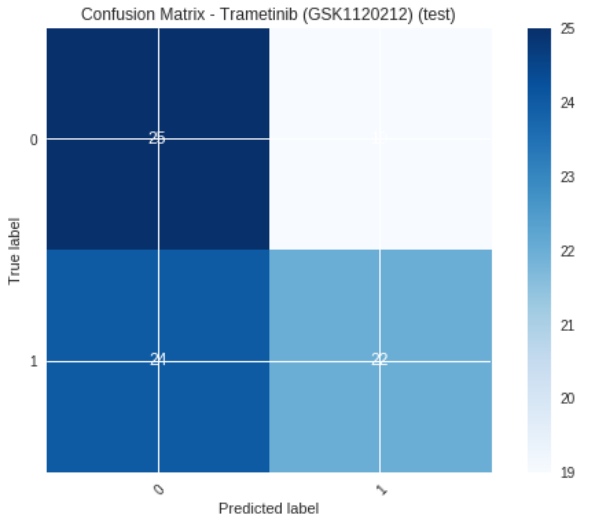
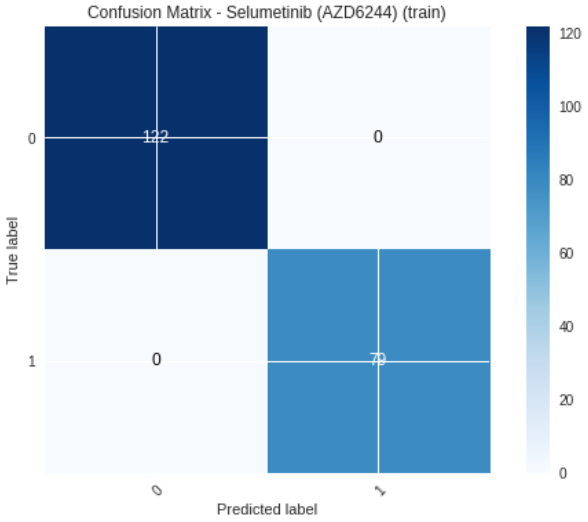
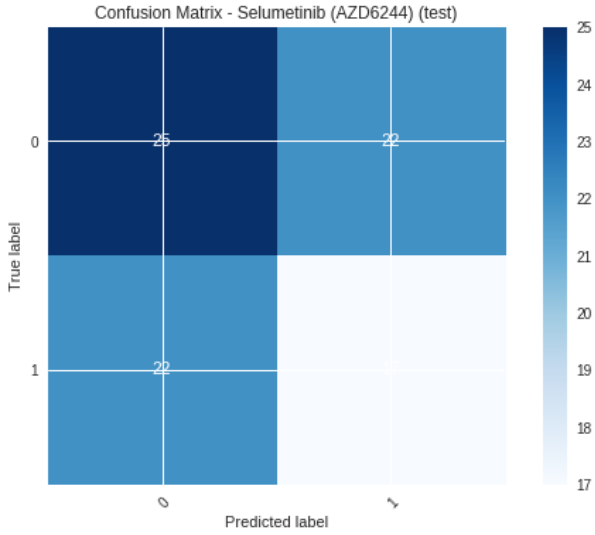


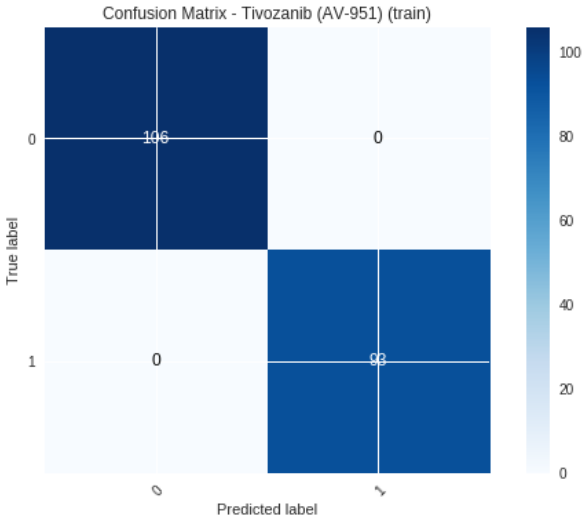
Figure 6: Confusion matrix plots for each drug tested, left shows training data results and right shows testing data results

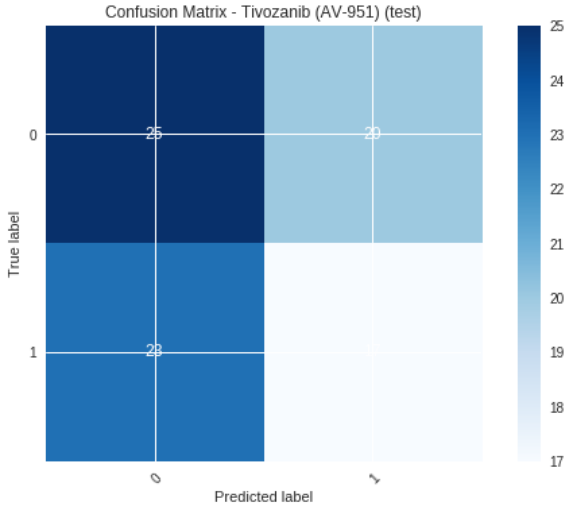

Confusion Matrix - Trametinib (GSK1120212) (train)
True label
Predicted label
123
0
0
86


Confusion Matrix - Trametinib (GSK1120212) (test)
True label
Predicted label
25
24
22


Confusion Matrix - Selumetinib (AZD6244) (train)
True label
Predicted label
122
0
0
79


Confusion Matrix - Selumetinib (AZD6244) (test)
True label
Predicted label
25
22
22


Confusion Matrix - Tivozanib (AV-951) (train)
True label
Predicted label
106
0
0
93


Confusion Matrix - Tivozanib (AV-951) (test)
True label
Predicted label
25
20
23

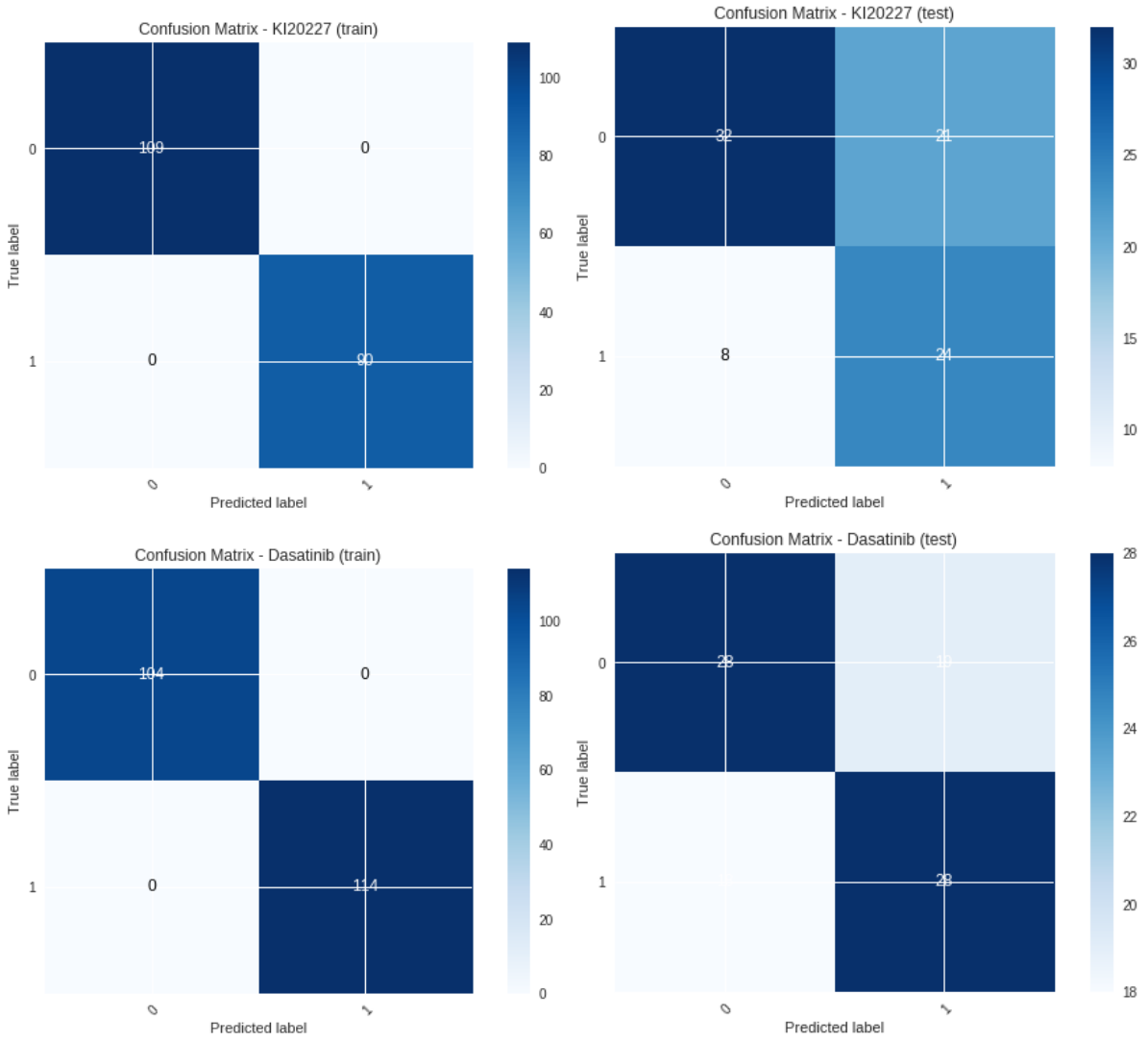


Figure 7: Confusion matrix plots for each drug tested, left shows training data results and right shows testing data results

4 Discussion

4.1 Conclusions

This work shows the importance of applying deep learning to problems with high dimensional data. The highly accurate training results that were calculated show how complex models can appropriately classify or predict values given small amounts of data if there are sufficient numbers of features. The models developed in this study show the difficulty in making predictions from such a dataset with small numbers of samples.

Feature importance evaluations show one method of interpreting so-called black box neural network models. In the drugs evaluated in the study, pathway analysis of drug target pathways mirrored the features that were most important in the network's prediction of drug sensitivity. Since this work is relevant to clinical settings, it is important to be able to explain why certain predictions are being made, instead of simply trusting the model predictions given a certain patient's genomic composition. Further work will be important in determining the factors which dictate why only certain drug predictions are able to be explained through pathway analysis.

Future experiments should be done to create datasets with greater numbers of samples in order to allow for prediction of drug sensitivity with higher accuracies. If future studies were able to sequence data for over 1000 patients and test their drug sensitivity as well for a similar number of drugs, it would be enormously helpful in training and creating more generalizable models for AML drug sensitivity predictions.

While analyzing and working with this problem, a couple of main challenges were encountered. First, was working with TensorFlow to build the graph model. This was difficult in creating the neural network model, learning about and implementing loss

functions, and creating the data pipeline. Furthermore, debugging in TensorFlow is quite difficult as normal print statements cannot be used in the graph flow. Tensorboard and the TensorFlow Print function were the two methods of debugging used. Memory errors and long training times were another issue due to the extreme high dimensionality of the data and the complexity of the model being implemented.

References

Barretina, Jordi et al. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity” Nature vol. 483,7391 603-7. 28 Mar. 2012, doi:10.1038/nature11003

Döhner, Hartmut, Daniel J. Weisdorf, and Clara D. Bloomfield. "Acute myeloid leukemia." *New England Journal of Medicine*373.12 (2015): 1136-1152.

Jain, Nitin, et al. "Phase II study of the oral MEK inhibitor selumetinib in advanced acute myelogenous leukemia: a University of Chicago phase II consortium trial." *Clinical Cancer Research* 20.2 (2014): 490-498.

Langfelder, Peter, and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis." *BMC bioinformatics* 9.1 (2008): 559.

Ley, TJ, Cancer Genome Atlas Research Network. "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia." *New England Journal of Medicine* 368.22 (2013): 2059-2074.

Lundberg, P. et al. Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. *Blood* **123**, 2220–2228 (2014).

Greenfest-Allen, Emily, et al. "iterativeWGCNA: iterative refinement to improve module detection from WGCNA co-expression networks." *bioRxiv* (2017): 234062.

Welch, J.S.et al. TP53 and decitabine in acute myeloid leukemia and myelodysplastic syndromes. N. Engl. J. Med. 375, 2023–2036 (2016).

Yang, Danwen, et al. "Tivozanib reverses multidrug resistance mediated by ABCB1 (P-glycoprotein) and ABCG2 (BCRP)." *Future oncology* 10.11 (2014): 1827-1841.

Yue, Tianwei & Wang, Haohan. (2018). Deep Learning for Genomics: A Concise Overview. Webb, Sarah. (2018). Deep learning for biology. Nature. 554. 555-557. 10.1038/d41586-018-02174-z.

Wanjuan Yang, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research, 41(D1):D955–D961, 2013. doi:10.1093/nar/gks1111. URL +http://dx.doi.org/10.1093/nar/gks1111.

Salama, April KS, and Kevin B. Kim. "Trametinib (GSK1120212) in the treatment of melanoma." *Expert opinion on pharmacotherapy* 14.5 (2013): 619-627.

Shibuya, Masabumi. "VEGFR and type-V RTK activation and signaling." *Cold Spring Harbor perspectives in biology* 5.10 (2013): a009092.

Sokolova, Marina, and Guy Lapalme. "A systematic analysis of performance measures for classification tasks." *Information Processing & Management* 45.4 (2009): 427-437.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* 15.1 (2014): 1929-1958.

Tyner, J. W, et al. (2018). Functional genomic landscape of acute myeloid leukaemia. Nature. https://doi.org/10.1038/s41586-018-0623-z