Section II: Methodology

Role of Student vs. Mentor

For 5 months, I (the student) executed the methodology described below under the mentorship of Dr. Jill Moore, a professor of biocomputational studies at UMASS Chan Medical School. I was responsible for identifying appropriate datasets, selecting, and developing machine learning models, and drawing biological significance from the feature selection. Dr. Moore shared her expertise as a researcher in computational biology and provided support for accurately normalizing datasets, debugging errors in the code, and accurate interpretation of the results.

Equipment and Materials

Robust datasets were obtained from the National Institute of Health (NIH) Gene Expression Omnibus (GEO) to use miRNAs as a predictive candidate for gynecological diseases. The following keywords were used to select datasets: “ovarian cancer” OR “breast cancer” OR “endometriosis” AND “microRNA” or “miRNA” (See Appendix 1b). The NIH GEO2R platform was used to determine differentially expressed miRNA for each set. Next, machine learning models were developed in Google Collab using the following packages: Matplotlib, pandas, Scikit-learn, Seaborn, NumPy, Keras, and TensorFlow. The miRNA features determined from these stages were used to draw biological significance from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and the DIANA-miRPath v4.0 webserver. The software and platforms used in this study were adapted from previous works by Hamidi et al. (2023) and Zhang and Hu (2022) by modifying the types of machine learning models used, varying the train-test split.

Differentially Expressed miRNA.

This study analyzed samples from 4,253 women from several studies. The GEO2R uses GEOquery, limma, and DESeq2 R packages to perform statistical tests, normalize datasets of varying

sizes, and apply multiple testing corrections to output a table of p-values and log fold change for each miRNA in the dataset. The top differentially expressed miRNAs for each disease were extracted from the results based on the ranking of their p-values in the up and down-regulated miRNAs (See Appendix 3). The findings of this stage will serve a greater purpose in the final stage of drawing biological connections between different diseases.

Environment Packages

The TensorFlow machine learning frameworks were used for preprocessing data, fitting/training, and evaluating the models. The “Python 3” programming language was used to process the data, and several packages of the version compatible with Python 3.12.1 were used to perform various functions: “Pandas” for processing CSV and table-like data, “NumPy” for processing number and data arrays, “Matplotlib” for visualizing graphs and plots, “Seaborn” for visualizing data distribution with heatmaps, and “Sci-kit Learn” for essential machine learning functions such as training and splitting data.

Machine Learning

Machine learning algorithms have proved effective as predictive models for cancer (Alharbi & Vakanski, 2023). Given the high dimensionality of the miRNA data, multiple machine-learning models were developed to evaluate their efficiency. First, logistic regression and deep learning models were trained in Python to perform binary classification between one disease and healthy samples. Next, the model was validated on a different dataset, and three train-test splits (20-80, 30-70, and 40-60) were tested to determine which provided the most accurate results. Once the binary classification models were validated and evaluated, multi-classification models were developed using Random Forest and a Deep Neural Network, based on the dimensionality of miRNA datasets and the strong performance of such models in previous studies (Hamidi et al., 2023). Feature extraction techniques were used to draw out the miRNA profiles that play a significant role in classifying each disease. The Deep Neural Network

and the GEOR2 results were cross validated before being entered into the Kyoto Encyclopedia of Genes and Genomes (KEGG) database for pathway modeling.

Evaluation of Models

The following metrics were evaluated for each model to compare the efficiency: Accuracy, Sensitivity, and Specificity. Each metric is calculated using the following case results:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Sample}} \quad (1)$$

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

$$\text{Specificity} = \frac{\text{True Negativite}}{\text{True Negative} + \text{False Positive}} \quad (3)$$

T-test

For this study, a 2-sample t-test was used to measure the difference in average accuracy between the DNN and Random Forest models by running the model ten times and recording the accuracies to determine if they are statistically significant. The test yielded a t-test statistic of 3.8027, corresponding to a p-value of 0.01, below the widely accepted threshold of . Therefore, the null hypothesis was rejected, and there is a statistically significant change in the accuracy of the Deep Learning Model and the Random Forest. Therefore, only the Deep Learning model was used to analyze the significant features as it performed with higher accuracies.

Pathway Modeling

Next, the DeepLIFT algorithm was applied to the deep learning model to extract the top miRNA profiles that played a key role in predicting each disease. The Deep Learning Model was selected over the Random Forest model due to its ability to automatically determine significant features and prevent weight bias with high-dimensional datasets (Alharbi & Vakanski, 2023). A panel of 60 miRNAs for each disease was developed through feature extraction and cross-validation from previous works (Sivajohan et al., 2022). A panel of 20 shared miRNAs was determined and used to model shared pathology and miRNA-mediated pathways through DIANA-miRPath v4.0 and KEGG databases.