Journal of Research High School

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

Predicting Breast Cancer Diagnoses using Supervised Classification Models

Kyle Wang[1] *

[1]Radnor High School, Radnor, PA, USA
*Corresponding Author: kylewang122205@gmail.com

Advisor: Daniela Perry, dsperry@ucsd.edu

Received June 14, 2023; Revised September 23, 2023; Accepted, October 17, 2023

Abstract

Machine learning has many applications in the healthcare industry with the potential to save lives, one of which is detecting and diagnosing diseases based on images or predicting the likelihood of breast cancer given gene expression data. As a result, researchers have considered using machine learning techniques for faster diagnoses, which is critical for diseases like cancer when early detection can lead to a better prognosis. This study utilized the impact of supervised classification models, RNA-seq data from control patients and breast cancer patients. Gene expression read counts were subsequently normalized during the exploratory data analysis phase and split into training and testing data to create models that would help doctors draw conclusions about the presence of breast cancer. The study then introduced a separate validation set, to which the model could be used to predict a diagnosis. The paper explored various techniques to improve accuracy, such as reducing the number of significant genes used, altering the hyperparameters of each model, and normalizing data with a zero-inflated negative binomial distribution. The research yielded results with a maximum accuracy of 90.1% was obtained with both logistic regression models, and their performances were further analyzed using sklearn (Python machine learning tool) metrics. The models also discovered that patients with the gene markers ENSG00000201908, ENSG00000216184, and ENSG00000221326 exhibited the greatest variation in gene counts between breast cancer patients and control patients, which could be worth exploring in future studies.

Keywords: Machine learning, Neural network, Logistic regression, Random forest

1. Introduction

Twelve percent of women in the United States will be diagnosed with breast cancer at some point in their lives (Waks & Winer, 2019). Though somewhat common, breast cancer remains difficult to predict and prevent. Early diagnosis is one of the best methods for a full recovery (Sun et al., 2017). However, for low-income countries, equipment and resources for detecting breast cancer are not readily available, which decreases the rate of survival. Using various models in machine learning, trends based on genetics and demographics can be used to create models that predict the likelihood that someone has breast cancer. Using models to predict breast cancer is not new; using mammogram images to predict breast cancer has yielded accuracies as high as 94.20% (Lin et al., 2022). If one could predict the occurrence of breast cancer based on quantitative gene counts even before the mammogram imaging that occurs at an older age, people could take more intentional actions in reducing future breast cancer risk.

The data and demographics containing both breast cancer patients and control patients comes from a public dataset that was previously used for different purposes. The genes were sequenced and analyzed through a process known as RNA-seq, which allows us to measure RNA transcripts that are transcribed. The transcribed genes are broken into small cDNA segments, upon which adaptors are attached, further allowing us to duplicate these segments through a polymerase chain reaction (PCR) and analyze accordingly (Wang, Gerstein, & Snyder, 2009). When comparing

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

expression levels between breast cancer patients and normal patients, a significant difference in gene counts could often be found and used to create a model for prediction.

The best prediction type of model that can be used to analyze this data is the supervised classification model, meaning the model has predetermined and labeled input values (specific genes and respective gene counts), and the output values are discrete variables (presence or absence of breast cancer). Three main classification models were used to find a relationship between genes and diagnosis: neural networks, logistic regression, and a random forest. Neural networks randomly select variables and add random weight values to calculate a predicted value, which is compared to the actual value to determine whether the prediction was accurate. If not, the model begins backpropagation, regenerating new weights for the variables and repeating the process for many epochs until the model converges. Logistic regression can be used with data with binary outcomes, converting continuous probabilities into a discrete 0 or 1. Random forests generate many decision trees and randomly selecting variables, maintaining trees that predict most accurately in the model. With many decision trees, the chance of having just one strong predictor is very high (Ren, Cheng, & Han, 2017). After determining the probability of a patient having breast cancer, a diagnosis can be made based on which outcome is more likely. These models will be employed to predict the probability of the presence of breast cancer based on gene expression data.

2. Materials and Methods

The data for this project was obtained through SILVER-seq (Small Input Liquid Volume Extracellular RNA Sequencing), which utilizes extracellular RNA found in human serum to compare expression levels of about 25% of the human genes (Zhou et al., 2019). The original training data consisted of 128 patients: 96 of the patients were breast cancer patients, while 32 of them were normal control patients. Using this dataset, various models created their parameters based on the gene counts for each patient. Molded from the training data, the new models were later compared to a validation set to determine how well each model predicts breast cancer diagnosis of a new set. For the validation data set, there were 161 total patients: 86 breast cancer patients and 75 normal control patients. Throughout this study, both training accuracies and validation accuracies were compared: the classification processes strive to find models for the gene count data that would correctly predict the breast cancer diagnoses from the training set to obtain a high training accuracy. However, generating a high validation accuracy is more important, as the previous model is then used on a new validation dataset to check the applicability of the model in a slightly different context. The challenge is generating a model with high training and validation accuracy, meaning the model cannot overfit the training data so heavily that it is useless and inaccurate in a slightly different context.

2.1 Neural Network

In order to effectively analyze high-dimensional data, the deSEQ-2 package, a program in the Python programming language that allows for easier processing of raw data like gene counts for use in a supervised model, was used to normalize and visually display the gene expression data. In the first iteration of the neural network, the dataset for breast cancer patients and normal patients were imported and transposed in order to ensure that each patient served as the sample with the characteristics of gene counts. Lots of data cleaning was required to remove the unnecessary data and add a diagnosis column so that the model can compare the model results to the actual diagnosis to determine the proportion of diagnoses that was correctly predicted. Gene count values were also converted from integer values to proportions, so that larger SILVER-seq samples won't affect the dataset. Using the standard scaler normalization from the sklearn package, data was inputted into a sequential neural network model. This model contained 4 dense models with 16, 8, 4, and 2 neurons, with the final output layer having 1 neuron and a sigmoid activation. One hundred epochs of training were run, with gradual improvement in performance after each epoch (high val_acc and low val_loss). After fitting this model into a new test dataset, the model could now try to predict the validation set.

In iteration two of the neural network, it was observed that some genes seemed to be extraneous, and the addition of these genes could add extra variables into the model that would only ruin its performance. As a result, genes were

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

filtered and only included genes with the most drastic count difference between normal patients and breast cancer patients (having a p-value less than 0.05 and suggesting a strong correlation between the gene and the occurrence of breast cancer), which narrowed the number of genes from 60675 to 2109. The neural network model's accuracy improved by 14.9% with this change.

2.2 Logistic Regression

The process also involved experimenting with other model types, such as the logistic regression model. The data cleaning process was similar to that of the neural network, with a similar train-test split of 0.50, meaning half of the data values are split between the training and test data. The implementations of the actual models were as simple as importing the logistic regression function from the sklearn linear model class and fitting the initial model using the gene counts and diagnosis of the 64 training values. Comparing the model-predicted diagnoses of the test data and the actual diagnoses had an 100% accuracy, but a comparison to a different validation set was necessary to ensure that the model was not overfit and feasible in a slightly different context. The validation set's predicted results were compared to its actual results, and the accuracy was slightly less than 100%. Therefore, the models were not a perfect fit, but the data was also not significantly overfit.

2.3 Random Forest

The random forest model had a similar initial setup, but the sklearn ensemble model class imported the random forest regressor model. The model had two hyperparameters that were controlled: n_estimators and random_state. In a random forest model, the n_estimators determine the number of trees that will be included in the random forest, with a greater number of trees typically correlated with a greater accuracy but slower model loading times. However, adding more trees also makes the model more prone to overfitting, leading to a lower accuracy when comparing the training data model to a different validation set. Choosing 20 trees seemed to maintain the highest accuracy without overfitting. The random_state keeps the end accuracy relatively consistent after every trial.

2.4 Zero-inflated Negative Binomial Normalization

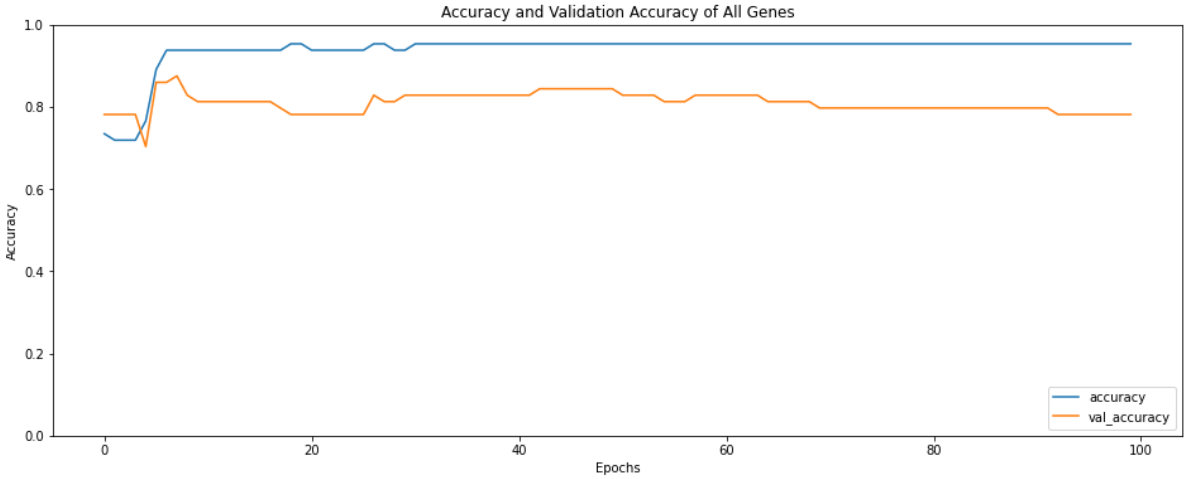
In the third iteration, a new normalization technique is employed that could potentially increase accuracy. A new normalization method – zero-inflated negative binomial normalization – was used, in which the excess zeros are modeled independently of the rest of the data. This normalization is used often in scientific research in overdispersed data that contains lots of zeros, as the zeros increase the amount of variability, which distorts the standard deviation and prevents an accurate normalization from being created (Alam, Al Mahi, & Begum, 2018). This new normalization method created slightly different normalized values with varying results in its predictions using the neural network model, logistic regression model, and random forest model.

3. Performance and Results

The model prediction of the training set data for the neural network model tracked accuracy and validation accuracy most closely. The first neural network model iteration that contained 60675 different genes had the highest training accuracy of 95.31%, but the validation accuracy never reached higher than 78.12%, decreasing as the model ran, which is a sign of overfitting. Admittedly, final accuracy and validation accuracy for all neural networks had variation due to the nature of neural networks, especially the original model, as 60,675 variables are randomly assigned weights with a small sample size to which the models are trained and validated. Using the same model for the significant genes only showed both a higher training accuracy of 100%, but the validation accuracy also continued to increase, with a maximum test accuracy of 96.88%. This contrast can be viewed more clearly in figure 1.

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

1a



1b

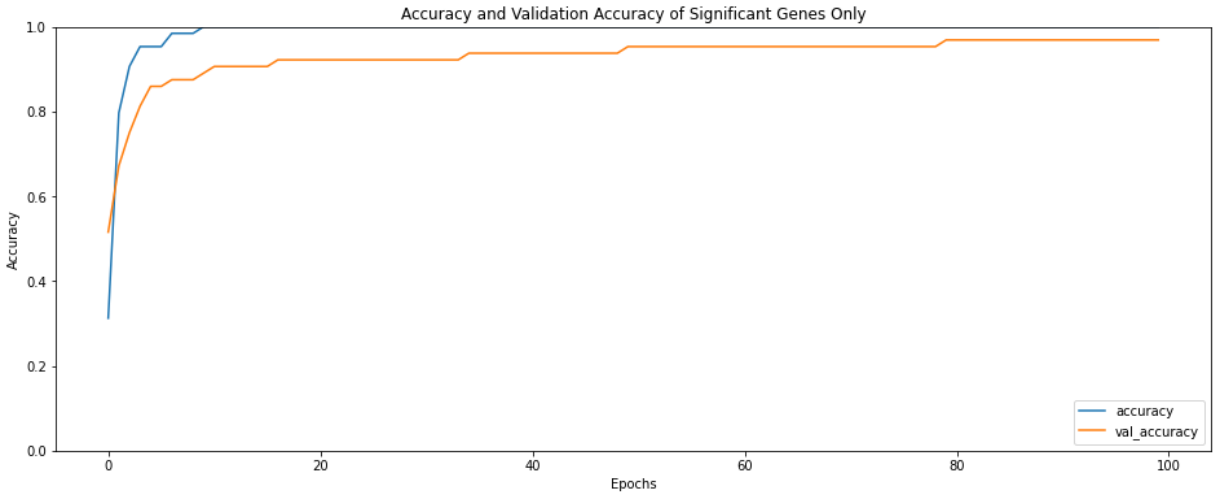


Figure 1: Comparison of accuracy and validation accuracy increase in neural networks with 1a: 60,675 genes (top) to 1b: 2,109 genes (bottom)

The first iteration of the neural network contained too many different types of gene counts to yield an accurate neural network model, resulting in a test accuracy of 55.9% - just barely better than guessing. However, the effect of reducing the number of variables to include only the significant gene was massive, as the accuracy rate increases to 70.8% for the neural network model with significant genes only.

At this stage, the random forest and logistic regression models were also introduced, and accuracy rates were markedly higher than the neural network rate, correctly predicting the breast cancer diagnosis 82.0% and 90.1% of the time, respectively.

Although higher accuracies for the zero-inflated negative binomial distribution were expected, accuracies actually seemed to remain relatively the same, if not worse. Using this normalization for both the training and validation sets, decreased the neural network model to an accuracy of 62.1%, slightly increased the random forest model's test accuracy to 82.6%, and kept the logistic regression's test accuracy at the same value of 90.1%. The various validation accuracies for each iteration are shown in table 1.

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

3.1 Confusion Matrix

Table 1: Final validation accuracies in seven models used.

Model	Accuracy Scores
Iteration 1: Neural Network (60,675 genes)	55.9%
Iteration 2a: Neural Network (2,109 genes)	70.8%
Iteration 2b: Random Forest (2,109 genes)	82.0%
Iteration 2c: Logistic Regression (2,109 genes)	90.1%
Iteration 3a: Neural Network (Zero-inflated negative binomial normalization)	62.1%
Iteration 3b: Random Forest (Zero-inflated negative binomial normalization)	82.6%
Iteration 3c: Logistic Regression (Zero-inflated negative binomial normalization)	90.1%

The accuracies of each model can be visually represented using a confusion matrix. Each prediction made by the model is split into four categories. The top-left box represents true-negative, meaning that the patient was correctly predicted to be a normal control patient, and the bottom-right box represents true-positive, meaning that the patient was correctly predicted to be a breast cancer patient. The top-right box represents a false-positive result, in which the model predicts that the patient has breast cancer, but the actual diagnosis of the patient is normal. Finally, the bottom-left box represents false-negative, in which the model predicts that the patient is normal when the patient actually does have breast cancer (Scikit-learn Developers., n.d.). Figure 2 shows the accuracy of each prediction made by each model in a visual manner, with colors indicating the number of predictions that fell under each category. Accuracy can be determined from the matrix by summing the values in the top-left and bottom-right boxes and dividing that value by the sum of all four values. An ideal model should have the majority of samples in the top-left and bottom-right boxes.

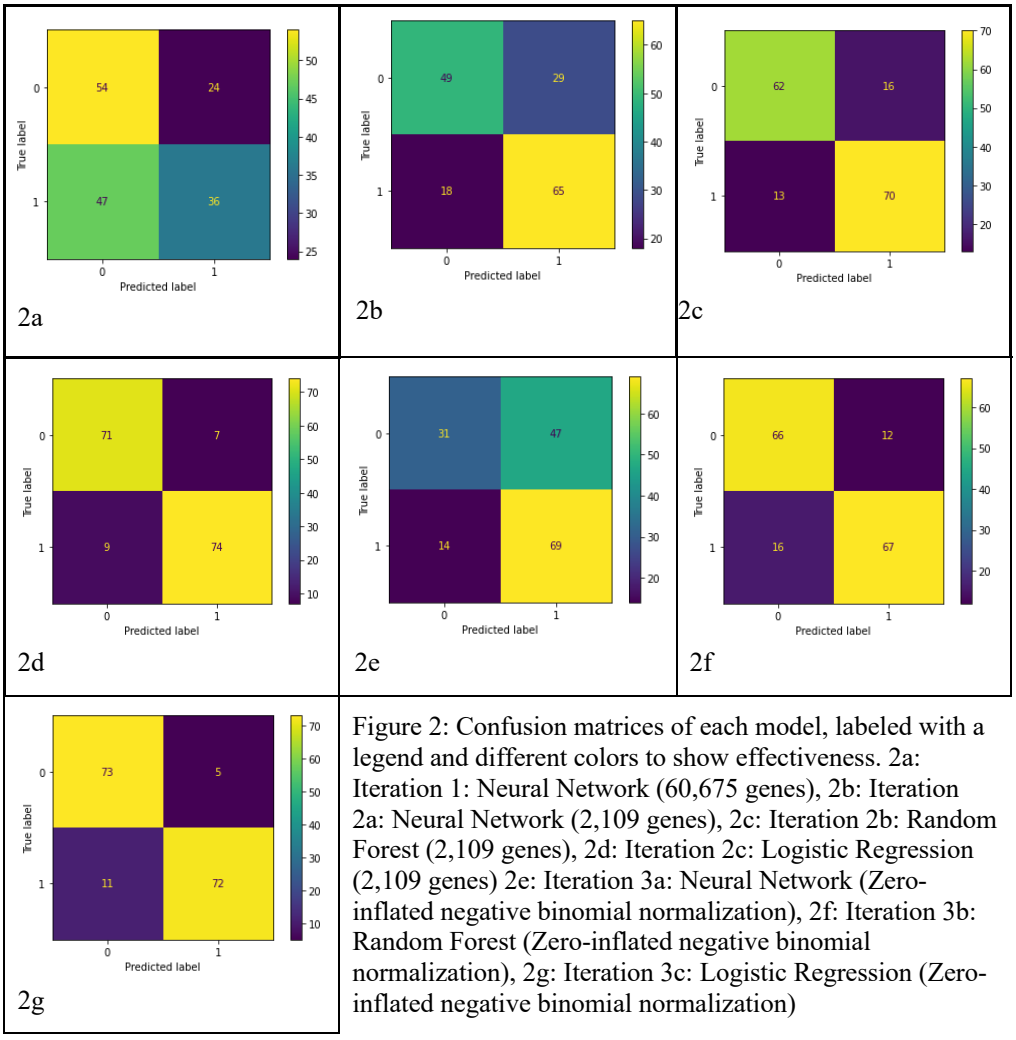


Figure 2: Confusion matrices of each model, labeled with a legend and different colors to show effectiveness. 2a: Iteration 1: Neural Network (60,675 genes), 2b: Iteration 2a: Neural Network (2,109 genes), 2c: Iteration 2b: Random Forest (2,109 genes), 2d: Iteration 2c: Logistic Regression (2,109 genes) 2e: Iteration 3a: Neural Network (Zero-inflated negative binomial normalization), 2f: Iteration 3b: Random Forest (Zero-inflated negative binomial normalization), 2g: Iteration 3c: Logistic Regression (Zero-inflated negative binomial normalization)

The confusion matrix in figure 2 is further labeled with colors that indicate the number of samples in each category, with yellow being the greatest and dark purple being the least. In a strong model like the one in figure 2g, a mix of yellow and dark purple indicates that the model correctly predicts the diagnosis most of the time and rarely makes mistakes. Looking at both the accuracy and the confusion matrix, one can see that the logistic regression models are the strongest, while the neural networks are

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

the weakest because there are lots of blue colors in the matrices, indicating that the matrix is about equally likely to predict correctly as incorrectly, resulting in a lower accuracy.

The performance of each model can be further investigated by utilizing various classification metrics.

3.2 Receiver Operating Characteristic Curve

Exploration began with the ROC curve and the ROC_AUC score. The ROC curve, which stands for a receiver operating characteristic curve, plots the sensitivity of the model compared 1 - specificity on the x-axis (Mandrekar, 2010). The sensitivity model represents how often the model correctly predicts an outcome correctly marked as positive or having a "cancer" diagnosis, compared to the total number of positive predictions, which would include normal patients who were predicted to have breast cancer. In contrast, the specificity model represents an outcome that is correctly marked as negative with a "normal" diagnosis, compared to the total number of negative predictions, which includes cases in which a patient with breast cancer is predicted to not have breast cancer.

It's often difficult to have a model with perfect sensitivity and specificity, so the model should typically prioritize one over the other. In this case, it's better to take precautions and do tests on a patient without breast cancer versus neglecting a breast cancer patient by claiming that they do not have cancer. Having a false negative is not ideal, so a model with higher sensitivity, or recall, is preferred (Lalkhen & McCluskey, 2008).

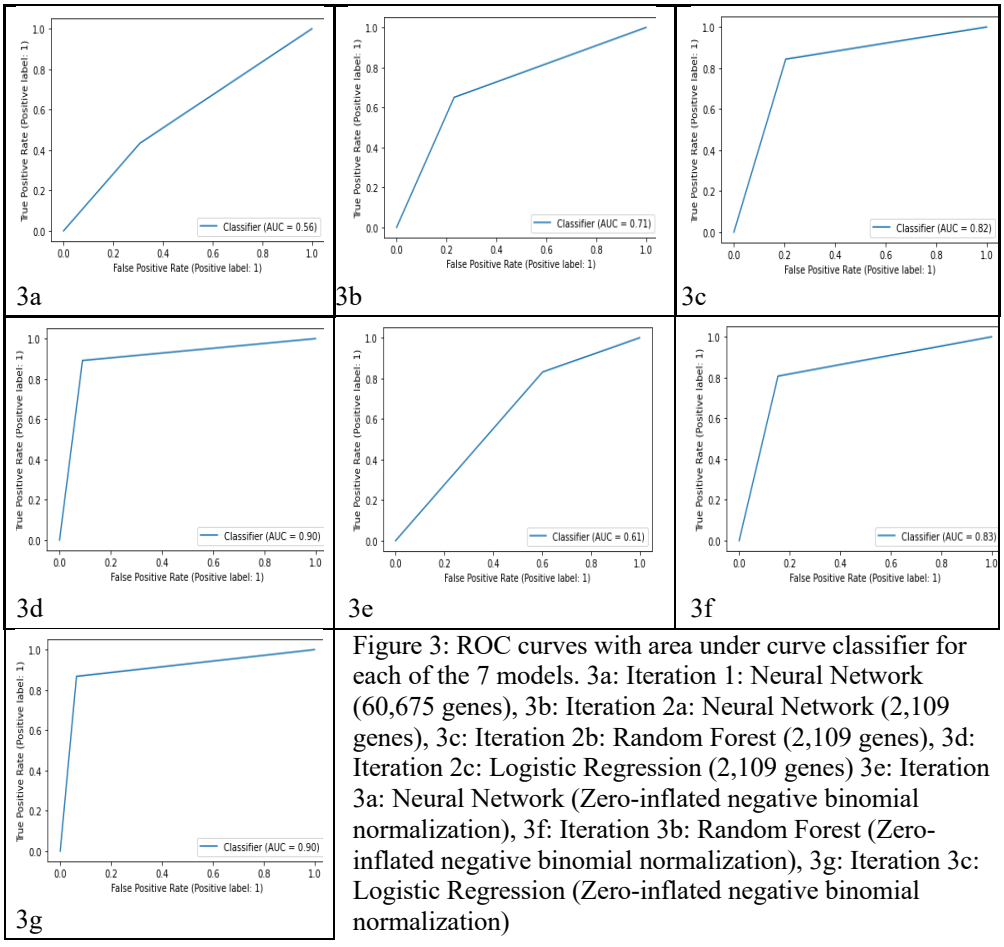


Figure 3: ROC curves with area under curve classifier for each of the 7 models. 3a: Iteration 1: Neural Network (60,675 genes), 3b: Iteration 2a: Neural Network (2,109 genes), 3c: Iteration 2b: Random Forest (2,109 genes), 3d: Iteration 2c: Logistic Regression (2,109 genes) 3e: Iteration 3a: Neural Network (Zero-inflated negative binomial normalization), 3f: Iteration 3b: Random Forest (Zero-inflated negative binomial normalization), 3g: Iteration 3c: Logistic Regression (Zero-inflated negative binomial normalization)

The ROC curves for the models in figure 3 were generated by comparing results from the actual diagnoses in the validation set and the predicted diagnoses in the validation set according to the corresponding model. The curves in these models are composed of two line segments with different slopes, beginning at (0.0, 0.0) and ending at (1.0, 1.0)

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

in all cases. Models, like the one seen in figure 3d, that start with a large slope are typically better-performing curves, whereas curves with a generally constant slope like 1 are poorer-performing curves. The ROC_ AUC score (located in the bottom left of each graph) also ranges from values between 0.5 and 1, with numbers closer to 1 also indicating better performance (Hoo, Candlish, & Teare, 2017).

3.3 Recall and Precision

The recall and precision are two other metrics that are worth noting. The recall is the same as sensitivity, measuring the number of breast cancer patients who were correctly predicted as having breast cancer, as some breast cancer patients were predicted to have a normal diagnosis, causing a type II error, as seen in equation 1.

$$TP \,/\, (TP + FN) \qquad \text{eq. (1)}$$

where TP = true positives and FN = false negatives

On the other hand, precision measures the number of breast cancer patients who were correctly predicted as having cancer as a percentage of the total number of patients who were given a positive diagnosis, causing a type I error, as seen in equation 2.

$$TP \,/\, (TP + FP) \qquad \text{eq. (2)}$$

where TP = true positives and FP = false positives

As mentioned previously, a type II error would be more detrimental in this model, so the number of false negatives should be mitigated, so recall and sensitivity should ideally have a proportion closest to 1.

Table 2 highlights the precision and recall values found after generating each model. As expected, model 1 performed poorly with the lowest precision and recall of 60.0% and 43.3%. This model is extremely prone to predicting normal diagnoses in patients with breast cancer, which would be dangerous, since people would not receive the proper treatment. Iteration 2c had the highest precision and recall rate of 91.4% and 89.2%, respectively, meaning out of these seven models, logistic regression currently looks like the best model to predict breast cancer. It should be noted that iterations 2a and 3a, neural network models, have recall percentages that are significantly higher than the precision percentages: in 3a, the difference is 23.6%. Neural networks that prioritize recall over precision would be more likely to have false positives that lead to a higher cost of additional testing but more breast cancer screening that ensures that patients are healthy.

Table 2: Precision and recall scores for each of the seven models.

Model	Precision	Recall
Iteration 1: Neural Network (60,675 genes)	60.0%	43.3%
Iteration 2a: Neural Network (2,109 genes)	69.1%	78.3%
Iteration 2b: Random Forest (2,109 genes)	81.3%	84.3%
Iteration 2c: Logistic Regression (2,109 genes)	91.4%	89.2%
Iteration 3a: Neural Network (Zero-inflated negative binomial normalization)	59.5%	83.1%
Iteration 3b: Random Forest (Zero-inflated negative binomial normalization)	84.8%	80.7%
Iteration 3c: Logistic Regression (Zero-inflated negative binomial normalization)	93.5%	86.7%

The precision recall curve for each iteration in figure 4 visually represents the relationship between precision and recall and is a metric that is suitable for skewed breast cancer gene count data, in which the proportions are typically zero or very close to zero (Davis & Goadrich, 2006). The precision is labeled on the y-axis, and the recall is labeled on the x-axis, with the graph starting from the top-left part of the graph and ending at the bottom-right part of the

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

graph. In an ideal situation, both the precision and recall should be high, so the precision-recall graph should have a horizontal line as high on the graph as possible.

The classifiers in these graphs are found using average precision, which is the area under the curve (Kielwagen, Grosse, & Grau, 2014). A higher number correlates with a better model in terms of the precision and recall metric.

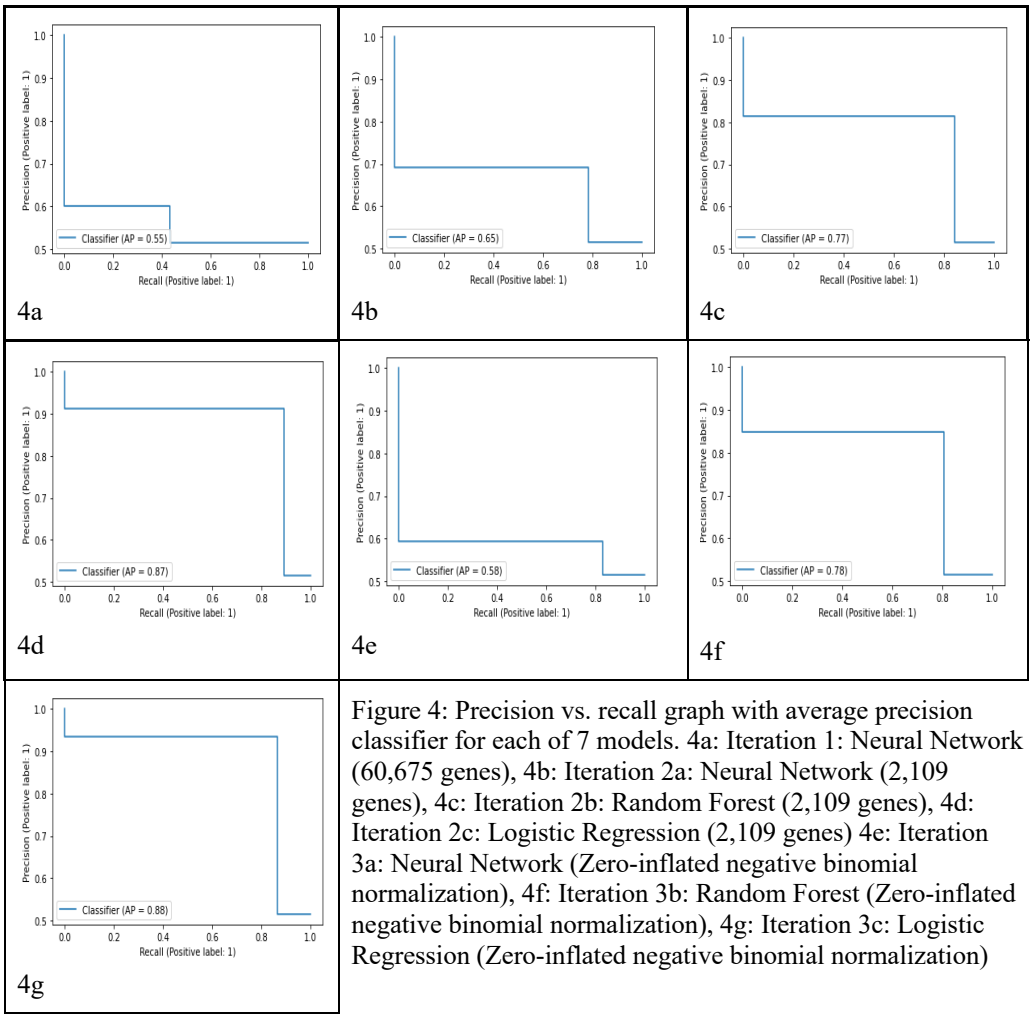


Figure 4: Precision vs. recall graph with average precision classifier for each of 7 models. 4a: Iteration 1: Neural Network (60,675 genes), 4b: Iteration 2a: Neural Network (2,109 genes), 4c: Iteration 2b: Random Forest (2,109 genes), 4d: Iteration 2c: Logistic Regression (2,109 genes) 4e: Iteration 3a: Neural Network (Zero-inflated negative binomial normalization), 4f: Iteration 3b: Random Forest (Zero-inflated negative binomial normalization), 4g: Iteration 3c: Logistic Regression (Zero-inflated negative binomial normalization)

As anticipated from previous metrics, the original neural network with 60,675 genes performed the worst, as the precision drops to 0.6 immediately, and the average precision is the lowest of the seven models at 0.55. Both logistic regression models performed well, as average precision was the highest at 0.87 and 0.88 in figures 4d and 4g, respectively. Average precision is correlated with accuracy, as the two metrics never deviated by more than 0.05 for each corresponding model, so this is an effective metric for this dataset.

4. Discussion

4.1 Limitations

Although the models seem to have lower numbers, the zero-inflated negative binomial distribution did not effectively normalize the data because of the nature of the zeros. The use of this model typically assumes that both structural zeros (zeros that occur because of some restriction that forces a value to be null) and random zeros (zeros that occur randomly in the dataset without possible restrictions) are present in the model. All genes are present in each

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

patient, so all zeros that occur in the sample are random and not structural, making the use of this model questionable (Hawinkel et al., 2020). To further increase accuracy rate, it is necessary to try to find a better normalization technique that makes sense in the context of the data. A larger sample size should also be considered, as having only 128 samples limits the amount of data that can be used to more effectively train the model. A larger sample size would also mitigate the amount of standard deviation that made accuracies in each trial slightly different.

4.2 Future Directions

While many genes were considered in combination to generate the model, it is important to take note the gene counts with the greatest difference between breast cancer patients and normal patients: ENSG00000201098, ENSG00000216184, ENSG00000221326, all of which had the lowest p-value. ENSG00000201098 is also referred to as the RNY1 gene that plays a role in chromosomal DNA replication, which makes sense as cancer is often caused by mutations in the DNA (GeneCaRNA., 2023, May 22). Future investigation into these genes may discover additional links to breast cancer that could serve as the starting point for developing a drug that could combat it.

5. Conclusion

Artificial intelligence has made significant advancements in the healthcare industry in terms of diagnosing diseases like cancer and finding ideal drugs and solutions to remedy these diseases (Davenport & Kalakota, 2019). In addition to the neural networks, logistic regression, and random forest models that can be used, machine learning can continue to be optimized to convert data and numbers on a spreadsheet into actionable steps that can help machines make the most reasonable and unbiased diagnoses and treatments. Though hyperparameters and normalizations could be further optimized, the logistic regression models had the highest accuracies of 90.1% when determining breast cancer diagnoses from gene counts. Normalizing the logistic regression models with the standard scaler or using the zero-inflated negative binomial distribution had no significant effect on the performance, as verified by metrics such as precision, recall, and receiver operating characteristic curves.

References

Alam, M., Al Mahi, N., & Begum, M. (2018). Zero-Inflated Models for RNA-Seq Count Data. *Journal Articles: Biostatistics*. 4, 1. https://digitalcommons.unmc.edu/coph_biostats_articles/4

Davenport, T. & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthc J.* 6(2), 94-98. 10.7861/futurehosp.6-2-94.

Davis, J. & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *ICML '06: Proceedings of the 23rd international conference on Machine learning*. 233-240.

GeneCaRNA. (2023, May 22) *RNY1 Gene - RNA, Ro60-Associated Y1*. GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RNY1

Hawinkel, S., et al. (2020). Sequence count data are poorly fit by the negative binomial distribution. *PLoS One.* 15(4), e0224909. https://doi.org/10.1371/journal.pone.0224909

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve?. *Emergency medicine journal : EMJ.* 34(6), 357-359. 10.1136/emermed-2017-206735.

Keilwagen, J., Grosse, I., & Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data. *PloS One.* 9(3), e92209. 10.1371/journal.pone.0092209.

Lalkhen, A. G. & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain.* 8(6), 221-223. https://doi.org/10.1093/bjaceaccp/mkn041

Vol. 2023 (2) 318 – 327
ISSN 2688-3651 [online]

Lin, R.-H., et al. (2022) Application of Deep Learning to Construct Breast Cancer Diagnosis Model. *Applied Sciences*. 12(4):1957, https://doi.org/10.3390/app12041957

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*. 5(9), 1315-1316. 10.1097/JTO.0b013e3181ec173d.

Ren, Q., Cheng, H., & Han, H. (2017). Research on machine learning framework based on random forest algorithm. *AIP Conference Proceedings*. 1820, 080020. https://doi.org/10.1063/1.4977376

Scikit-learn Developers. (n.d.) *sklearn.metrics.ConfusionMatrixDisplay*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html

Sun, Y.-S., et al. (2017). Risk factors and preventions of breast cancer. *Int J Biol Sci.* 13(11), 1387-1397. 10.7150/ijbs.21635.

Waks, A. G. & Winer, E. P. (2019). Breast cancer treatment: a review. *JAMA.* 321(3), 288–300. 10.1001/jama.2018.19323.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57–63. https://doi.org/10.1038/nrg2484

Zhou, Z., et al. (2019). Extracellular RNA in a single droplet of human serum reflects physiologic and disease states. *National Academy of Sciences.* 116(38), 19200-19208. 10.1073/pnas.1908252116.