Using MicroRNAs and Deep Learning to Noninvasively Diagnose Gynecologic Conditions

Palak Yadav

Massachusetts Academy of Math and Science

Advanced STEM with Scientific and Technical Writing

Instructor: Kevin Crowthers, Ph.D.

Worcester, MA. 01605

Abstract

Women's health is a growing public health crisis. 1 out of 10 women will experience some type of chronic gynecological disease in their lifetime, yet confirmative diagnosis can take up to 5-7 years, due to the dismissal of the symptoms, lack of access to adequate resources, and social stigma. MicroRNAs are segments of non-coding RNA that play a vital role in gene expression and have shown as a noninvasive diagnostic candidate detected in bodily fluids. The goal of this study is to evaluate serum miRNA expression levels to predict gynecologic conditions, specifically ovarian cancer, breast cancer, and endometriosis. A predictive machine learning model was trained on public miRNA datasets to generate unique miRNA profiles for each condition. Significant miRNAs in the predictive model were extracted through feature selection techniques and inserted in pathway modeling software to determine the pathways most affected in each disease, with *[insert prevalent and unique miRNAs]*. By identifying the unique and shared pathology of ovarian breast cancer, and endometriosis, miRNA prevalence can be used as a noninvasive diagnostic tool, potentially reduce waiting periods, and guide future therapeutic development.

Keywords: machine learning, microRNAs, gynecology, endometriosis, ovarian cancer, breast cancer, diagnosis

Acknowledgments

I would like to extent my deep gratitude to Dr. Jill Moore and Dr. Kevin Crowthers for their guidance and mentorship on this project, and all the seniors and STEM professionals who provided their thoughts and insights throughout the process.

Using MicroRNAs and Deep Learning to Noninvasively Diagnose Gynecologic Conditions.

Women's health is a growing public health issue due to a lack of funding, accessibility, and effective diagnostic tools. Unfortunately, 1 out of 10 women will experience a chronic gynecological condition in their lifetime, yet the average diagnostic period is about 5-7 years due to the dismissal of the symptoms, lack of access to adequate resources, social stigma, and other factors (Ahn et al., 2017). These numbers are even more dire for African American, Hispanic, and other marginalized communities. There is an urgent need to address this inequity in healthcare and provide female-identifying patients with the right services at the right time.

Current State of Gynecological Diseases

An analysis of the National Institute of Health (NIH) funding reported that conditions that negatively affect women receive significantly less funding in proportion to the burden they exert on individuals and society at large (Smith, 2023). For example, ovarian cancer ranks 5th for lethality in a selection of 19 prevalent cancers, yet it ranks 12th in terms of funding. This discrepancy in funding and research has limited the availability of effective and accessible screening tools and therapeutics, resulting in 80% of the ovarian cancer cases being diagnosed at an advanced stage (Smith, 2023; Mogensen et al., 2016). The lack of funding coupled with the challenges of the vagueness of symptoms, lack of awareness, understanding of the pathology, and accessibility to resources, makes it challenging to diagnose detrimental gynecologic conditions at the right time. This study aims to better understand the pathology and etiology of three major gynecological conditions - breast cancer, ovarian cancer, and endometriosis - and work towards developing a noninvasive diagnostic tool as the current screening practices generally involve laparoscopy and biopsies for confirmative diagnoses, and there is a need to find noninvasive ways to diagnose conditions such as endometriosis and ovarian cancer (Ginsburg et al., 2017)

Breast Cancer, Ovarian Cancer, and Endometriosis

Endometriosis results from the abnormal growth of the uterine lining and may contribute to ovarian cancer and breast cancer. Many studies have speculated a correlation between these three detrimental conditions; however, the results have been varied (Mogensen et al., 2016). A special link between breast and ovarian cancer has been established through the mutations in the BRCA1 and BRCA2 genes (Yoneda et al., 2011). Endometriosis is often managed with oral contraceptives and hormonal therapy, which can put one at a higher risk for ovarian cancer due to the imbalance of hormone levels. The correlation between breast cancer and endometriosis has been unreliable, as some studies indicate a positive relationship, while others report either the opposite or no significant difference (Ye et al., 2022). There is a need to better understand the correlation between these three conditions beyond just population studies, as a molecular understanding is key to identifying potential therapeutics and strategies to diagnose them effectively.

TABLE 1 | miRNA signaling pathways involved in gynecological cancers.

miRNA	Signaling pathway	Target	Target expression	Action	Pathology	Reference
miR-433	MAPK	RAP1A	Overexpression	Cell migration, proliferation, apoptosis	Breast cancer	(76)
miR-99a	mTOR	PI3-AKT	Overexpression	Invasion, proliferation, apoptosis	Cervical cancer	(77)
	FGFR3				Breast cancer	
miR-155	AKT	LKB1	Overexpression	Autophagy	Cervical cancer	(78)
miR-21	TNFR1	Caspase 3	Overexpression	Apoptosis	Breast cancer	(79)
	PI3K/AKT/mTOR	TNF-alpha			Cervical cancer	
	RAS/MEK/ERK	PTEN			Ovarian cancer	
		RASA1				
miR-200	NOTCH	ZEB1 ZEB2	Overexpression	Invasion, metastasis	Ovarian cancer	(80)
miR-141	TGF-beta	E cadherin				
miR-200a		EMT				
miR-200b						
miR-200c						
Let-7	RAS	P53	Overexpression	Apoptosis	Ovarian cancer	(81)
*miR let-7d-5p	HGMA1					
miR-34a	p53	HNRNPA1		Cell proliferation	Breast cancer	(82)
					Endometrial cancer	
miR-424	p53	HNRNPA1	Overexpression	Cell proliferation, apoptosis	Breast cancer	(82)
miR-503	p53	HNRNPA1	Overexpression	Cell proliferation, apoptosis	Breast cancer	(82)
miR-142-3p	Bach-1	EMT	Overexpression	Invasion, migration	Breast cancer	(83)
miR-205	ZEB1, ZEB2	EMT	Overexpression	Apoptosis, cell differentiation, and proliferation	Endometrial cancer	(84)
		PTEN				
miR 4712-5p	PTEn/AKT/GSK3beta/cyclin D1	PTEN	Overexpression	Cell invasion, metastasis	Vulvar cancer	(85)
miR-3147	TGF-β/Smad	TGFβ RII	Overexpression	Invasion, cell proliferation, migration	Vulvar cancer	(86)
		EMT				
miR-146a		BRCA1	Overexpression	Cell proliferation	Breast cancer	(87)

ZEB1 and ZEB2 Zinc finger E-box-binding homeobox 1/2 HNRNPA1 Heterogeneous nuclear ribonucleoprotein A1.

Table 1: miRNA signaling pathways involved in gynecological cancers (Duică et al., 2020). Some miRNAs can be observed in multiple conditions, demonstrating the need to identify unique markers for improved diagnosis.

MicroRNA

Given the current state of gynecologic conditions, there is a need to find a noninvasive way to diagnose such diseases quickly and accurately. Tracing gene expression data as indicators for diseases is an emerging area of research. RNA-seq data examines large datasets of RNA

extracted from patient samples of blood, saliva, urine, and other bodily fluids. MicroRNAs are 22 nucleotides-long non-coding segments of the RNA. They derive from the transcription of the DNA, and the interaction between 3' untranslated regions (3' UTR) of target mRNAs is regulated by the interaction between various miRNA transcriptional factors to either suppress or express specific genes (J. Gilabert-Estelles et al., 2012). They play a significant role in gene expression as mutations in a few miRNAs can have a cascading effect on mRNA transcription and protein production, leading to dysregulation in numerous biological pathways and signaling (O'Brien et al., 2018). MicroRNAs are easy to analyze due to their abundance and accessibility in bodily fluids and hold promise as noninvasive diagnostic candidates. Many previous works have found that certain diseases have unique miRNA expression levels in comparisons to other samples, however many miRNAs are found across several diseases and there is a need to discover miRNA specific to each gynecologic disease as well as evaluating biological interconnectedness from miRNAs found significantly across several diseases (Zhao et al., 2014).

Machine Learning

Machine learning models can be trained on large sets of patient miRNA data to build predictive models. Several studies have attempted to use this approach for classifying various types of cancers and have proved successful in classifying diseases based on their miRNA expression (Alharbi & Vakanski, 2023). Unlike previous works, this study aims to target gynecologic conditions specifically, and evaluate biological significance between three prevalent diseases. Having a greater understanding of the miRNAs involved in gynecological conditions and their expression levels will shed light on their unique and shared pathology. All of this will help work towards identifying promising diagnostic and therapeutic targets.

Problem Statement: Diagnosing gynecological conditions quickly and effectively is a challenge due to the lack of funding, knowledge gaps, and accessibility to screening tools.

Research Question: How can miRNA expression data and machine learning be used as a diagnostic candidate and provide a greater understanding of the pathology and etiology of three major gynecological conditions: ovarian cancer, breast cancer, and endometriosis?

Objective:

Obj. 1: Collect miRNA expression samples associated with the target diseases and healthy control to train the machine learning models. For the scope of this project, the diseases will include breast cancer, ovarian cancer, and endometriosis.

Obj. 2: Design a deep learning binary classification model to differentiate miRNA expression levels of a specific disease samples and control. The goal is to achieve greater 80% accuracy to prove that each disease has a unique miRNA expression profile, and use determine the significance levels (p-value) of the top feature miRNA to identify miRNA unique to each disease.

Obj. 3: Implement multiclass predictive models to classify all the target diseases with an accuracy of at least 75% and extract a panel of miRNA to be significant found across all three diseases.

Obj. 4: Use the miRNA profiles developed to model miRNA-mediated pathways and draw biological significance about the pathology and etiology of the three diseases.

Hypothesis

Given the effectiveness of miRNA expression data and deep learning models in classifying various of types of cancer in previous works, it can be hypothesized that deep learning models will produce accurate classification of gynecologic diseases for early disease prognosis and identifying potential therapeutic targets.

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. 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 the datasets, debugging errors in the code, and accurately interpretation the results.

Equipment and Materials

To use miRNAs as a predictive candidate for gynecological diseases, robust datasets were obtained from the National Institute of Health (NIH) Gene Expression Omnibus (GEO). 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 miRNAs 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 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 preprocess the data and several packages 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 important machine learning functions such as training and splitting data.

Machine Learning

Machine learning algorithms have proved to be 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 used to determine which one proves proved the most effective. Once the binary classification

models were validated and evaluated, multi-classification models were developed using Random Forest and a Deep Neural Network given the dimensionality of the miRNA datasets and the strong performance of such models in previous studies (Hamidi et al., 2023). Feature extraction techniques be used to draw out the miRNAs that play a significant role in the classification of each disease. The findings of the Deep Neural Network and the GEOR2 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:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Sample}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negativite}{True\ Negative + False\ Positive}$$

T-test

For this study, a 2-sample t-test can measure if the difference between the average accuracy of two models is statistically significant. Therefore, a 2-sample t-test was used to compare the performance of the multiclass of Deep Neural Network and the Random Forest model, by applying the test on the accuracies for 5 trials. The test yields a t-test statistic of 3.8027, corresponding to a p-value of 0.01, which is below the widely accepted threshold of $a \leq 0.05$. Therefore, the null hypothesis is rejected, and there are statistically significant changes in the accuracy of the Deep Learning Model and the Random Forest.

Pathway Modeling

Next, the DeepLIFT algorithm was applied on the deep learning model to extract the top significant feature miRNA that played a key role in the prediction of 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. A panel of 60 miRNAs for each disease was developed through feature extraction and cross-validation from previous works. A panel of 20 miRNAs were found across all three disease and used to model their shared pathology. The DIANA-miRPath v4.0 accepts inputs of up to 200 miRNAs and is linked to the KEGG databased to model miRNA-mediated pathways and gene unions.

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

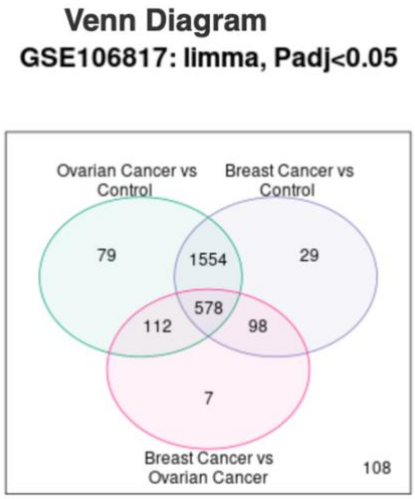


Figure 1: The distribution of miRNAs that are unique and shared amongst breast cancer, ovarian cancer, and control in one dataset. The center highlights 578 miRNAs found across all three cases, 7 miRNAs to differentiate breast cancer from ovarian cancer, 79 to classify ovarian cancer from control, and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

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.

Binary Classification

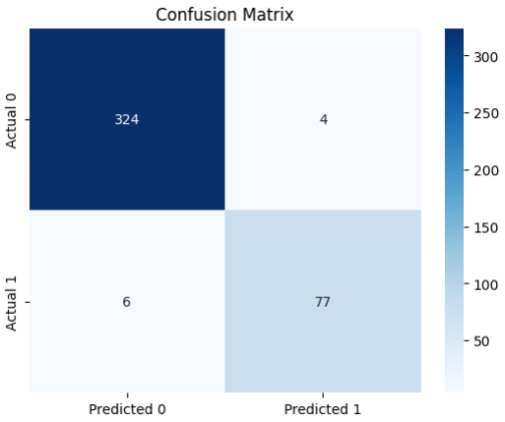


Figure 2: Confusion matrix of binary classification model for ovarian cancer set. Achieved an accuracy of 96%.

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.

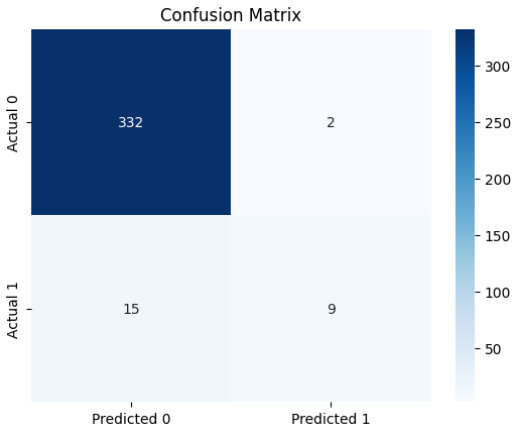


Figure 3: Confusion matrix of binary classification model for breast cancer set. Achieved an accuracy of 94%.

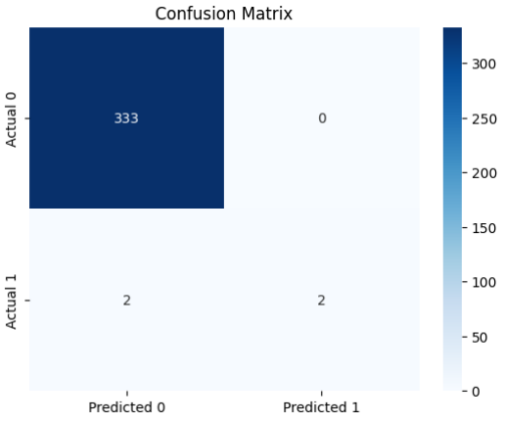


Figure 4: Confusion matrix of binary classification model for endometriosis set. Achieved an accuracy of 97%.

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.

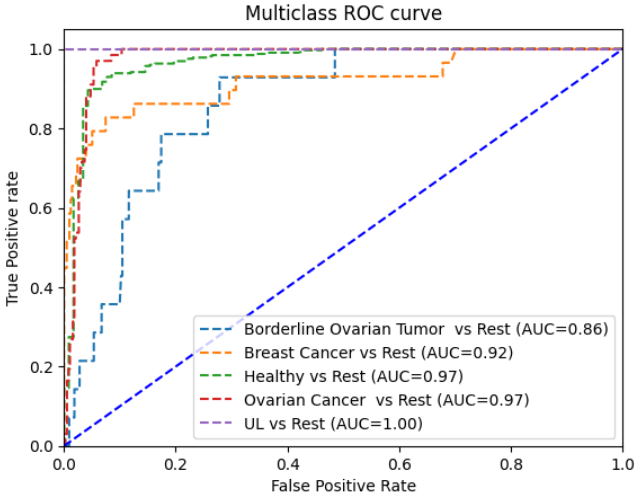


Figure 5: A Receiving Operator Curve (ROC) illustrates the performance of binary classification models. In this case, each line demonstrates the accuracy levels of classifying the condition out of all the possible outcomes in that dataset.

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%.

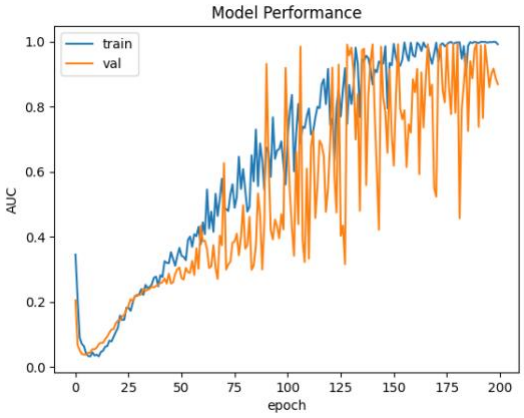


Figure 6: A model performance graph depicting the growth of accuracy as the model is trained over 200 epochs.

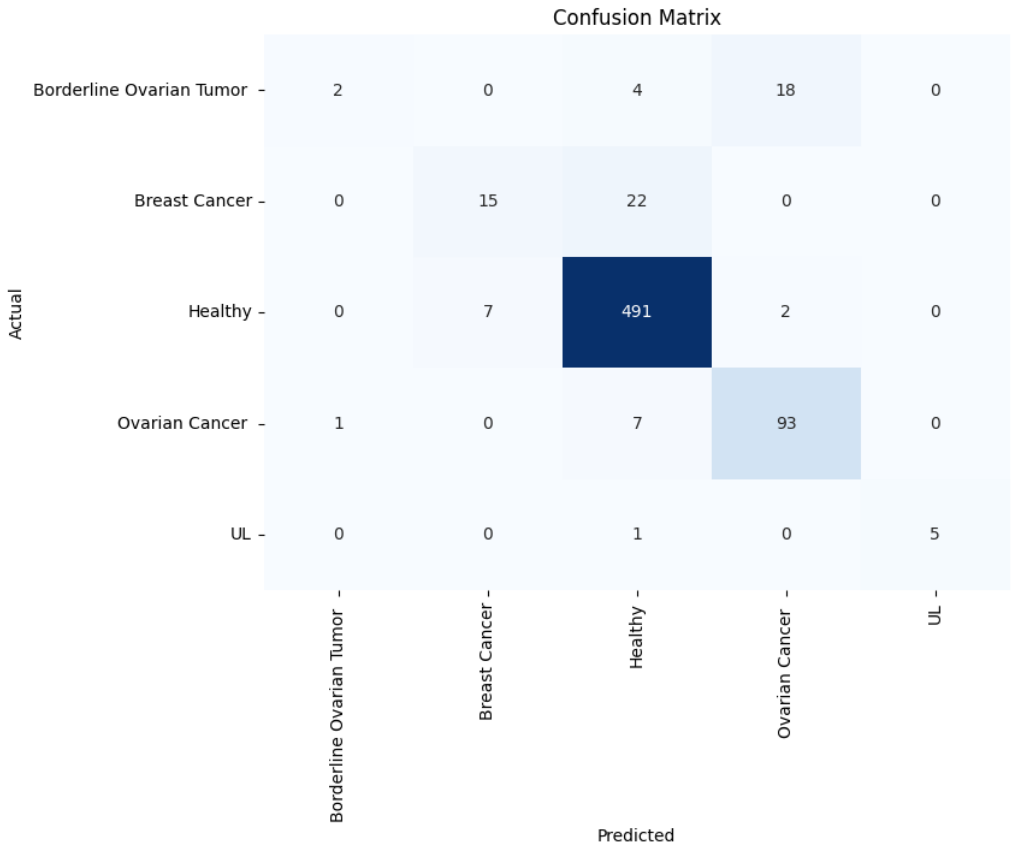


Figure 7: A confusion matrix highlighting the accuracy of each disease classification, and an overall model accuracy of 85%.

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.

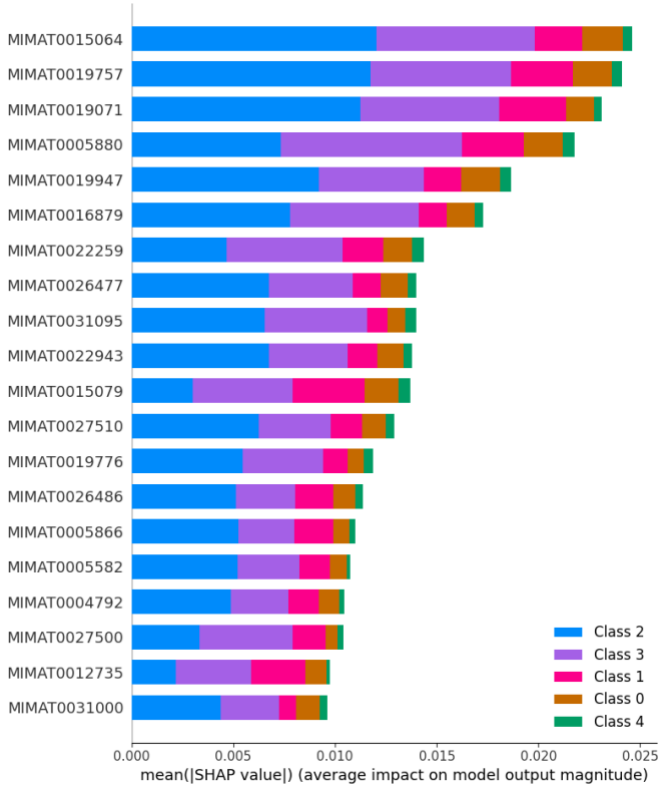


Figure 8: The Shapley values of the significant feature miRNA in the Deep Neural Network and their contribution to the prediction of each class.

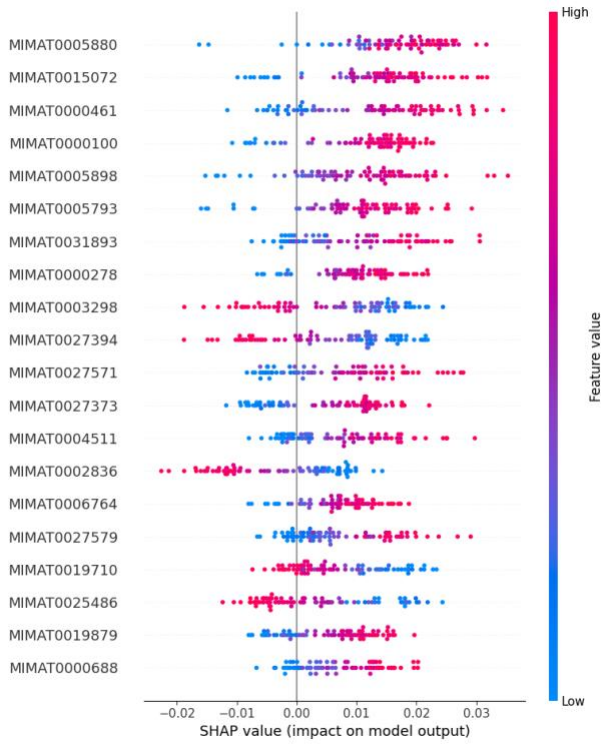
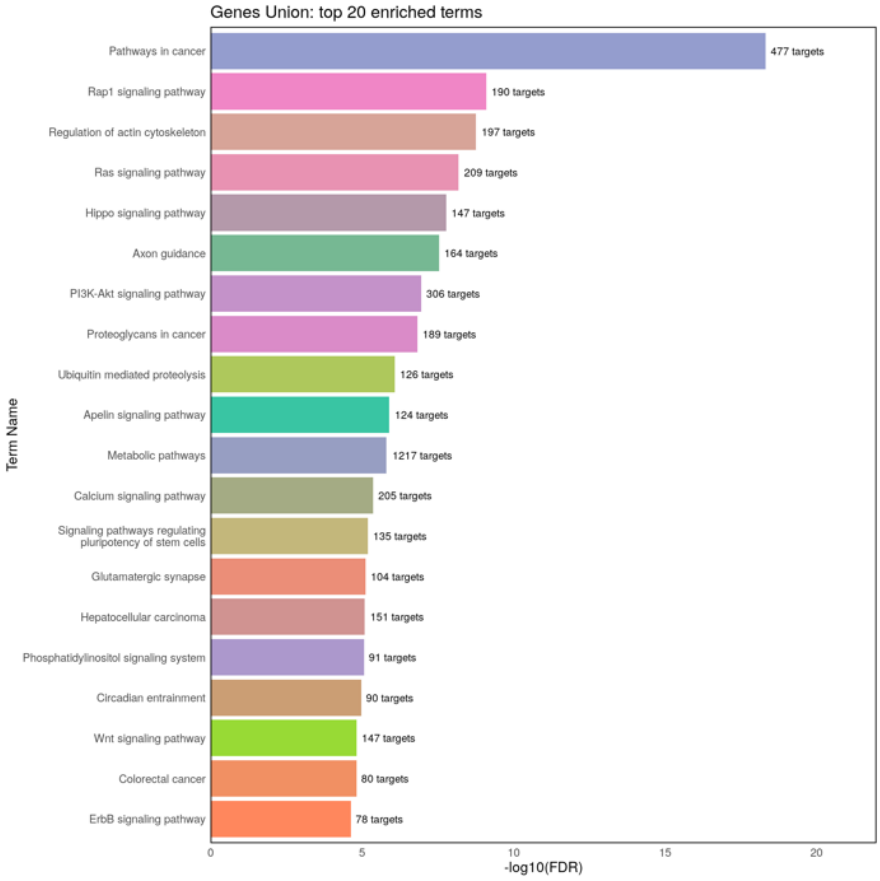


Figure 9 The Shapley values of the significant feature miRNA in the binary classification model of ovarian and their up and down regulation in comparison to the control. This algorithm was also applied on the breast cancer and endometriosis binary model and the miRNA, and their significance value can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



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.

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

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

profiles were cross validated with works to ensure that there is a consistency of the differentially expressed miRNAs identified and evaluate the accuracy of the new miRNAs identified in this study. Mitigating potential discrepancies in the miRNA expression profiles is key as the machine learning model and ultimately the pathway analysis rely on the accuracy of the miRNA panels. The aggregated multiclass deep learning model performed with an accuracy greater than 80%. Although the accuracy of the Deep Neural Network was lower than Random Forest, Deep Neural Networks are better apt at balancing complex data, whereas Random Forest are prone to bias prediction due to its inability to accurately balance class sizes. Overall, the models developed performance the same or better than previous models used for miRNA expression levels analysis (Alharbi & Vakanski, 2023). Lastly, the Deep LIFT algorithm was applied on the deep neural network's top features and their contribution to the classification of each disease. The miRNA panel extracted from each of these three models provided a unique profile for each disease as well as insights into their shared pathology. Comparing the panels of miRNAs identified in this study to previous works reveals some similarity and some new miRNAs that should be explored more in future experiments.

Biological significance

The pathway modeling software miRNA-DIANA Tools accepts at most 200 miRNA entries. Approximately 60 miRNAs from the unique disease panel, and 20 miRNAs from the shared panel were used to analyze the miRNA mediated pathways. miRNA expressed across all three diseases targeted genes found in the pathways of cancer (P=6.04E-18), followed by the Axon Guidance, regulation of the actin cytoskeleton, and the RAP1 signaling pathway, all with significance levels of lower than P=3.15E-08. A PubMed search of the miRNA targeting these pathways revealed the interconnected pathology of these three diseases. miRNA-Let-7d was found across all three diseases and its aberrant expression has shown to play a key role in female malignancy (Zhang, et. al, 2017). The human lethal-7 (let-7) family of

miRNA are found on chromosome 9 and mediate cell proliferation and carcinogenesis. The expression of this family of micro-RNAs are regulated by the transcriptional and post-transcriptional through the OCT4, MYC, and p53 mutations, and the aberrant expression of let-7d ultimately targets mRNAs involved in various tumor hallmarks as well as conditions such as endometriosis and uterine fibroids (Zhang, et. al, 2017). miRNA-let-7d sheds light on the interconnected pathology of these three diseases, indicating that its aberrant expression could predict the presence of one disease and the risks of developing another gynecologic disease.

The under expression of miR-320a was found to be prevalent in breast cancer samples. The miR-320 family has shown to be linked with the EMT process, decreasing the expression of E-cadherin, and increasing the expression of N-cadherin through the targeting of the *FOXM1* gene and other signaling pathways such as P13K/AKT and TGF-β/Smad signaling (Liang et al., 2021). Targeting the under expression of miR-320a in breast cancer would inhibit the migration and invasiveness of the tumor.

miR-1307-3p upregulation was found with high significance in the ovarian tumor samples (p-value < 10e-8) and slightly lower yet still significance levels in breast cancer and endometriosis. The miR-1307 family of microRNAs are found in the USMG5 gene intron region in the chromosome 10, yet there is little understanding of its functionality (Saberianpour & Abkhooie, 2021). Upregulation of this class of miRNAs has been linked to the chemoresistance in ovarian tumor, as well as abnormal cell growth, differentiation, and metastasizes (Saberianpour & Abkhooie, 2021). Targeting the genes mediated by miRNA-1307 could serve as ovarian cancer therapy, and further exploration of its role in other female malignancies could improve survival for a wider range of diseases.

These are just a few examples of pathways and signal mechanisms that demonstrate the interconnected pathology of breast cancer, ovarian cancer, and endometriosis. All these conditions exhibit abnormal behavior in pathways associated with angiogenesis, rapid cell growth, infertility, and

insulin levels, indicating that having one of these diseases increases the risk of getting the other two diseases.

Future Research

Future investigation should focus on collecting larger sample tests using similar miRNA extraction methods to improve generalizability and reduce misinterpretation by collecting samples from women of different ethnicity, age, and stages of mensuration cycle. To test the effectiveness of serum-based miRNA, healthcare workers should perform miRNA extraction methods on samples of patients and compare the prediction of machine learning model to current diagnostic and screening practices. Furthermore, the miRNA-mediated pathways and genes identified through this approach should be targeted through *in vitro* and *in vivo* samples to evaluate their effectiveness in suppressing the disease. While this study focused primarily on evaluating deep learning and random forest algorithms as predictive models, ensemble learning models aggregate multiple types of models to improve performance. By collecting larger sets of samples and experimenting with various kinds of machine learning models, miRNAs can be used for prognosis, therapeutics targets, and advancing understanding of the genetic and epigenetic factors of more than just gynecologic diseases. Implementing this diagnostic technique in clinics will expediate diagnostic periods by predicting an array of diseases through simple blood tests, and ultimately provide individuals with the care they need at the right time. This technology will be especially beneficial for marginalized communities who lack access to adequate medical care, and provided personalized therapies as each the pathology and etiology of each disease may vary patient to patient.

Section V: Conclusion

This study demonstrates the potential use of miRNAs as non-invasive diagnostic markers for breast cancer, ovarian cancer, and endometriosis. Deep learning binary classification and multi-class models were designed to predict the like hood of a disease based on a patient's serum miRNA expression profile which were selected from publicly available datasets. The models provided strong accuracy, sensitivity, and specificity in predicting each of the diseases, as well as the two stages of ovarian cancer (ovarian tumor and borderline tumor), demonstrating the strengths of machine learning as predictive models in early detection. Feature extraction techniques applied to the deep learning models provided panels of unique and shared miRNAs found across the three selected diseases. miRNA-Let-7 was found across all three diseases and its dysregulation plays a key role in tumor growth and onset of several cancer hallmarks. Overall, miRNAs found across all three diseases primarily targeted pathways of cancer, RAP1, RAS, and Hippo signaling, all of which have been spectacled by previous studies in malignancies and found to be a cause of degradation of benign conditions such as endometriosis. These findings pave the way for developing miRNA-based diagnostics and personalized therapies, ultimately improving patient outcomes.

Section VI: Appendix

Appendix 1a.

A list of programming languages, applications, and libraries required to build, train, and test the intended machine learning models.

Applications	**Description**
Google Collaboratory Python 3	A product from Google Research that allows for accessible and efficient Python code with support for important libraries and functions for machine learning models.
TensorFlow/Keras	A Python-based framework for building, training, and testing machine learning models
Matplotlib	Python library for visualization and graphics.
Sci-Kit Learn	Python-based machine learning library that consists of commands and functions to create a wide range of models.
R/R-Studio	An integrated platform to use R and R-based packages and libraries for statistical analysis.
GEO2R	A web-based NCBI Gene Expression Omnibus (GEO) analytical tool with an in-built limma package, DESEQ-2 commands, and other Bioconductor projects to identify differentially expressed genes and miRNAs.

Appendix 1b.

A list of datasets selected from Gene Expression Omnibus and the Cancer Atlas TCGA used for the training and testing phase.

GEO Access Link	Description
GSE106817	333 ovarian cancers, 66 benign tumors, 29 of ben ovarian, 143 breast cancer, and 275f non-car controls.
GSE235525	34 high-grade serous ovarian cancer, 36 samples
GSE201712	64 ovarian cancer samples
GSE226445	350 samples of women with known BRCA mutation, which is known to cause breast and ovarian cancer and 30303 wild types.
GSE230956	4 endometriosis samples and 4 benign samples.
GSE113486	100 ovarian cancers and 100 breast cancer samples.

Appendix 2.

A decision matrix to evaluate the model in comparison to previous studies' models (Alharbi & Vakanski, 2023).

Criteria	Rank	Logistic Regression	Expected %	Random Forest	Expected %	Neural Network	Expected %
AUC	8	9	95%	9	95%	9	95%
Accuracy	9	9	80%	8	90%	9	87%

Sensitivity	8	7	82%	7	88%	8	90%
Specificity	7	7	85%	8	83%	8	90%
Total		**258**		**256**		**273**	

Appendix 3. Access to all the code for the machine learning model and the panel of miRNAs identified:

https://github.com/palak0503/STEM

Section VII: References

Ahn, S. H., Singh, V., & Tayade, C. (2017). Biomarkers in endometriosis: Challenges and opportunities. Fertility and Sterility, 107(3), 523–532. https://doi.org/10.1016/j.fertnstert.2017.01.009

Alharbi, F., & Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. Bioengineering, 10(2), 173. https://doi.org/10.3390/bioengineering10020173

Bornehag, C.-G., Sundell, J., Weschler, C. J., Sigsgaard, T., Lundgren, B., Hasselgren, M., & Hägerhed Engman, L. (2004). The association between asthma and allergic symptoms in children and phthalates in house dust: A nested case–control study. Environmental Health Perspectives, 112(14), 1393–1397. https://doi.org/10.1289/ehp.7187

Centers for Disease Control and Prevention. (2019, November 20). Leading causes of death-females-all races/origins. https://www.cdc.gov/women/lcod/2017/all-races-origins/index.htm

Cook, R. J., & Dickens, B. M. (2014). Reducing stigma in reproductive health. International Journal of Gynecology & Obstetrics, 125(1), 89–92. https://doi.org/10.1016/j.ijgo.2014.01.002

Duică, F., Carmen Elena Condrat, Cezara Alina Dănilă, Andreea Elena Boboc, Radu, M., Xiao, J., Li, X., Sanda Maria Crețoiu, Suciu, N., Dragoş Creţoiu, & Dragoş Predescu. (2020). MiRNAs: A powerful tool in deciphering gynecological malignancies. Frontiers in Oncology, 10. https://doi.org/10.3389/fonc.2020.591181

Gilabert-Estelles, J., Braza-Boils, A., Ramon, L. A., Zorio, E., Medina, P., Espana, F., & Estelles, A. (2012). Role of microRNAs in gynecological pathology. Current Medicinal Chemistry, 19(15), 2406–2413. https://doi.org/10.2174/092986712800269362

Ginsburg, O., Bray, F., Coleman, M. P., Vanderpuye, V., Eniu, A., Kotha, S. R., Sarker, M., Huong, T. T., Allemani, C., Dvaladze, A., Gralow, J., Yeates, K., Taylor, C., Oomman, N., Krishnan, S., Sullivan, R., Kombe, D., Blas, M. M., Parham, G., & Kassami, N. (2017). The global burden of women's

cancers: A grand challenge in global health. The Lancet, 389(10071), 847–860. https://doi.org/10.1016/s0140-6736(16)31392-7

Hamidi, F., Gilani, N., Reza Arabi Belaghi, Hanif Yaghoobi, Esmaeil Babaei, Parvin Sarbakhsh, & Jamileh Malakouti. (2023). Identifying potential circulating miRNA biomarkers for the diagnosis and prediction of ovarian cancer using machine-learning approach: Application of Boruta. Frontiers in Digital Health, 5. https://doi.org/10.3389/fdgth.2023.1187578

Horne, A. W., & Missmer, S. A. (2022). Pathophysiology, diagnosis, and management of endometriosis. BMJ, 379, e070750. https://doi.org/10.1136/bmj-2022-070750

Li, X., Dai, A., Tran, R., & Wang, J. (2023). Identifying miRNA biomarkers for breast cancer and ovarian cancer: a text mining perspective. *Breast Cancer Research and Treatment*, *201*(1), 5–14. https://doi.org/10.1007/s10549-023-06996-y

Li, X., Dai, A., Tran, R., & Wang, J. (2023b). Identifying miRNA biomarkers for breast cancer and ovarian cancer: a text mining perspective. *Breast Cancer Research and Treatment*, *201*(1), 5–14. https://doi.org/10.1007/s10549-023-06996-y

Liang, Y., Li, S., & Tang, L. (2021). MicroRNA 320, an Anti-Oncogene Target miRNA for Cancer Therapy. Biomedicines, 9(6), 591. https://doi.org/10.3390/biomedicines9060591

Mogensen, J. B., Kjær, S. K., Mellemkjær, L., & Jensen, A. (2016). Endometriosis and risks for ovarian, endometrial and breast cancers: A nationwide cohort study. *Gynecologic Oncology*, *143*(1), 87–92. https://doi.org/10.1016/j.ygyno.2016.07.095

Moustafa, S., Burn, M., Mamillapalli, R., Nematian, S., Flores, V., & Taylor, H. S. (2020). Accurate diagnosis of endometriosis using serum microRNAs. *American Journal of Obstetrics and Gynecology*, *223*(4), 557.e1–557.e11. https://doi.org/10.1016/j.ajog.2020.02.050

O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology*, *9*(402). https://doi.org/10.3389/fendo.2018.00402

Rout, A. R. (2020). Advantages and Disadvantages of Logistic Regression. GeeksforGeeks. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Saberianpour, S., & Abkhooie, L. (2021). MiR-1307: A comprehensive review of its role in various Cancer. Gene Reports, 101392. https://doi.org/10.1016/j.genrep.2021.101392

Spyros, T., Giorgos, S., Marios, M., Dimitra, K., Ioannis,K., Anna Karavangeli, Filippos S Kardaras & Artemis G Hatzigeorgiou. DIANA-miRPath v4.0: expanding target-based miRNA functional analysis in cell-type and tissue contexts (Nucleic Acids Research, DOI: 10.1093/nar/gkad431)

Sivajohan, B., Elgendi, M., Menon, C., Allaire, C., Yong, P., & Bedaiwy, M. A. (2022). Clinical use of artificial intelligence in endometriosis: a scoping review. *Npj Digital Medicine*, *5*(1). https://doi.org/10.1038/s41746-022-00638-1

Smith, K. (2023, May 3). *Women's health research lacks funding – these charts show how*. Www.nature.com. https://www.nature.com/immersive/d41586-023-01475-2/index.html

Srinivasulu, S., Tsai, M., Sanjay Kumar Shukla, & Ho, S.-Y. (2023). Artificial intelligence-driven pan-cancer analysis reveals miRNA signatures for cancer stage prediction. *Human Genetics and Genomics Advances*, *4*(3), 100190–100190. https://doi.org/10.1016/j.xhgg.2023.100190

Ye, J., Peng, H., Huang, X., & Qi, X. (2022). The association between endometriosis and risk of endometrial cancer and breast cancer: a meta-analysis. *BMC Women's Health*, *22*(1). https://doi.org/10.1186/s12905-022-02028-x

Yoneda, A., Lendorf, M. E., Couchman, J. R., & Multhaupt, H. A. B. (2011). Breast and Ovarian Cancers. *Journal of Histochemistry & Cytochemistry*, *60*(1), 9–21. https://doi.org/10.1369/0022155411428469

Zhang, A., & Hu, H. (2022). A Novel Blood-Based microRNA Diagnostic Model with High Accuracy for Multi-Cancer Early Detection. *Cancers*, *14*(6), 1450. https://doi.org/10.3390/cancers14061450

Zhang, Y. L., Wang, R. C., Cheng, K., Ring, B. Z., & Su, L. (2017). Roles of Rap1 signaling in tumor cell migration and invasion. *Cancer biology & medicine*, *14*(1), 90–99. https://doi.org/10.20892/j.issn.2095-3941.2016.0086

Zhao, Y.-N., Chen, G.-S., & Hong, S.-J. (2014). Circulating MicroRNAs in gynecological malignancies: from detection to prediction. *Experimental Hematology & Oncology*, *3*(1), 14. https://doi.org/10.1186/2162-3619-3-14