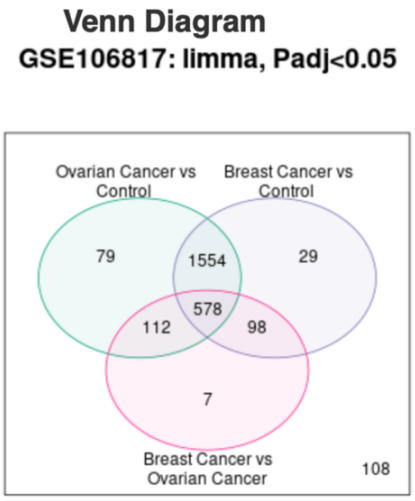
Section III: Results

The analysis of 4,253 samples provided a comprehensive understanding of the performance of various machine learning models as predictive tools for miRNA-based disease detection.

Differential Expression



Datasets selected from the Gene Expression Omnibus were processed and normalized for accurate analysis in the GEOR2 platform. The *limma* package and libraries from the Bioconductor Project in R-Studio were used to perform statistical tests such as t-tests and log2 fold change to identify which miRNAs are differentially expressed in each disease in comparison to the control sample. Two-sample t-tests were performed to determine which miRNAs are differentially expressed between multiple diseases and found common across all three. The Venn Diagram visualizes the distribution of the different groups and validates the claim that each disease has a unique miRNA profile that can be used for future classification models. Identifying the differentially expressed miRNAs through statistical tests will create a unique miRNA profile for each disease and help with the feature selection for the machine learning models. The miRNAs that are identified as most significant for each disease can be assigned greater weight to improve the accuracy of the predictive models. The table below highlights the top up and down regulated miRNA identified from each dataset. All the datasets were then aggregated into one file and performed the same tests to determine if there is a significant difference between the differential miRNAs identified.

Figure 1: The distribution of miRNAs that are unique and shared amongst breast cancer, ovarian cancer, and control in one dataset. The center highlights 578 miRNAs found across all three cases, 7 miRNAs to differentiate breast cancer from ovarian cancer, 79 to classify ovarian cancer from control, and 29 miRNAs to classify breast cancer from control.

Binary Classification

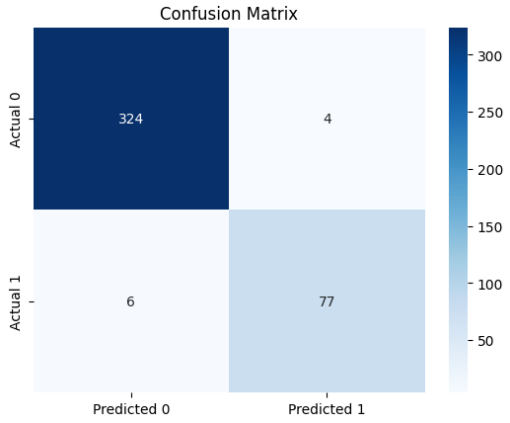


Figure 2: Confusion matrix of binary classification model for ovarian cancer set. Achieved an accuracy of 96%.

First, a logistic regression model was designed for binary classification as it is easy to implement and provides the coefficient of each predictor (Rout, 2020). The logistic regression models resulted in an accuracy of greater than 95%, which might indicate the occurrence of overfitting as logistic regression models are tend to overfit if the number of features is greater than the number of samples. Mitigating this issue would require selecting some feature over others, which could make the predictions biased.

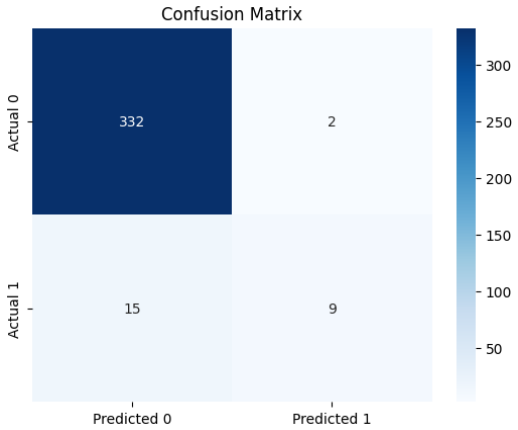


Figure 3: Confusion matrix of binary classification model for breast cancer set. Achieved an accuracy of 94%.

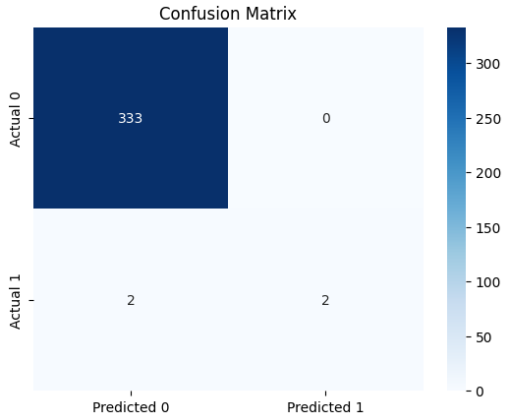


Figure 4: Confusion matrix of binary classification model for endometriosis set. Achieved an accuracy of 97%.

Therefore, a neural network was designed to perform dimensionality computations and overcome the limitations of a logistic regression model by automatically identify significant features. A 20-80 test and train split were selected, and the model was iterated through 50 epochs. The confusion matrices highlight the performance of the model on each of the diseases.

Feature importance algorithms were applied to determine which miRNA played a significant role in prediction. The findings of this step will be cross verified with the miRNA identified from the statistical tests to develop a robust miRNA panel for each disease.

Random Forest. A simple multiclass Random Forest algorithm was applied to classify 4 different types of gynecological conditions and control samples. The overall model achieved an accuracy of 92%, with the sub-class accuracy as described in Figure 5. The varying accuracy levels for each class validate that each disease contains a unique miRNA profile that can be harnessed to classify patient samples. It can also be hypothesized that Borderline Ovarian Tumor had the lowest accuracy due to its biological nature and miRNA expression levels being like control samples and ovarian cancer samples. The overall outcome of this aim is to develop machine learning models that can successfully classify multiple gynecologic diseases. Next, a Deep Neural Network was implemented, and each miRNA will be assigned a weight through feature extraction of the Random Forest model to improve the accuracy.

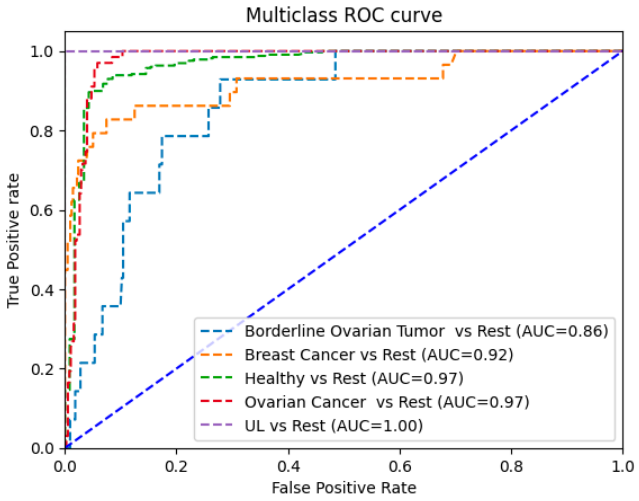


Figure 5: A Receiving Operator Curve (ROC) illustrates the performance of binary classification models. In this case, each line demonstrates the accuracy levels of classifying the condition out of all the possible outcomes in that dataset.

Deep Neural Network

The deep learning binary classification models were compiled into one model to predict multiple diseases. Multiclass predictive models provide an advantage in terms of time and feasibility over binary classification models in clinical settings due to their capabilities to predict several diseases at once. Several test and train splits were experimented with to produce the highest accuracy. After training it for 200 epochs, the model produced an accuracy of 85%.

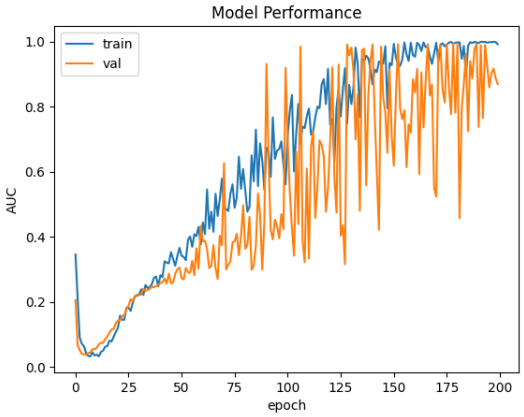


Figure 6: A model performance graph depicting the growth of accuracy as the model is trained over 200 epochs.

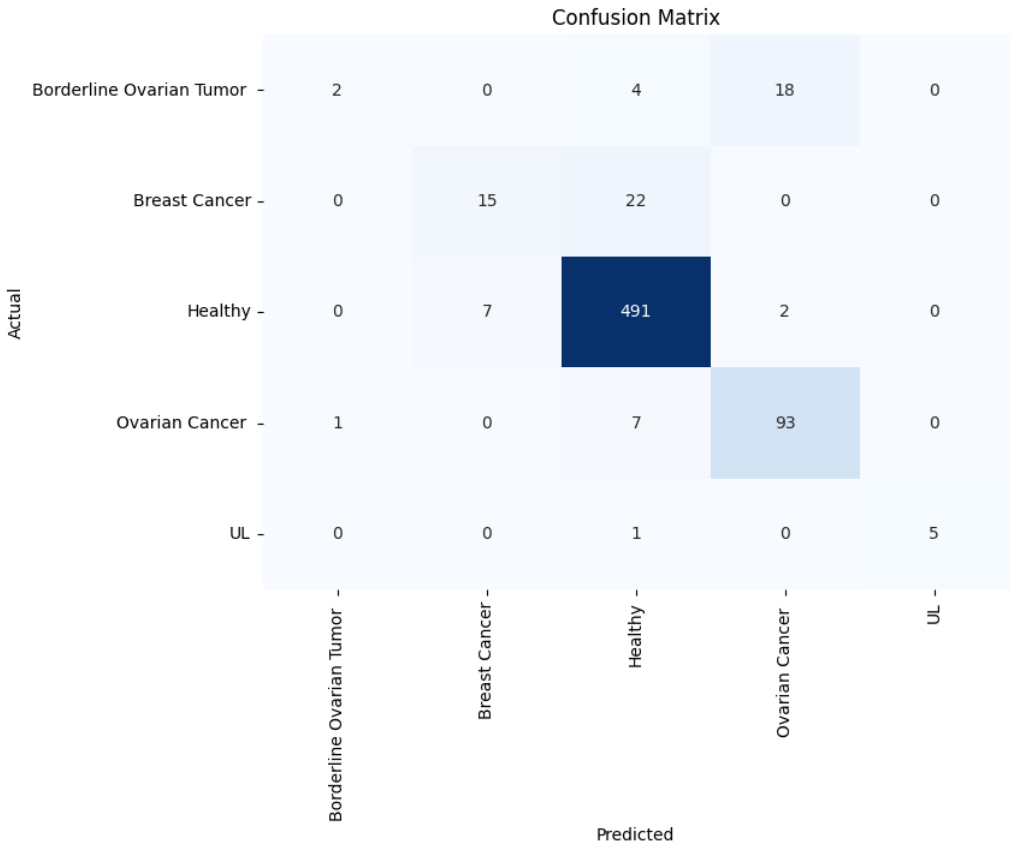


Figure 7: A confusion matrix highlighting the accuracy of each disease classification, and an overall model accuracy of 85%.

Pathway Modeling

The DeepLIFT Algorithm was applied to the Deep Neural Network model to determine the contribution of each miRNA to the prediction of each class. The miRNAs featured in Figure 8 highlight the most significant input features and each of their influence in predicting a certain disease in the multiclass model. These miRNAs can be cross verified by the unique miRNA panels found in previous works and be used to model the biological pathways in the three diseases being studied.

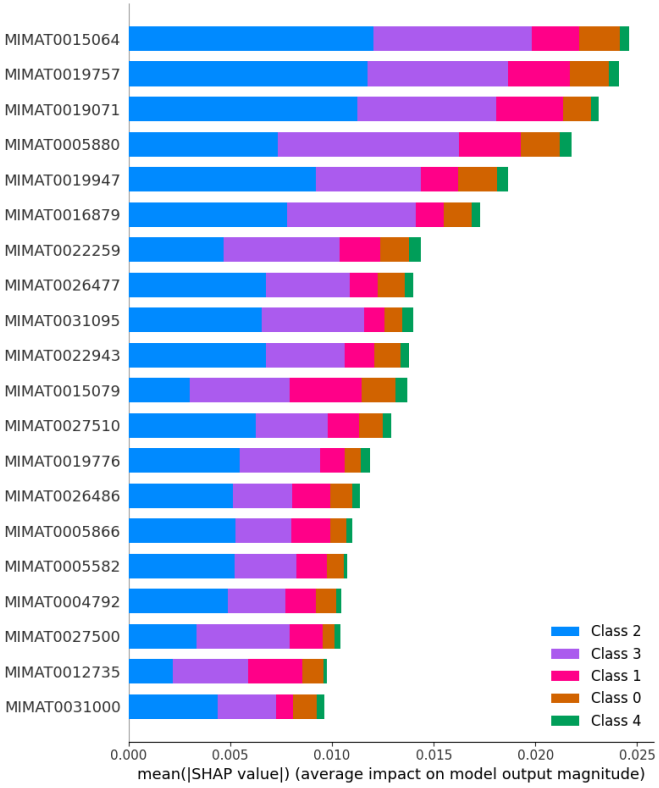


Figure 8: The Shapley values of the significant feature miRNA in the Deep Neural Network and their contribution to the prediction of each class.

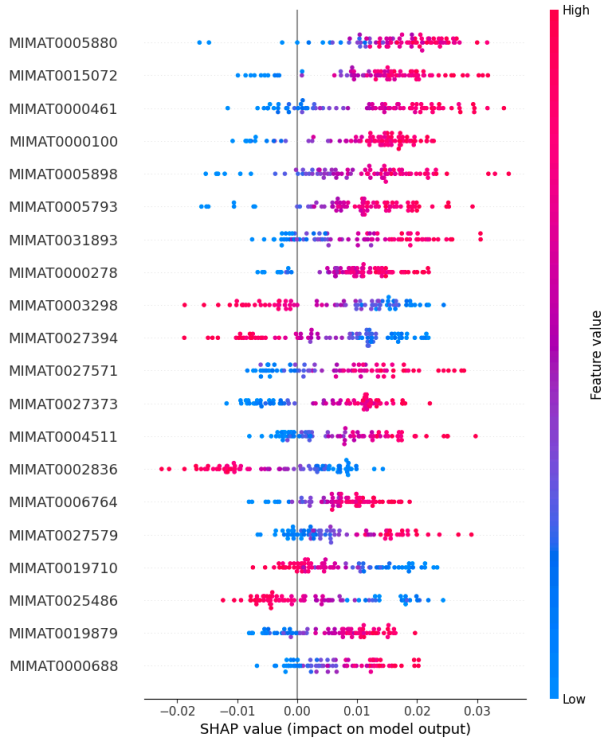


Figure 9 The Shapley values of the significant feature miRNA in the binary classification model of ovarian and their up and down regulation in comparison to the control. This algorithm was also applied on the breast cancer and endometriosis binary model and the miRNA, and their significance value can be found in Appendix 3.

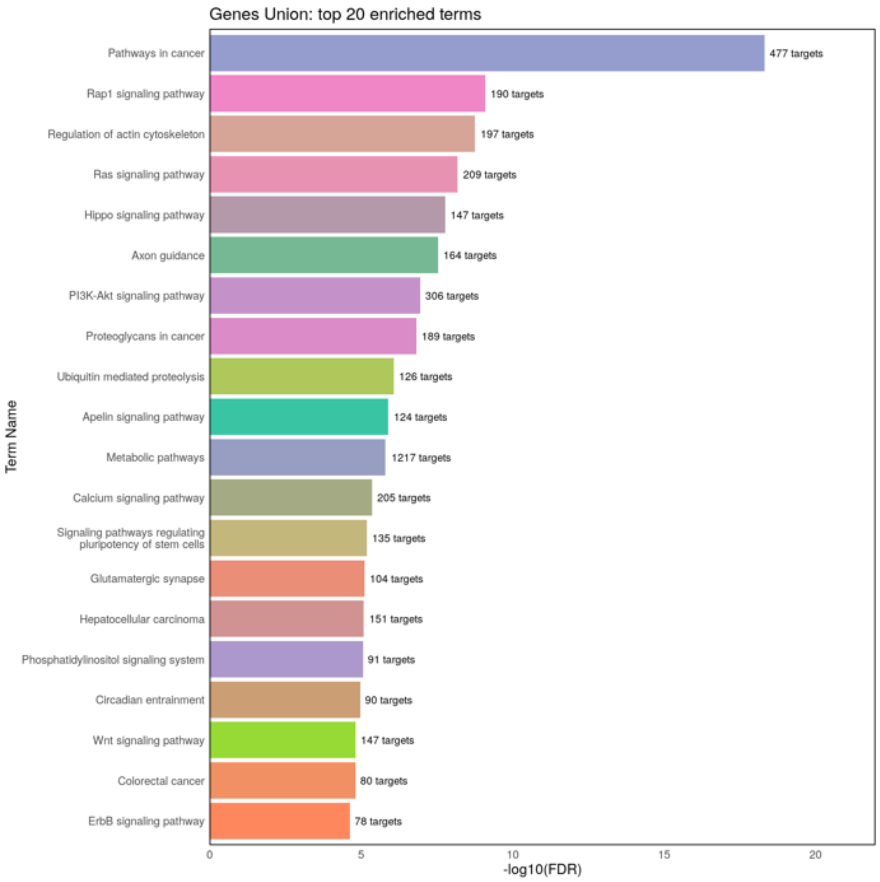


Figure 9: A bar graph showcasing the pathways of the genes targeted by the significant miRNAs identified by the machine learning models.

The extracted miRNAs were inserted into the DIANA/miRPath v.4 software, which models pathways influenced by specific miRNAs by applying statistical tests on data retrieved from the KEGG database. Figure 9 highlights the clusters of miRNAs and the biological pathways most impacted by the aberrant expression of these miRNAs.

Section IV: Discussion

This study demonstrates that miRNAs can be used as noninvasive candidates for disease detection and identifying therapeutic targets. In the preliminary stages of binary classification, the deep learning models provided high accuracy (over 90%) and reliability in comparison to the logistic regression model due to the neural network's stronger capabilities for processing high dimensional data. Although each dataset utilized different miRNA extraction techniques impacting the measured miRNAs and resulting in slightly different miRNA profiles for each disease than previous works, this pitfall was mitigated by compiling multiple datasets and using robust normalization methods. The resultant miRNA