Health and Wellness

Analyzing Machine Learning Models to Classify Triple-Negative and Non-Triple Negative Breast Cancer

Emma Heath, Rithika Nair, Youqing Luo

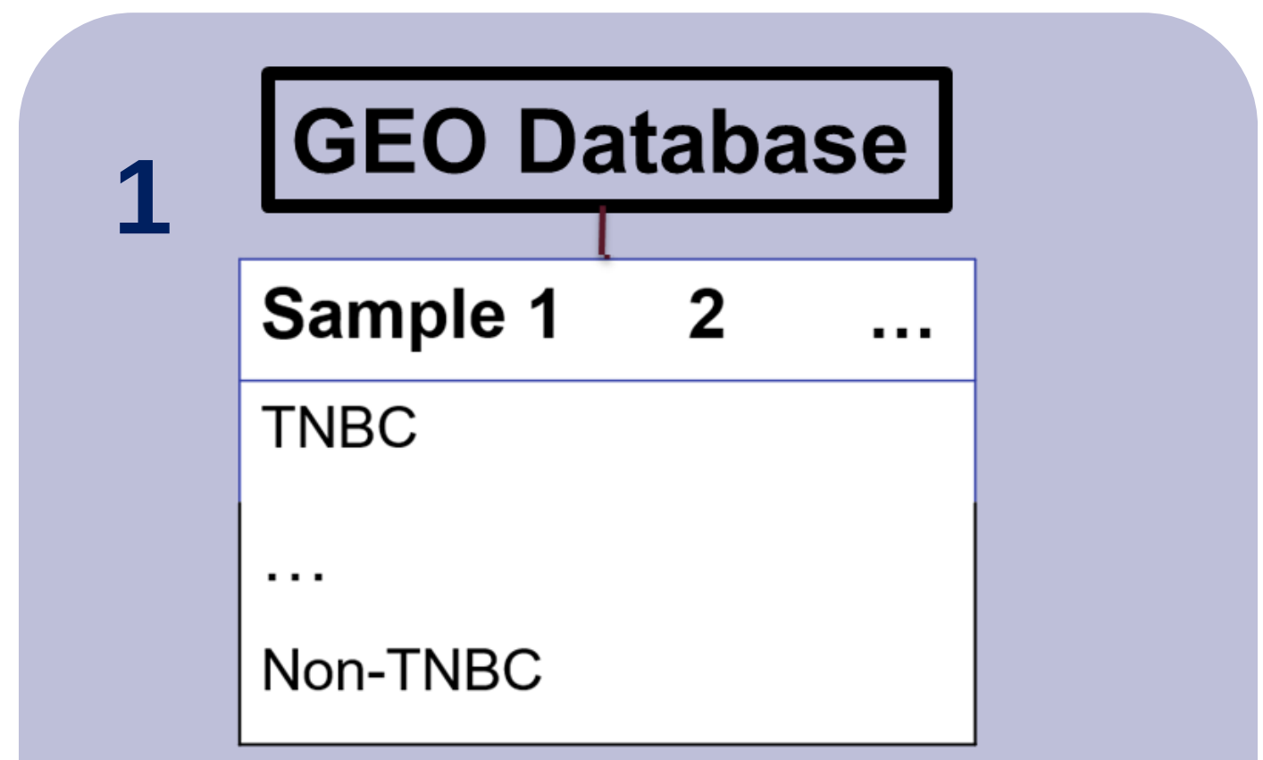
Background

- Breast cancer is one of the most common and lethal cancers in women. Triple-negative Breast Cancer (TNBC) is a specific subtype of breast cancer with a lower survival rate, accounting for ~15% of all breast cancers [1].
- Traditional diagnosis methods include histopathological assay, imaging methods, and benign features in images. However, these methods can be misinterpreted, so more accurate and specific methods are needed.
- RNA sequencing is a useful tool that can provide proxy measurements of gene expression to compare transcriptomic data (data involving RNA expression levels) between biological samples [2].
- Machine learning, combined with differential gene expression analysis, is receiving more attention in cancer identification and prognosis [2]. Recently, studies have compared basic binary statistical methods in breast cancer type classification to determine which models are most useful[4].

Objective: Use publicly available RNA-seq data to train different machine learning models to classify TNBC and non-TNBC. Afterwards, analyze the different models to determine which one has a higher accuracy.

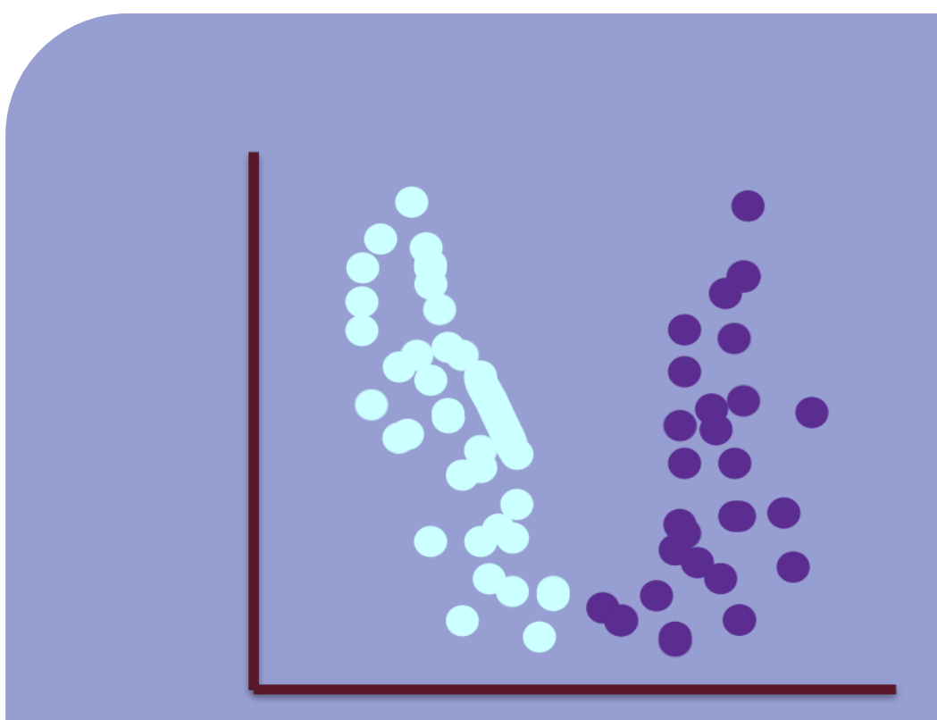
Methods

1



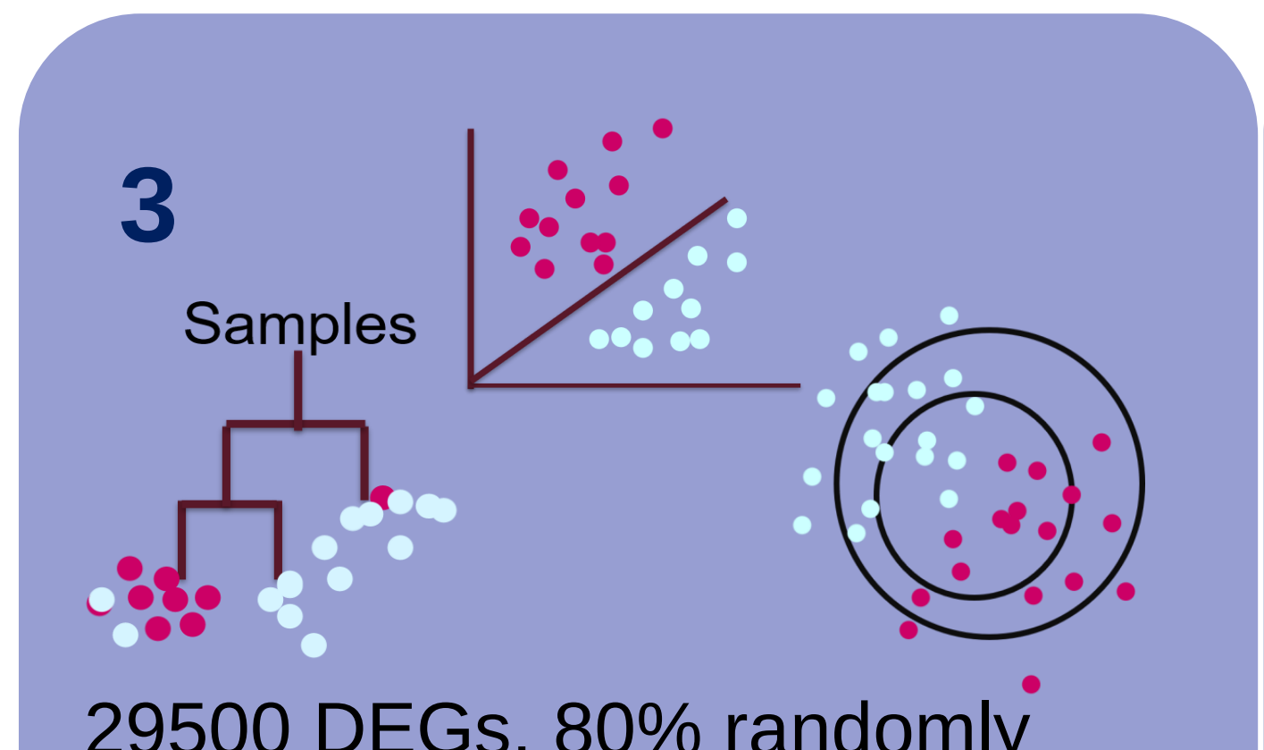
962 samples gathered from GEO (434 TNBC and 528 non-TNBC). Data selected that had clinical phenotyping and RNA seq information.

2



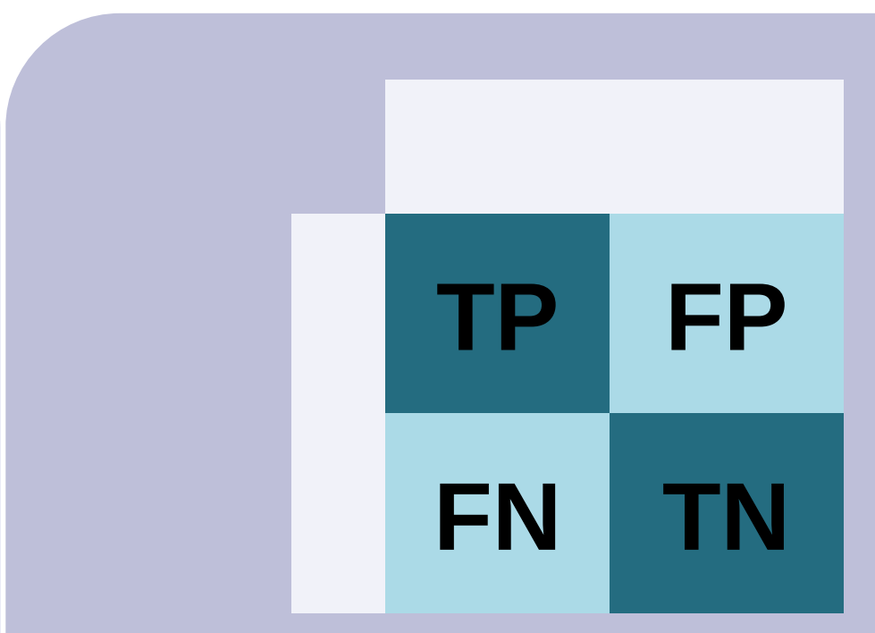
Differentially expressed gene analysis and feature selection was done with LIMMA. Criteria - $|\log_2 FC| > 1$ and p-value<0.05

3



29500 DEGs, 80% randomly selected data for training in 4 different modules: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT) and Naïve Bayes (NGB), and 20% for testing.

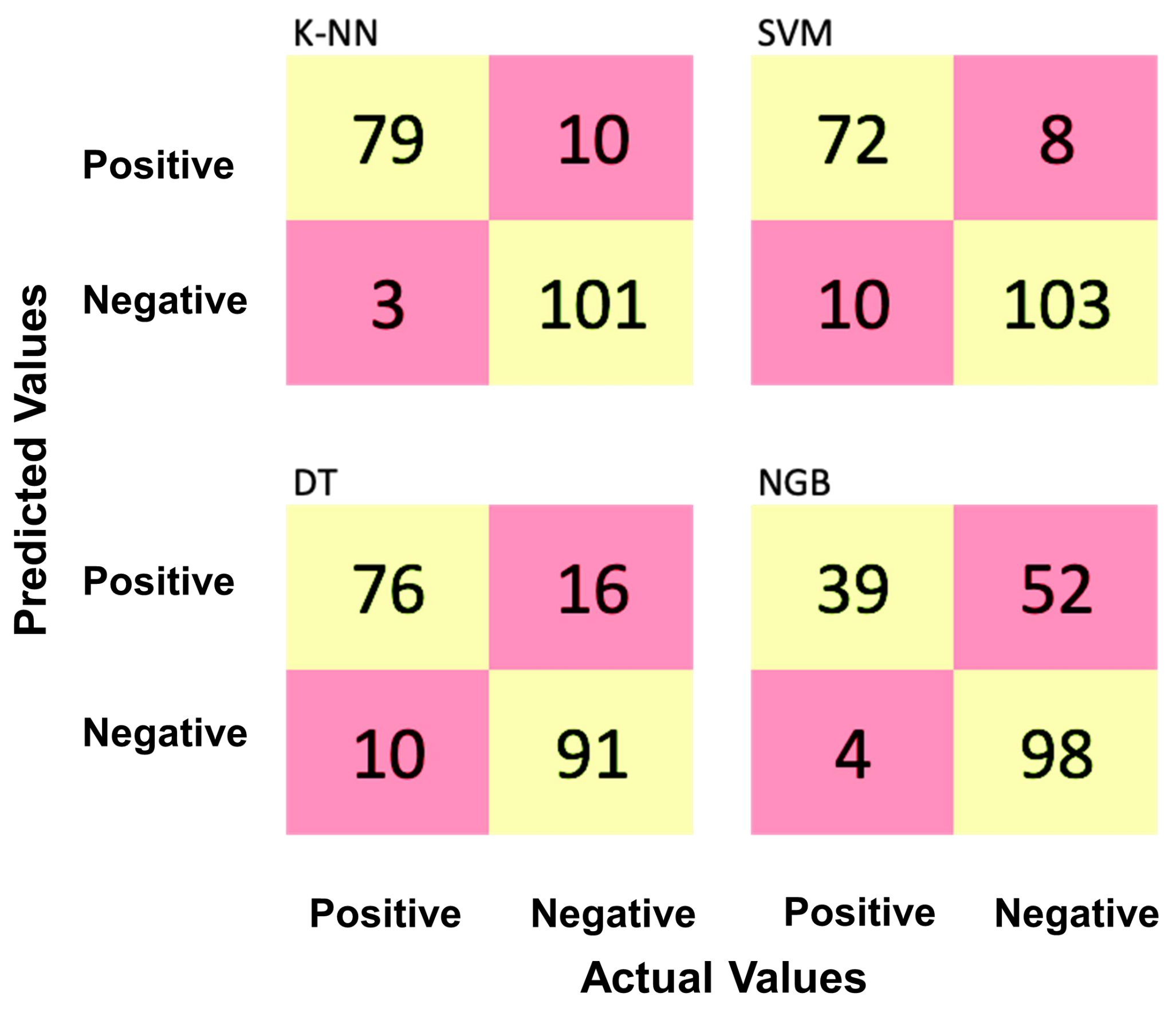
4



TP	FP
FN	TN

Evaluation: Calculated Accuracy, Precision, Recall, F1 score and Specificity. Further evaluated by using the False Positive Rate and False Negative Rate.

Results


Predicted Values
Actual Values

K-NN

	Positive	Negative
Positive	79	10
Negative	3	101

SVM

	Positive	Negative
Positive	72	8
Negative	10	103

DT

	Positive	Negative
Positive	76	16
Negative	10	91

NGB

	Positive	Negative
Positive	39	52
Negative	4	98

Figure 1: Confusion matrices. 20% samples were used to test the accuracy of each model. Yellow boxes are True Positive and True Negative. Pink boxes are False Positive and False Negative.

	Support Vector Machine	K-Nearest Neighbor	Decision Tree	Naives Bayes
Accuracy	90.67%	93.26%	86.53%	70.98%
F1 score	88.89%	92.40%	85.39%	58.21%
False Positive Rate	7.21%	9.01%	14.95%	34.67%
False Negative Rate	12.20%	3.66%	11.63%	9.30%

Figure 2: Results Comparing Models. Red text refers to the model that performed the best and blue text refers to the model that performed the worst. F1 was calculated based off of precision and recall calculations.

- Our investigation revealed that KNN is the best model when compared to SVM, Decision Tree, and Naïve Bayes.

Discussion

Comparison between different models:

- KNN has the highest accuracy followed by SVM, Decision Tree, and Naive Bayes models. However, using accuracy as the only evaluation metric for the models is insufficient.
- Precision depicts the proportion of samples that are truly TNBC amongst all the samples predicted to be TNBC. Recall depicts the proportion of how many samples were correctly predicted as TNBC amongst all the TNBC samples in the data. It is important to consider the two evaluation metrics as well and thus, we use the F1 score which is the harmonic mean of recall and precision. KNN yet again has the highest F1 score followed by SVM, Decision Tree, and Naive Bayes models.
- False Positive Rate (FPR) and False Negative Rate (FNR) further show the accuracy of the model. The lower the false positive rate, the more specific the model, as the amount of non-TNBC samples misclassified as TNBC will be less. SVM has the lowest FPR followed by KNN, Decision Tree, and Naïve Bayes models. Similarly, the lower the false negative rate, the more sensitive the model is, as the number of TNBC misclassified as non-TNBC will be less. KNN has the lowest FNR followed by Naïve Bayes, Decision Tree, and SVM models.

Limitations

- Only 962 data samples were used, which is a relatively small amount of data.
- Only 4 different algorithms were evaluated. Many more algorithms exist that can be employed to classify breast cancer subtypes.
- Different subtypes within TNBC exist, and the models used did not account for this.

Future Research

- Future research could focus on using larger amounts of data in newly built machine learning models that take into account the different subtypes of TNBC.

Conclusion

The combination of molecular biology and machine learning offers new insight to breast cancer classification methods. Four commonly used algorithms were used to classify TNBC and non-TNBC, then analyzed to determine which are the most accurate. Out of all four models, KNN was found to be the most accurate in classifying TNBC and non-TNBC. Overall, this study could help to understand basic knowledge of bioinformatics in cancer classification, but more accurate models are needed to better conduct disease diagnosis and prognosis.

Acknowledgment

We would like to thank Eric Lee, Wayne Fan, and Dongkyu Lee for their mentorship and support. We also thank the Undergraduate Research Committee for the opportunity to present in this research conference.

References

[1] DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Goding Sauer, A., Jemal, A., & Siegel, R. L. (2019). Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians, 69*(6), 438-451.
[2] Advanced machine learning approaches in cancer prognosis: Challenges and applications. Nayak J., Favorskaya M. N., Jain S., Naik B. and Mishra M.(Eds.). Springer International Publishing.
[3] Fei, H., Chen, S., & Xu, C. (2020). RNA-sequencing and microarray data mining revealing: The aberrantly expressed mRNAs were related with a poor outcome in the triple negative breast cancer patients. *Annals of Translational Medicine, 8*(6), 363-363.
[4] Naorem, L. D., Muthaiyan, M., & Venkatesan, A. (2019). Integrated network analysis and machine learning approach for the identification of key genes of triple-negative breast cancer. *Journal of Cellular Biochemistry, 120*(4), 6154-6167.