

# Using MicroRNA and Deep Learning to Noninvasively Diagnose Gynecological Diseases.

Palak Yadav

Massachusetts Academy of Math and Science

STEM Project

Instructor: Kevin Crowthers, Ph.D.

**Table of Contents:**

<b>GLP Record-Keeping Contract</b>	<b>1</b>
<b>Logbook Etiquette Date</b>	<b>2</b>
<b>Brainstorming</b>	<b>3</b>
Pie Diagrams:	3
Mind Maps	3
Scamper	3
Fishbone Diagrams	3
5 Whys	3
<b>Project Introduction:</b>	<b>3</b>
<b>Professional Communication:</b>	<b>4</b>
<b>Materials and Methods:</b>	<b>4</b>
<b>Background:</b>	<b>4</b>
<b>Daily Entries:</b>	<b>4</b>

# GLP Record Keeping Contract

I, Palak Yadav, commit to record keeping in accordance with Good Laboratory Practices.

- My experiments and records will be reproducible, traceable, and reliable.
- I will NOT write my notes on scraps of paper, post-it notes, or other disposable items. My notes will go directly into my laboratory notebook.
- My data will be recorded in real-time. If I cannot record data in real-time, I will record raw data as soon as physically possible.
- I will record both qualitative and quantitative observations in my laboratory notebook and laboratory reports.
- My laboratory notebook will include information on the materials and instruments utilized during experimentation.
- I will initial and date over the edge of any material that is taped into my laboratory notebook.
- I will provide a real-time record of any analysis I perform.
- I will use a blue or black pen to make entries in my laboratory notebook. I will NOT use a pencil.
- I will define ALL abbreviations.
- If I make a mistake in my laboratory notebook, laboratory worksheets, or other written material, I will not obliterate or obscure the mistake. Instead, I will cross out the mistake using a single line. Any empty spaces in tables or partially used notebook pages will be crossed out using a single diagonal line.
- If I record information online (ex. In Google Drive), I will login so that my contributions are traceable.
- I will initial and date each page in my notebook and the front of each laboratory report.

**Palak Yadav**

---

Printed name



Signature

**10/26/2023**

---

Date

A more detailed description of GLP is located here:

<https://docs.google.com/document/d/1zeYoNSniKTc7MIBgTG1SEnhJiCK3UimCvTcKPQcyHGw/edit?usp=sharing>



# Logbook Etiquette

# Date

For research and engineering purposes, a logbook is considered a legal document and will help in providing documentation for the origin of ideas.

1- When adding something written in Pen- Blue or Black ~~not a Pencil~~ (and DO NOT USE WHITEOUT- mistakes can be corrected by adding the information above the crossed out material and adding your initials and date

2- Don't worry about neatness- it is a living document but **should be legible but understandable**

3- Page Numbers should be consecutive and located on the top corner of the page- outer edge

4- Do not remove pages

5- Put a line through empty space

6- Neat handwriting

7- **Make an entry every time you work on your project**

8- Make sure your entries are verified by a mentor/ teacher signature and your signature

9- Organize your Notebook: Format

A: Table of Contents

B: Brainstorming and Topic Ideas

C: Project Introduction: Topic, Phrase 1 (Testable Question/Engineering Need/Mathematical Conjecture),  
Phrase 2 + Timeline

D: Communications (i.e. to corresponding authors, mentors, and expert consultation, etc)

E: Draft of Materials and Methods (this can be performed for daily entries if variations occur over the course of the project).

F: Background- ie. competitor/market analysis, criteria/constraints

G: Daily Entries (every time you complete work on the project)

1: Title and Date

2: Short Introduction (putting the experiment/observations into context/objectives)

3: Methods/Materials (if not included in the beginning of the notebook)

Materials become important when someone needs to repeat your experiments

4: Observations/Experimental Data (both RAW and ANALYZED)-

A: graphs/figures

B: data tables

C: pictures

D: sketches or proof of concepts and prototypes (with labels)

E: Decision matrices

E: Ethical responsibility

5: Calculations and Data Analysis (STATISTICS)

6: Final Concluding Remarks

## Things to keep in mind:

-You don't want to have too much blank space

-If you are adding a pre-printed graph or sketch, paste in and sign + date.



Time Log

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	T
			<b><u>SEPTEMBER</u></b>				
27 - 1 hr Read a Journal	28	29 - 1 hr Mind map, create next steps plan	30 - 1 hr Read articles from local labs	31 - 1 hr Identify list to contact	1 - 1 hr Article reading	2	5
3	4	5	6 - 2 hr Brainstorming with peers & practicing elevator pitch	7	8	9	2
10 - 2 hr Read articles	11 - 1 hr Research about current medicine	12 - 1 hr Read articles	13 - 2hr Drafting emails, updating forms, reading articles	14	15	16	6
17 - 1 hr Update notes and read articles	18	19	20	21	22	23 - 3 hr Finish 1 article Contact Isabella Pailt	4
24	25	26	27 - ½ hr Professor email list	28	29	30 - 2 hr Professor email list	2.5
			<b><u>OCTOBER</u></b>				
1 - 2 hr Send emails, read more	2	3 - ½ hr Send emails	4	5	6	7 - 2 hr Read articles	4.5
8 - 1 hr Send emails	9 - 1 hr Send emails	10 - ½ hr Update notes	11	12 - 3 hr Zoom call, send emails	13 - ½ hr Update notes	14 - 1 hr Update notes	7
15 1hr Update documents	16	17 - 2 hr Read article, start presentation	18 - 2 hr Speak with Dr. Jytoi	19	20 - 1hr Send emails	21	6
22 - ½	23	24	25	26	27	28	7

Send emails	Prep + meet with Dr. Mattson 2hr	Read about Phosphoinositide and endometriosis connection  1 hr	Meeting with Dr. Zhuge  Understand Whittington work  1 hr		Read articles 1 hr	Talk to Dr. Simon 1.5 hr	
29 Read articles -develop a list of what is lacking based on the endo review article  Understanding the molecular basis of EC & endo will shed light on their correlation → improve patient outcomes  1 hr	30 Read other projects in Women's health (30 mins)  Linear Regression - AI Club (1.5 hr)	31  1.5 hr – outreach  1 hr mins - alternative project ideas + generate a list of articles and ideas to explore	1	2 - 3 hr -Update emails -Read articles -Meeting plan w/ Whittington	3 - 1 hr Dr. Whittington Meeting	4	9
			<b><u>NOV</u></b>				
5 - 3 hr microRNA article AI & gyno research	6- 2 hr Article research	7 - 2 hr Article research			10 - ½ Professor Wu meeting	11 - 2 hr Research articles Understand lipid structures	9.5
12 - 1.5 hr  Continue reading from yesterday  Goal: - ML basics	13 - 1.5 hr  Alexandra Harrison mtg + placenta research (mh & stem cell ideas)	14 - 2 hr  Meeting prep Nanoparticles application	15 - 1.5 hr Meeting with Prof. Obayemi (½) Research	16	17 - ½ Meeting with Dr. Terry	18 - 3hr Learn about ML & identify endo data sets	11

- GWAS papers - Epigenetics lab?	(lunch + 6-7)						
19 (1.5 hr) Meeting with Rianna	20 - ½ AI Club - 1 hr Meeting with Isabella - Look at melanoma	21	22 - 1hr Meeting with Prof. Ambaddy	23 - 3hr Breast cancer research	24 - 3hr Machine Learning Model Basics	25 - 4 hr REad articles Identify Databases	
26 - 4 hr Grant Proposal Draft Pitch	27 - 2 ht Research + notes	28	29	30 - 2hr Meeting w/ Dr. Moore Research	1	2	8
			<b><u>DEC</u></b>				
3 - 4 hr Meeting w/ Joseph (1) STEM meeting prep (1) Understanding stats + R (2)	4 - 2 hr Update notes Meeting prep	5 - 1 hr poster development + Articles	6 Poster development	7 Endo data	8	9 - 1.5 hr Improve Bioconductor analysis on endo data	8
10 - 1 hr Patent search	11 - 1 hr Article notes	12 - 2 hr Finished 1 patent and 1 article	13 - 1 Finish 2 articles	14 - 1 Finish all articles	15 - 1 final review	16	8
17	18	19 - 1 hr Find datasets	20	21 - 2 hr Create outline Find R packages Meeting with Professor	22	23	3
24	25 - 1 hr	26	27	28	29 - 2 hr		
			<b><u>Jan</u></b>				
	1 - 1 hr	2 - 1hr Meeting with Professor	3	4 - 2 hr	5	6 - 1 hr	4.5

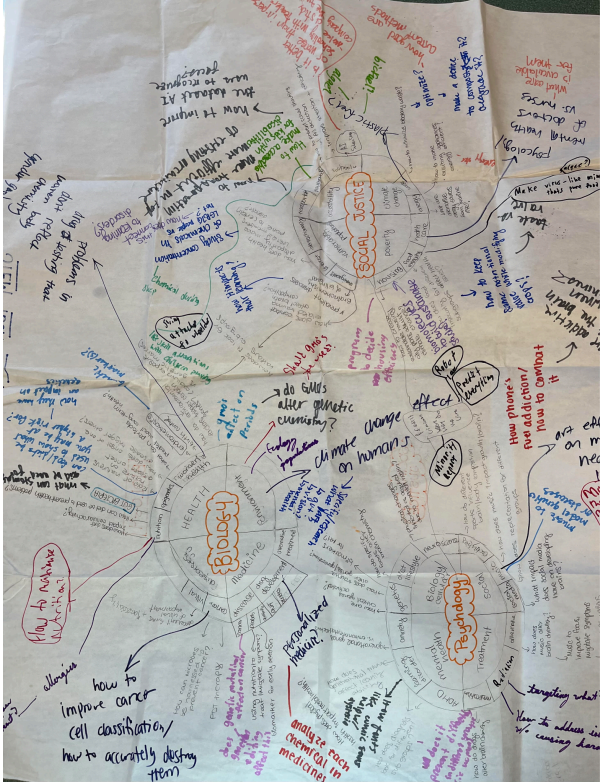
7	8 - 2hr Fix BC Meet with professor	9	10	11 - 1.5 Neural Network for Breast Cancer data	12	13	3.5
14-2.5 hr	15	16	17 - 2 hr Random Forest Frameworks of DNN	18	19	20 0.5 hr Fix DNN	
21 - 2 hr Clean binary classification model	22	23	24 - 2.5 hr Finish DNN model	25 - 1 hr 1- professor	26	27 - 2 hr Feature extraction	7.5
28	29 - 1 hr Pathway modeling	30 - 1 hr DeepLift algorithm on Random Forest	31	1 - 1 hr Deeplift algorithm on DNN binary classification + meeting with a professor	2	3 - 2 hr Compile data and experiment with pathway modeling data	5
4 - 3 hr Finish pathway modeling and pathway analysis							



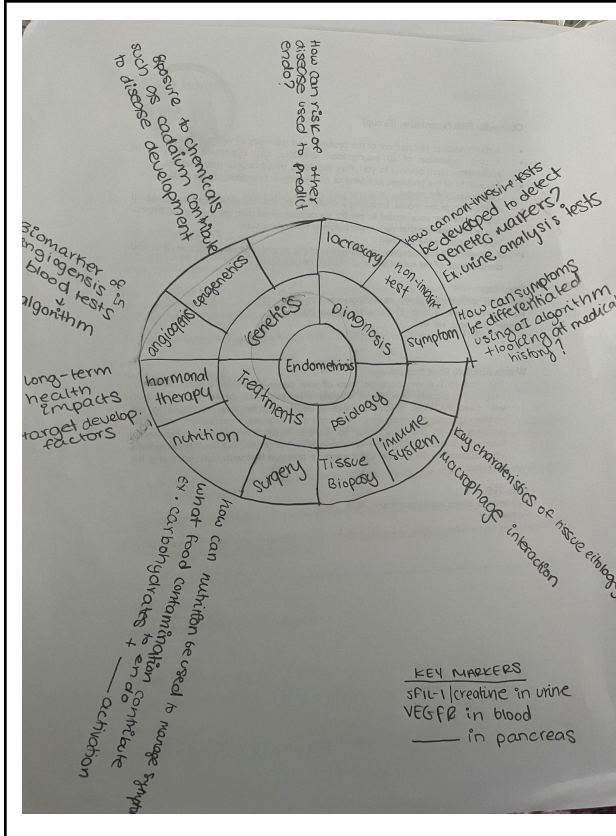
# Brainstorming

## Gantt Chart:

## Pie Diagrams:

Diagram	Explanation
	<p>8/31/23 PY.</p> <p>This is a brain dump of all my ideas. I am interested in three major areas: biology, psychology, and social justice. I wrote down all the topics I am interested in and asked for my peer's inputs as well.</p>

Parak



9/12/23 PY

This mind map pulls together all the health-related topics I am interested in and potential avenues for a project. I am interested in looking at the accessibility and environmental factors that influence a certain population's health.

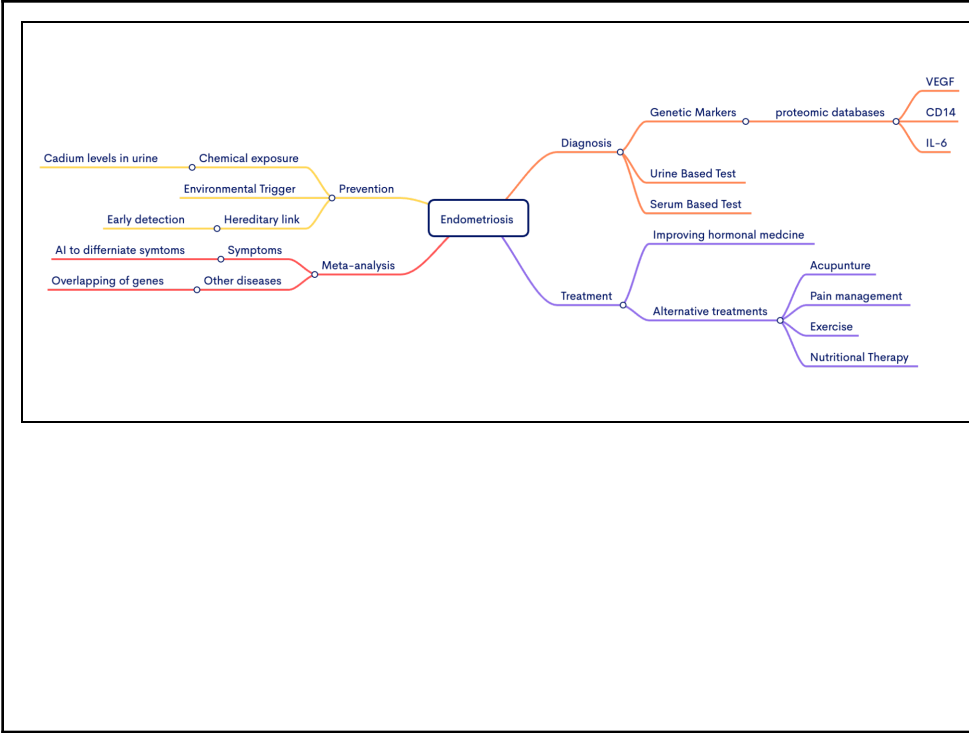
Mind Maps

Diagram	Explanation
	<p>9/2/23 PY</p> <p>This is a mind map for my interest in healthcare. I condensed my piedrams into tangible topics and research questions. I find the psychology questions interesting as well as the intersection with the environment.</p>

Parak

9/23/23 PY

This mind map dives deeper into endometriosis and potential projects I could pursue in this field. I am especially interested in building a diagnostic tool as surgical methods are the only available and effective methods, however, it can take up to 5-7 years to diagnose it. I am hoping to identify potential genetic markers that can be targeted in serum-based tests. Another avenue is urine-based tests that detect the metabolites of the major biomarkers.



## Fishbone Diagrams

Diagram	Explanation
	<p>8/23/2023</p> <p>Endometriosis is a reproductive condition where the lining of the uterus grows on the outside instead of the inside. Currently, there are no effective and noninvasive diagnostic tools for this condition. A major cause of this is the lack of funds and research. Therefore, diagnosis There are six major angles I looked at to this problem. The presence of endometriosis has been linked to various other</p>

Palak

	<p>9/1/2023</p> <p>Learning disabilities have been on a rise in the past few decades. This fishbone diagram looks at different aspects of misdiagnosis in learning disabilities, from protocols followed in schools to the understanding of the brain morphology. Developing a thorough understanding of the psychological aspect of learning disabilities will be key to developing efficient diagnosis tools and protocols.</p>
	<p>9/13/2023</p> <p>Current building structures are not environmentally friendly and are susceptible to damage by natural disasters. By considering different systems found in a building, nature inspired designs can be implemented to promote safety and reduce carbon footprint.</p>

## Pre Project Planning Document

[W Palak\\_Pre-Project Planning.docx](#)

## Project Abstract

Women’s health is a growing public health crisis as 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 (miRNAs) are promising non-invasive biomarkers detectable in bodily fluids and known to play key roles in disease development. This study aimed to develop a deep neural network that utilizes serum miRNA expression profiles to predict these gynecological conditions. A predictive model was trained on public miRNA datasets to generate unique miRNA profiles for each condition. Key miRNA signatures associated with each disease were identified

Palak

and used to model pathways potentially involved in disease development and progression of breast cancer, ovarian cancer, and endometriosis. This approach offers a non-invasive diagnostic tool with the potential for early disease detection, provides valuable insights into disease pathology, and paves the way for future therapeutic development.

## Systems Map

### Project Introduction:

#### **Problem Statement**

Diagnosing gynecological conditions can take a prolonged time, due to the lack of adequate sources, social stigma, and more. miRNAs hold promise as noninvasive and accessible screening candidates in gynecology. My research objective is to evaluate the role of miRNA as a strong diagnostic candidate and develop a robust classification model.

#### **Research Question:**

- How can miRNA classification be used as a noninvasive diagnostic candidate?
- What implication do miRNAs have about the pathology and comorbidities of gynecological conditions?

#### **Research Objective:**

The goal of this project is to design a machine-learning model that analyzes upregulated and downregulated microRNA expression from publicly available serum sample datasets to predict the likelihood of the following diseases: endometriosis, ovarian cancer, and breast cancer.

#### **Hypothesis:**

I hypothesize that miRNAs will prove to be an effective marker for gynecological conditions, and a combination of Random Forest and Deep Neural Network will prove to be the most efficient at predicting the likelihood of specific gynecological disease. Additionally, a shared pathway between the three diseases will be identified which will shed insight on their shared pathology and interconnectedness.

#### **Brief Overview**

Endometriosis is a gynecological condition that impacts over 190 million women of reproductive age around the world. It is marked by abnormal ectopic and eutopic growth of endometrial tissue outside of the uterus and often results in implants to other nearby organs. This results in chronic pelvic pain, painful periods, infertility, and other comorbidities (Horne, A. W., et al., 2022). Despite its high prevalence, there is little understanding of the pathology and etiology of this condition.



Confirmative diagnosis can take up to 8-10 years, for multiple reasons, including but not limited to, the dismissal of the symptoms due to regular menstrual pain, lack of access to adequate resources, social stigma, and more (Ahn, S. H., et al, 2017). This can impose significant challenges on the patient and healthcare system. The goal of my project is to evaluate the geographic variance of endometriosis levels and the correlation to factors, such as UV radiation, water sanitation, diet, exposure to chemicals, and more. These findings will help better understand the correlation between environmental triggers and the pathology and etiology of endometriosis.

### Initial Project Plan

- Identity datasets + Data Samples\_Gynecological Conditions
- Develop a unique miRNA panel for each condition
- Determine the ideal cut-off point for the top differentially expressed miRNAs
- Binary Classification between healthy vs unhealthy miRNA samples
  - a. Random Forest
  - b. Logistic Regression
  - c. Feature importance of each miRNA – which ones are most predictive?
- Multi-Disease Classification
  - a. Consider Random Forest vs DNN
- Analysis
  - a. Measure:
    - i. Sensitivity
    - ii. Specificity
    - iii. AUC
    - iv. Accuracy
- Run pathway modeling database to determine biological significance
- Verify from previous studies

### Professional Communication:

Template
<p>Hello Dr.____,</p> <p>I hope you are doing well. My name is Palak Yadav and I am a junior at the Mass Academy of Math and Science at Worcester Polytechnic Institute. I am conducting a five-month research project on endometriosis and I recently came across your work.</p>



I am inspired by your research on analyzing diverse data sets to identify genetic markers for endometriosis. I would love to learn more about your work and potential ways to get involved in endometrial research.

If this sounds possible, I am happy to schedule a meeting to discuss further or share my questions via email if that is more convenient for your schedule.

Thank you for your time and consideration,  
Palak Yadav (she/her)

UMass Chan/other

Name	Communication	Reponse
Marc Laufer (Boston Children)	Sent 9/21 + 2 times	No response
Tiffany A. Moore Simas (UMass)	Sent 10/1	Received future contacts
Kristen Anne Matteson MD	10/17 10/31	No response
Julia V Johnson MD	10/31	
William F Flynn Jr. MD	10/28	
David Robertson MD	10/27	
Susan Zweizig MD	10/27	
Christopher Marshall MD	10/30	
Christopher N Otis MD	10/25	
S. Jean Emans MD	10/20	
Catherine M. Gordon, MD MS	10/10	
Heidi C. Fantasia PhD Fantasi	10/10	

PD  
Palak

Mahdi Garelnabi BS		
Elizabeth R Bertone-Johnson ScD SM		
Brian W Whitcomb Ph.D., B.A.		
Carly Detterman CNM, MSN		

WPI

Reeta Prusty Rao	10/20	
Catherine Whittington	11/1	Did not have lab opportunity but offered insights
Tanja Dominko	10/20	
Anita Mattson	Ovarian cancer background 10/4	
Joseph B. Duffy	Doesn't have capacity	

## STEM Update Meetings

1/22/24

- Binary classification is a method of supervised learning that categorizes datasets into two classes 1 and 0, and in this case, there will be two outcomes: disease vs healthy.
- Based on previous studies of what binary classification models proved to be the most effective, I found logistic regression and random forest to have the highest accuracy.





- I developed a logistic regression model and random forest for each of the three conditions, experimented with three test and train splits to determine which one had the highest accuracy.
- Logistic regression uses the sigmoid function to map the function of best fit that will allow for the binary classification of a sample set. It outputs a values of 1 or 0 and we get to set the threshold, which is generally 0.5
- Random state can be any number, and its goal is to regenerate the data that many times to test it
- It's important to scale the values between one range so all miRNA are considered equally

### Logistic Regression Classifier in Python - Basic Introduction

In logistic regression... Basically, you are performing linear regression but applying a sigmoid function for the outcome.

### Sigmoid / Logistic Function

$$p=1/1+e^{-y}$$

### Properties of Logistic Regression

The dependent variable follows a Bernoulli Distribution

Estimation is maximum likelihood estimation (MLE)

### Advantages

Straight forward, easy to implement, doesn't require high compute power, easy to interpret, used widely.

Doesn't require feature scaling and provides a probability score for observations.

### Disadvantages

Not able to handle a large number of category features/variables.

Vulnerable to overfitting.

## Materials and Methods:

Materials:

Applications	Description
Google Collaboratory	A product from Google Research that allows for accessible and efficient
Python 3	Python code with support for important libraries and functions for machine learning models.

<b>TensorFlow/Keras</b>	A Python-based framework for building, training, and testing machine learning models
<b>Matplotlib</b>	Python library for visualization and graphics.
<b>Sci-Kit Learn</b>	Python-based machine learning library that consists of commands and functions to create a wide range of models.
<b>R/R-Studio</b>	An integrated platform to use R and R-based packages and libraries for statistical analysis.
<b>GEO2R</b>	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.
<b>DIANA-miRPath v4.0</b>	microRNA pathway modeling software that extracts genes and signaling information from databases such as KEGG and performs analytical tests. <a href="http://62.217.122.229/app/miRPathv4">http://62.217.122.229/app/miRPathv4</a>

### Datasets:

<b>GEO Access Link</b>	<b>Description</b>
GSE106817	333 ovarian cancers, 66 benign tumors, 29 of ben ovarian, 143 breast cancer, and 275 of 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.

The following steps will be taken to meet the specified aims of the project:

1. Accurate datasets and miRNA profiles from at least 50 samples of each condition must be obtained. The National Institute of Health (NIH) Gene Expression Omnibus and the Cancer Atlas TCGA contain publicly available miRNA profiles for several diseases. The GEOR2 platform can be used to apply statistical tests to determine the differential expression of miRNAs.
2. All the miRNA sample profiles will be compiled into one large dataset to allow for efficient computation. Firstly, all data must be normalized using the mix-max normalization method and clean irrelevant features for efficient and accurate evaluation. Previous studies have developed a thorough feature selection procedure to prevent overfitting and preserve necessary values to reduce complexity and time (Hamidi et al., 2023).
3. Develop a model that performs binary classification on each condition dataset. Utilize Logistic Regression, Random Forest, and K-Nearest Neighbor Algorithm to differentiate the miRNA markers of samples with the conditions and healthy samples.

4. miRNA markers identified for each gynecological condition from the last step will be utilized to design a Deep-Learning Neural Network that can classify three different gynecological conditions. Python packages such as TensorFlow, NumPy, Pandas, Matplotlib, and Scikit Learn are useful in developing neural networks.
5. Identify which miRNA markers are most influential in the onset of each gynecological condition using feature extraction algorithms (Srinivasulu et al., 2023).
6. Evaluate the biological significance of the differential expressed microRNAs by modeling pathways in the KEGG Pathway Database to draw biological significance, provide insights for potential therapeutic and diagnostic targets, and advance the current understanding of the pathology of gynecological diseases (Alharbi & Vakanski, 2023).
7. The model will be validated through comparison to previous studies to verify which miRNA and pathways were found to be common and which were newly identified.

## Background:

Women's health has historically been dismissed and stigmatized in society, and even today, painful periods and chronic pelvic pain are deemed as "ladies' problems" (Cook & Dickens, 2014). An analysis of the National Institute of Health 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 ovarian cancer cases being diagnosed at an advanced stage (Mogensen et al., 2016). The lack of funding coupled with the challenges of the vagueness of symptoms, lack of awareness and accessibility to resources, and little



understanding of the pathology, make 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.

### **Diagnosing Gynecological Conditions**

Although many past studies have attempted to identify noninvasive diagnostic tools, there is still no screening or non-invasive tool available for several gynecological conditions such as ovarian cancer and endometriosis (Ginsburg et al., 2017). Endometriosis results from the abnormal growth of the uterine lining and may develop into ovarian cancer and breast cancer if not treated effectively. 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). Although endometriosis is a benign condition, it 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). Due to the prevalence and detrimental impacts of these conditions, understanding their pathology and etiology is key to identifying potential therapeutics and strategies to diagnose them effectively.

### **microRNAs and Gene Expression**

Harnessing gene expression data as biomarkers for many diseases is an emerging area of research. RNA-seq data examines large datasets of RNA extracted from blood serum, plasma, and tissue biopsies to identify mRNA (protein-coding) or microRNA (non-coding) profiles. MicroRNAs are 22 nucleotides long and play a key role in regulating gene expression and vital biological processes. They function as post-transcription regulators in the 3' untranslated regions and cause the degradation of mRNA (Li et al., 2023). There are multiple mechanisms behind the functionality of miRNA, starting from the transcription of the primary miRNA by RNA polymerase II to the migration of the precursor miRNA from the nucleus to the cytoplasm to be transformed by the Dicer to gain more functionality. Ultimately, each miRNA targets specific genes by attaching to the mRNA molecules, leading to the activation or inhibition of the intended protein (O'Brien et al., 2018).

Compared to other types of non-coding RNAs, microRNAs are easy to analyze due to their abundance and accessibility in bodily fluids such as blood, urine, saliva, and plasma, making them a promising candidate as noninvasive biomarkers for a variety of diseases (Zhao et al., 2014). They also 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 the dysregulation of numerous biological pathways and signaling (O'Brien et al., 2018). Many previous works have found that certain diseases have different miRNA expression levels in comparison to control samples, however many gynecologic diseases have shared pathology and therefore overlapping aberrant miRNA expression (Zhao et al., 2014). Therefore, developing unique miRNA profile panels for each disease will allow for the identification of strong biomarkers for diagnostic development.

### **Role of 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). A systematic review of machine learning and miRNAs in cancer classification stated that the following algorithms are most used in machine learning models: Decision Trees, Support Vector Machines, Random Forest Trees, KNN algorithms, and Artificial Neural Networks (Sivajohan et al., 2022). Such algorithms allow for binary classification of having a condition or not having a condition and subtype differentiation. Deep Neural Networks allow for more sophisticated computation and can differentiate between multiple input classes and determine the weight of each input in the final prediction (Alharbi & Vakanski, 2023).

## Daily Entries:

This section should include specific components to the Engineering Design Process (Build, Test/Evaluate/Revise, Reflection) or Research Process

Experiment 1: Title, Date, eSignature

Introduction:

Methods/Materials

Observations and Experimental Data:

Calculations and Data Analysis:

Concluding Remarks

2021\_STEMProject\_Ex1Test1\_Data\_Crowthers\_v1-01 (link to data file)

**Experiment 0: 12/25/23**

SIGNED BY: PALAK YADAV

I download necessary packages and libraries in R Studio to build a binary classification

The steps are outlined below:

1. Download R Studio and R on Mac 11: <https://posit.co/download/rstudio-desktop/>
2. Download packages
  - a. `library(ggplot2)`
  - b. `library(cowplot)`
  - c. `library(randomForest)`
  - d. `library(ROCR)` # Includes function to calculate the AUROC metric
  - e. `library(PRROC)` # Includes function to calculate the AUPR metric
  - f. `library(tidyverse)`
  - g. `library(tidymodels)`
  - h. `library(lmtest)`

Binary Classification will allow the prediction of a gynecological condition based on the expression of miRNAs. Given a set of variables, the probability of the outcomes is determined

- Test the ISLR library using the Default dataset
  - Format the data in readable format
    - Tissue Type (Ovarian Cancer vs Healthy) → default column
    - miRNA levels
- Apply a similar procedure to ovarian cancer dataset



**Experiment 1: 12/29/23**

SIGNED BY: PALAK YADAV

**Objective:** Identity differentially expressed miRNA

- Run GEOR2 statistical analysis on all the datasets to identify differential expression for each disease.
- Compile data from GEO to R
- Experiment with multiple ways of retrieving data
- Generated a new list of top up and down regulated miRNA using a different set for ovarian cancer <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE235525>

**Spreadsheet to the tables for each data****Insert volcano plots**

Next steps: start building machine learning models using feature selection methods

**Experiment 1: 1/3/24**

SIGNED BY: PALAK YADAV

**Objective:** Build logistic regression models for binary classification**Methods:**

- Normalize dataset to allow for accurate computation
- Clean out missing data points by interesting mean values of that column
- Write code for logistic regression model in R using templates by resources such as WW3 Schools, University website, and those provided by Dr. Moore
- Featured engineering by ranking the top 5 miRNAs identified by the last statistical tests

[Logit Regression | R Data Analysis Examples](#)[Python Machine Learning - Logistic Regression](#)**Observation:**

- Errors in accurately splitting the data for train and test sets





- split list [49 x 21 x 70 x 823] (S3: List of length 4
- data list [70 x 823] (S3: data.frame A data.frame with 70 rows and 823 columns
- in\_id integer [49] 50 46 13 10 59 29 ...
- out\_id logical [1] NA
- id list [1 x 1] (S3: tbl\_df, tbl, dat A tibble with 1 row and 1 column

Data	
model	List of 6
mydata	70 obs. of 823 variables
mylogit	List of 30
newdata1	4 obs. of 7 variables
predictions	21 obs. of 1 variable
split	List of 4
test	21 obs. of 823 variables
tidy_data	823 obs. of 3 variables
train	49 obs. of 823 variables

Newdata after ranking of the top 5 miRNAs:

	hsa.miR.5 79.5p	hsa.miR.124 7.5p	hsa.miR. 4455	hsa.miR.145. 5p	hsa.miR.142 .5p	ra n k	rank P
<b>1</b>	553.7571	126.3571	103.3143	120.5143	167.6714	1	0.88813 69
<b>2</b>	553.7571	126.3571	103.3143	120.5143	167.6714	2	0.88813 69
<b>3</b>	553.7571	126.3571	103.3143	120.5143	167.6714	3	0.88813 69
<b>4</b>	553.7571	126.3571	103.3143	120.5143	167.6714	4	0.88813 69

Coefficients value of the feature miRNA:

```
Call:
glm(formula = Condition ~ hsa.miR.579.5p + hsa.miR.1247.5p +
     hsa.miR.4455 + hsa.miR.145.5p + hsa.miR.142.5p, family = "binomial",
     data = mydata)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.590024	2.633200	0.224	0.82270
hsa.miR.579.5p	0.001895	0.001749	1.084	0.27849
hsa.miR.1247.5p	0.060096	0.027313	2.200	0.02779 *
hsa.miR.4455	0.029666	0.026549	1.117	0.26381
hsa.miR.145.5p	-0.012134	0.015902	-0.763	0.44545
hsa.miR.142.5p	-0.052269	0.019881	-2.629	0.00856 **

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 96.983 on 69 degrees of freedom  
Residual deviance: 13.915 on 64 degrees of freedom  
AIC: 25.915

Number of Fisher Scoring iterations: 9

# A tibble: 823 × 3

term	estimate	penalty
<chr>	<dbl>	<dbl>
1 (Intercept)	0.840	0
2 Patient	-0.00289	0
3 hsa.let.7b.5p	-0.00106	0
4 hsa.let.7c.5p	-0.0000277	0
5 hsa.let.7d.5p	0.0000234	0
6 hsa.let.7e.5p	-0.00143	0
7 hsa.let.7f.5p	0.00260	0
8 hsa.let.7g.5p	-0.00194	0
9 hsa.let.7i.5p	-0.000107	0
10 hsa.miR.1.3p	-0.00000954	0

# i 813 more rows

```
> # Confidence intervals using profiled log-likelihood
> confint(mylogit)
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	-4.667438526	6.454940628
hsa.miR.579.5p	-0.001447820	0.005749137
hsa.miR.1247.5p	0.021671590	0.138012123
hsa.miR.4455	0.004282107	0.106539469
hsa.miR.145.5p	-0.053405331	0.010857088
hsa.miR.142.5p	-0.105722632	-0.021416213

There were 50 or more warnings (use warnings() to see the first 50)

>

\*\*Need to fix error in the sizing

of the data

Warnings:

Warning messages:

1: glm.fit: fitted probabilities numerically 0 or 1 occurred

↑  
Tarak

According to the sources, this warning occurs when the outputs are too close to 0 or 1, and is often called quasaki or separation problem.

term	estimate	penalty
1 (Intercept)	8.396905e-01	0
2 Patient	-2.892084e-03	0
3 hsa.let.7b.5p	-1.058050e-03	0
4 hsa.let.7c.5p	-2.773855e-05	0
5 hsa.let.7d.5p	2.343171e-05	0
6 hsa.let.7e.5p	-1.433872e-03	0
7 hsa.let.7f.5p	2.597280e-03	0
8 hsa.let.7g.5p	-1.935912e-03	0
9 hsa.let.7i.5p	-1.072943e-04	0
10 hsa.mir.1.3p	-9.543766e-06	0
11 hsa.mir.1.5p	-2.401419e-03	0
12 hsa.mir.100.5p	6.801341e-04	0
13 hsa.mir.101.3p	-1.823210e-03	0
14 hsa.mir.103a.3p	1.583880e-04	0
15 hsa.mir.105.5p	-9.733493e-04	0
16 hsa.mir.106a.5p.hsa.mir.17.5p	1.843669e-03	0
17 hsa.mir.106b.5p	-7.172465e-04	0
18 hsa.mir.107	-1.166478e-03	0
19 hsa.mir.10a.5p	9.731177e-05	0
20 hsa.mir.10b.5p	8.259201e-04	0
21 hsa.mir.1178.3p	-4.731734e-04	0
22 hsa.mir.1180.3p	-1.811798e-03	0
23 hsa.mir.1183	3.507202e-04	0
24 hsa.mir.1185.1.3p	-4.939797e-05	0
25 hsa.mir.1185.2.3p	-2.242521e-03	0
26 hsa.mir.1185.5p	-1.662792e-03	0
27 hsa.mir.1193	-2.876002e-04	0
28 hsa.mir.1197	5.627332e-04	0
29 hsa.mir.1200	-7.009509e-04	0
30 hsa.mir.1202	-1.405486e-03	0
31 hsa.mir.1203	8.145850e-04	0
32 hsa.mir.1204	2.161497e-04	0

### Next steps:

fix the errors found in the fitting of the data. Researching solutions on StackOverflow suggested the following methods to fix the errors:

Here are a few potential steps to address or investigate the issue:

- Check for Separation:
  - Inspect your data for potential separation. Separation occurs when there is a perfect (or near-perfect) relationship between the predictor variables and the response variable, leading to infinite or very large coefficient estimates.
  - Review the distribution of the outcome variable and the relationship between predictor variables and the outcome.
- Consider Regularization:
  - If you have a large number of predictor variables, regularization techniques (e.g., L1 or L2 regularization) can be applied to prevent extreme coefficient estimates and address issues with separation. In R, you can consider using the glmnet package for regularized logistic regression.
- Address Collinearity:

- Check for multicollinearity among your predictor variables. High collinearity can lead to unstable coefficient estimates and numerical instability.
- If collinearity is present, consider addressing it through variable selection or regularization.
- Use Firth's Method:
  - Firth's method is a modification of logistic regression that can be used to handle separation. In R, you can use the `logistf` package, which implements Firth's bias reduction method.

**Experiment 2: 1/7/24**

SIGNED BY: PALAK YADAV

**Objective:** Start building random forest model frameworks

**Methods:**

Random forest algorithm

- Needs 3 parameters:
  - Nodes size
  - Features sampled
  - Number of trees
- Helps reduce overfitting and bias
- `set.seed()` functions allows for random creation of numbers

**Observation:** faced challenges in normalizing the datasets and differentiating the multiple classes of disease. R-Studio kept crashing

**Next Steps:** Try building models in python (Google Colab) to see if the machine learning parameters function better.

**Experiment 3: 1/11/24**

SIGNED BY: PALAK YADAV

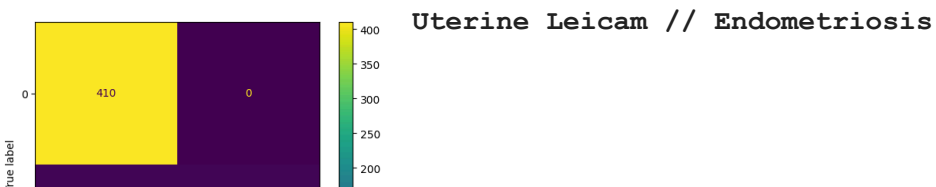
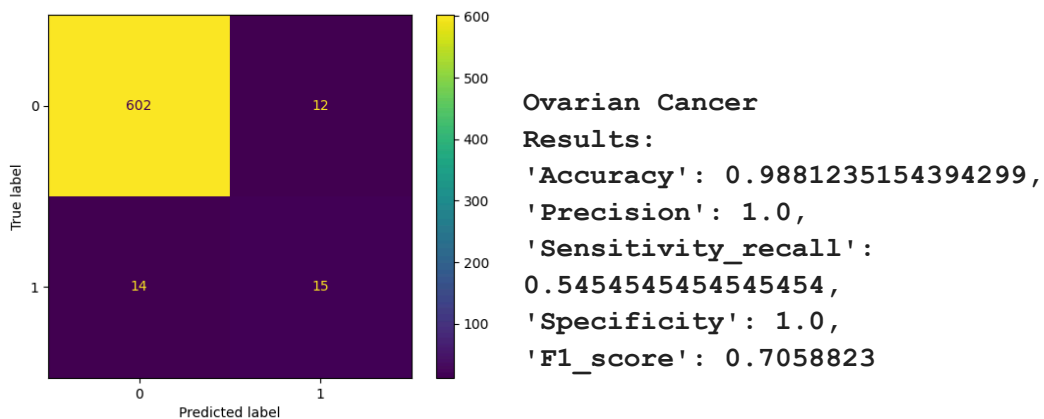
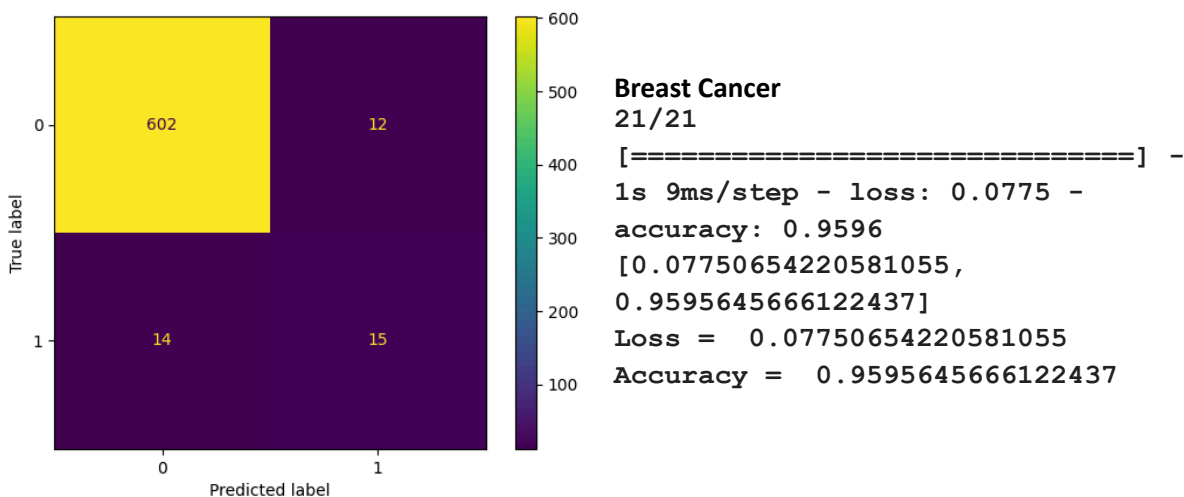


**Objective:** Create Logistic Regression Binary Model in Python

**Methods:**

- Use Sci-kit learning functions
- Determine train and test split
- Set epoch levels
- One-hot encoding to turn disease label and healthy label to numerical values
- Normalize dataset (not including the Condition column)

**Observation:**



Prabhat

Experiment 4: 1/14/24	SIGNED BY: PALAK YADAV
-----------------------	------------------------

<b>Objective:</b> Build multi-class Random Forest model
<b>Methods:</b> <ul style="list-style-type: none"> <li>- Compile all data into one large set</li> <li>- Random train seeds</li> <li>- Differentiate</li> </ul>
<b>Observation</b>
<p>The figure is a Multiclass ROC curve plot. The y-axis is labeled 'True Positive rate' and ranges from 0.0 to 1.0. The x-axis is labeled 'False Positive Rate' and ranges from 0.0 to 1.0. A diagonal dashed blue line represents a random classifier. Five curves are plotted, each representing a different class:         <ul style="list-style-type: none"> <li>Borderline Ovarian Tumor vs Rest (AUC=0.86): Blue dashed line</li> <li>Breast Cancer vs Rest (AUC=0.92): Orange dashed line</li> <li>Healthy vs Rest (AUC=0.97): Green dashed line</li> <li>Ovarian Cancer vs Rest (AUC=0.97): Red dashed line</li> <li>UL vs Rest (AUC=1.00): Purple dashed line</li> </ul>         All curves are significantly above the diagonal line, indicating high performance. The UL vs Rest curve is the highest, reaching a True Positive Rate of 1.0 at a False Positive Rate of approximately 0.1.       </p>
<b>Next steps:</b> High accuracy levels across all datasets - likely due to overfitting. Try applying the same model on datasets collected from different studies <ul style="list-style-type: none"> <li>- When applied to another data set 78.8%</li> </ul>

Experiment 5: 1/17/24	SIGNED BY: PALAK YADAV
-----------------------	------------------------

<b>Objective:</b> Build Deep Neural network - Binary Classification
<b>Methods:</b> <ul style="list-style-type: none"> <li>- Tensorflow/Keras functions</li> </ul>

*Palak*

- Create layers
- Normalization

### Observation

Deep Neural Network  
 Test loss: 0.5793562531471252  
 Test AUC: 0.7293312549591064  
 500 epochs

Next steps: feature extraction techniques

Experiment 6: 1/20/24

SIGNED BY: PALAK YADAV

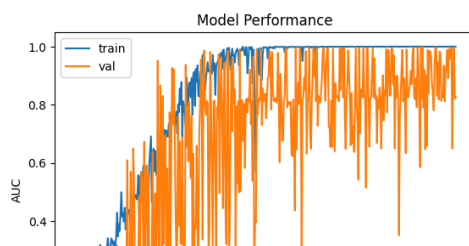
Objective: Build Deep Neural network - Multiclass

### Methods:

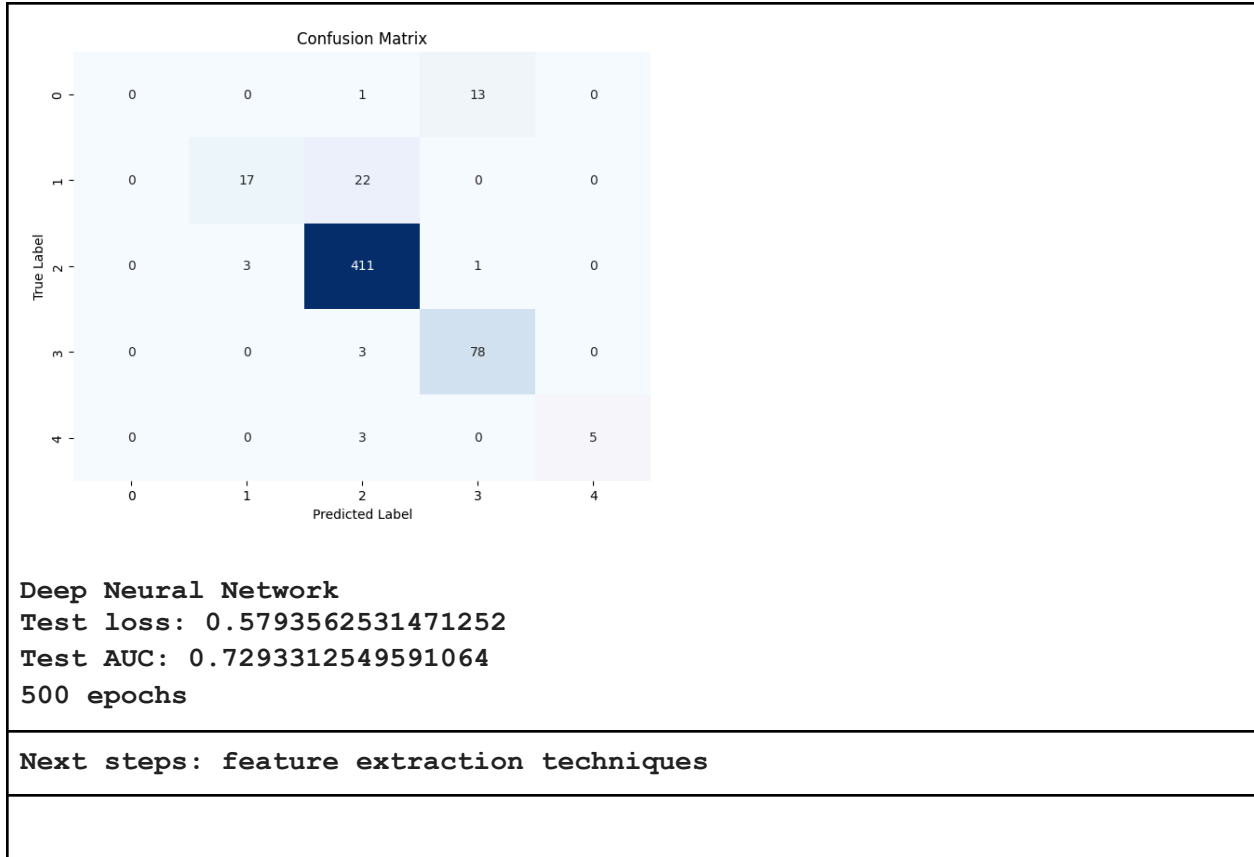
- Tensorflow/Keras functions
- Create layers
- Normalization
- Created neural network on Google Collab for Breast Cancer binary classification using Integrated Dataset
- Had some initial challenges of fitting the model to 2567 miRNA entries
- For preliminary results, applied R statistical tests using GEOR2 to input top 28 differentially expressed miRNA between breast cancer and control
  - Below are the results of this trail
- Next steps: increase the input value, understand what the results indicate

Neural network frameworks: [Binary Classification with TensorFlow Tutorial](#)

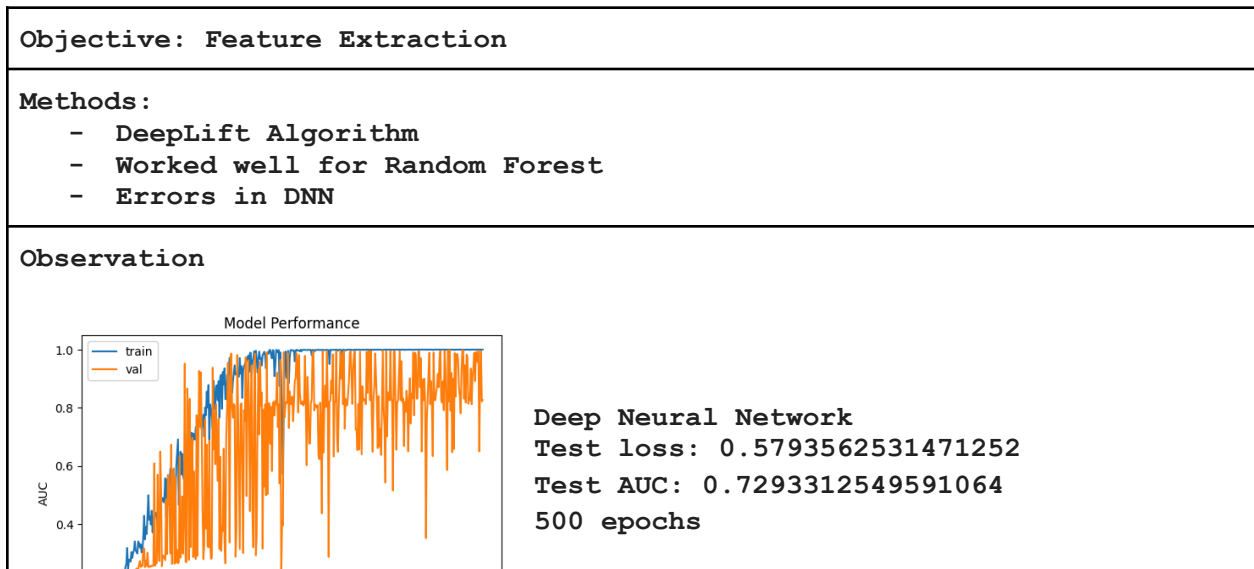
### Observation



Palak



<b>Experiment 7: 1/24/24</b>	SIGNED BY: PALAK YADAV
------------------------------	------------------------



Palak



Next steps: feature extraction techniques

Observation: Random Forest

Feature importance

Random Forest

Top 10 Features:

	Feature	Importance
1773	MIMAT0022259	0.358973
713	MIMAT0005792	0.150006
2241	MIMAT0027500	0.102886
1347	MIMAT0019071	0.041051
701	MIMAT0005582	0.030534
2209	MIMAT0027468	0.017957
1257	MIMAT0018978	0.017559
745	MIMAT0005880	0.016409
2411	MIMAT0027670	0.014648
965	MIMAT0015064	0.010478

Top 10 Features:

	Feature	Importance
1477	MIMAT0019757	0.019348
965	MIMAT0015064	0.015854
1347	MIMAT0019071	0.015106
745	MIMAT0005880	0.015067
1666	MIMAT0019947	0.012000
2543	MIMAT0031095	0.011243
2027	MIMAT0026486	0.011125
1028	MIMAT0016879	0.010507
2018	MIMAT0026477	0.009730
2251	MIMAT0027510	0.009275

Updated Random Forest Code

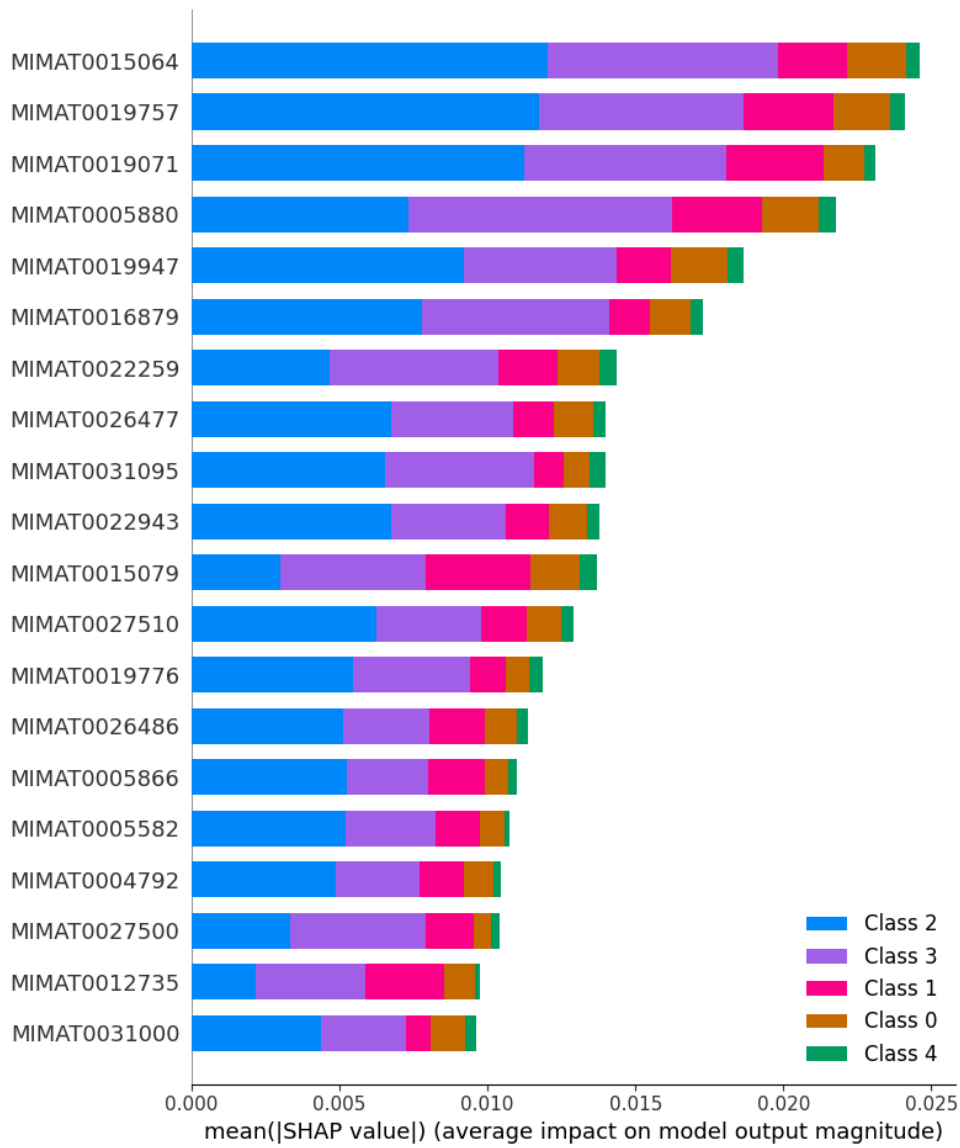
Accuracy: 0.92

Classification Report:

	precision	recall	f1-score	support
Borderline Ovarian Tumor	0.00	0.00	0.00	14
Breast Cancer	0.85	0.44	0.58	39
Healthy	0.93	0.99	0.96	415
Ovarian Cancer	0.85	0.96	0.90	81
UL	1.00	0.62	0.77	8
accuracy			0.92	557

PD  
Tarak

macro avg	0.73	0.60	0.64	557
weighted avg	0.89	0.92	0.90	557



This graph

shows the top ten features in the random forest model and their contribution to the prediction of each disease.

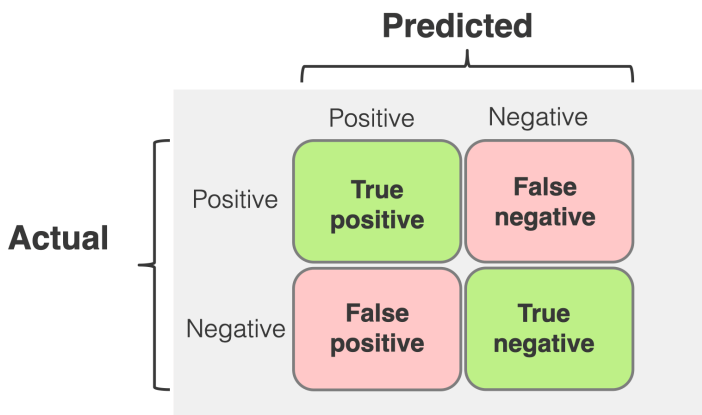
PD  
Talak

Experiment 8: 1/25/24	SIGNED BY: PALAK YADAV
-----------------------	------------------------

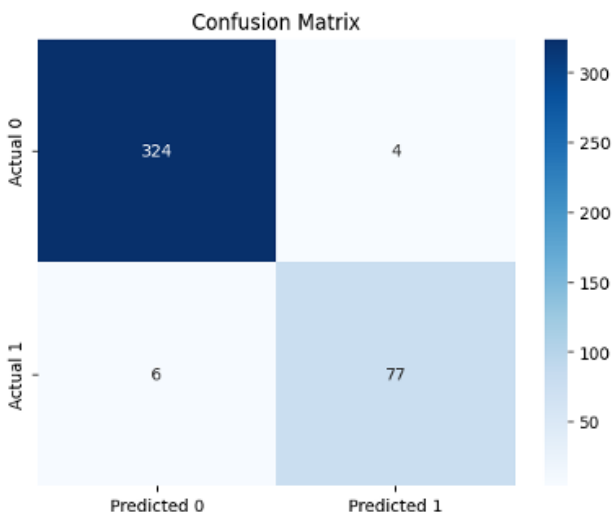
Objective: test all models on external dataset

The classification models were tested on other datasets of miRNA and their accuracy proved to be above 80% for the majority of the cases. Below are confusion matrices for each disease type and the significance of the values in each square.

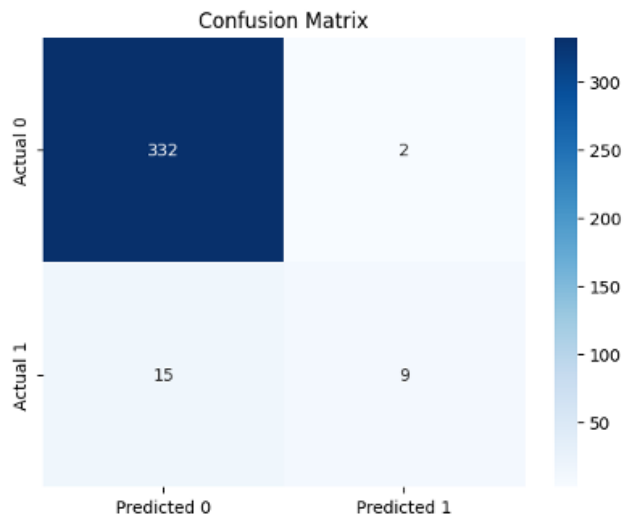
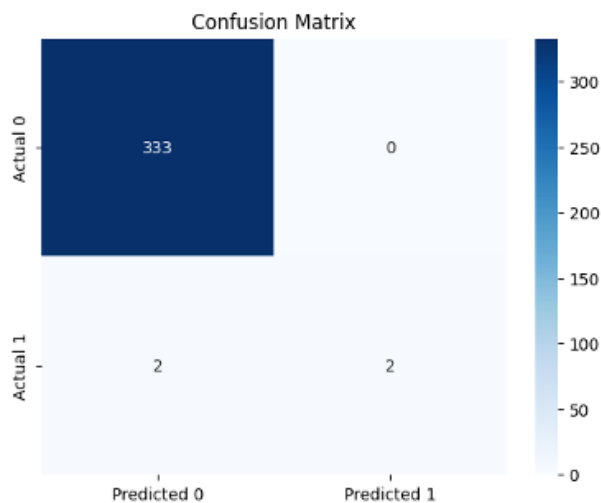
[How to interpret a confusion matrix for a machine learning model](#)



Ovarian Cancer



*Palak*

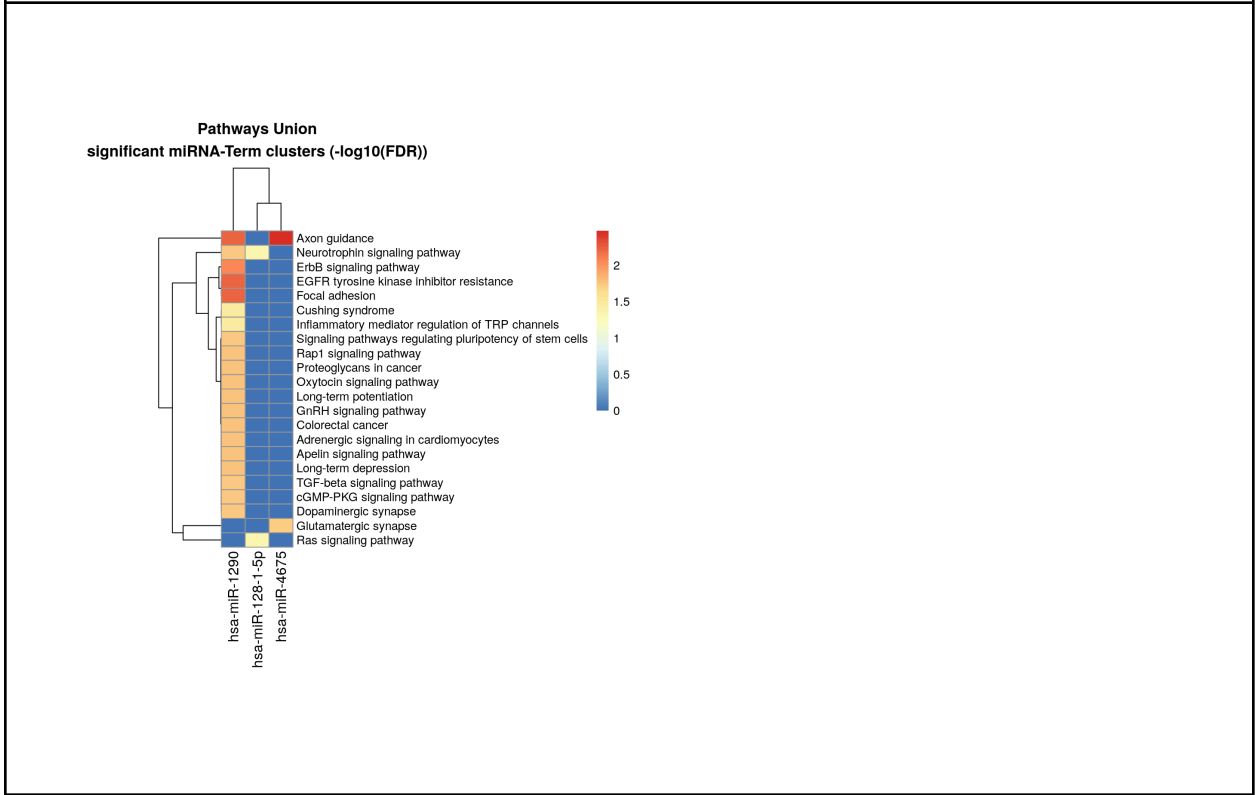
**Breast Cancer****Endometriosis****[Advantages and Disadvantages of Deep Learning - GeeksforGeeks](#)**

A deep neural network was used given that it prevents the likeliness of overfitting, which is often a limitation of simpler models such as logistic regression and random forest. Deep learning also allows for automatic feature selection which will be useful in the last stage for extracting significant miRNAs. It is also better apt at analyzing large and complex datasets, which is beneficial in this case with over 2000 samples from 5 different disease types

PD  
Palak

<b>Experiment 9: 1/28/24</b>	SIGNED BY: PALAK YADAV
------------------------------	------------------------

**Objective: Inset significant miRNA in pathway modeling**



<b>Experiment 10: 2/3/24</b>	SIGNED BY: PALAK YADAV
------------------------------	------------------------

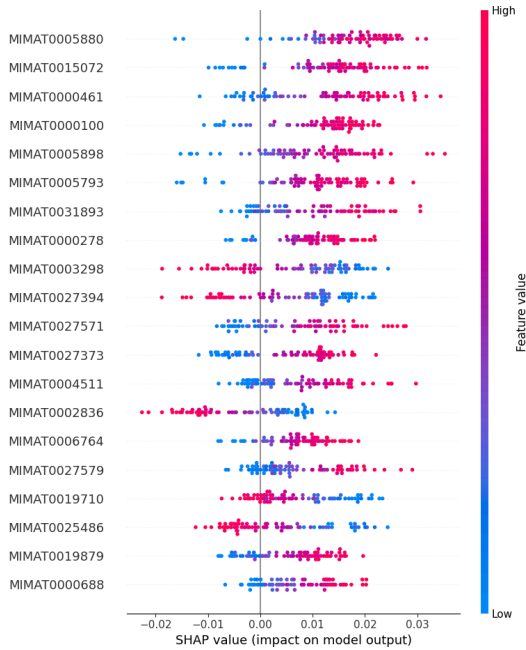
**Objective: Cros-verify significant feature miRNA from machine learning model to those from past studies to determine if there is consistency to some extent. Next, use pathway modeling software to explore miRNA-mediated pathways in select diseases.**

*Palak*

The DeepLIFT algorithm was applied on each predictive model to determine the unique miRNA profile of each disease as well as the shared miRNAs.

The graphs below show the top 10 features and the impact of their up or down-regulation in predicting each disease.

Ovarian cancer:



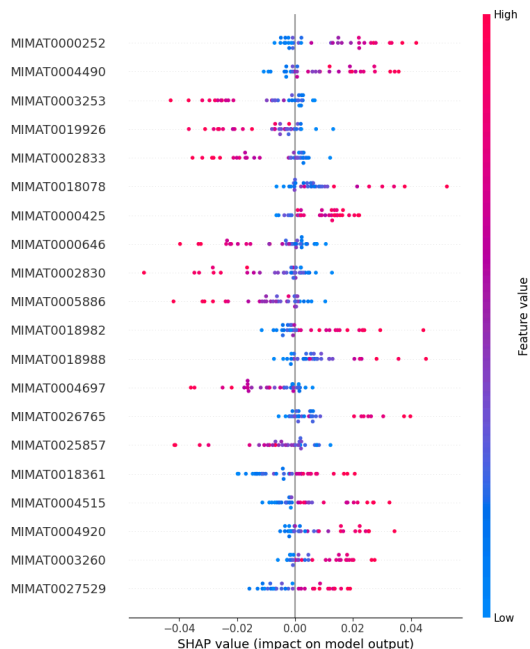
Ovarian Cancer Set

miRNA access key	miRNA name from <a href="https://mirbase.org/">https://mirbase.org/</a>
'MIMAT0019710'	hsa-miR-4648
'MIMAT0019879'	hsa-miR-4745-3p
MIMAT0025486	hsa-miR-6515-5p
'MIMAT0027579'	hsa-miR-6838-3p
'MIMAT0019710'	
'MIMAT0006764'	hsa-miR-320d
'MIMAT0002836'	hsa-miR-526b-3p
'MIMAT0004511'	hsa-miR-99a-3p

Palak

'MIMAT0027373'	hsa-miR-6736-5p
'MIMAT0027571	hsa-miR-6835-3p
MIMAT0005880	hsa-miR-1290
'MIMAT0015072	hsa-miR-320e
'MIMAT0000100	hsa-miR-29b-3p
MIMAT0000461	hsa-miR-195-5p
MIMAT0005898	hsa-miR-1246
'MIMAT0005793	hsa-miR-320c
'MIMAT0031893	hsa-miR-181b-2-3p
MIMAT0003298	hsa-miR-629-3p
MIMAT0000278	hsa-miR-221-3p
MIMAT0027394	hsa-miR-6747-5p
MIMAT0027571	hsa-miR-6835-3p
MIMAT0027373	hsa-miR-6736-5p
MIMAT0004511	hsa-miR-99a-3p
MIMAT0002836	hsa-miR-526b-3p
MIMAT0019864	hsa-miR-3064-5p
MIMAT0019879	hsa-miR-4745-3p
MIMAT0025486	hsa-miR-6515-5p
MIMAT0027579	hsa-miR-6838-3p

Breast Cancer feature miRNA:



Breast Cancer

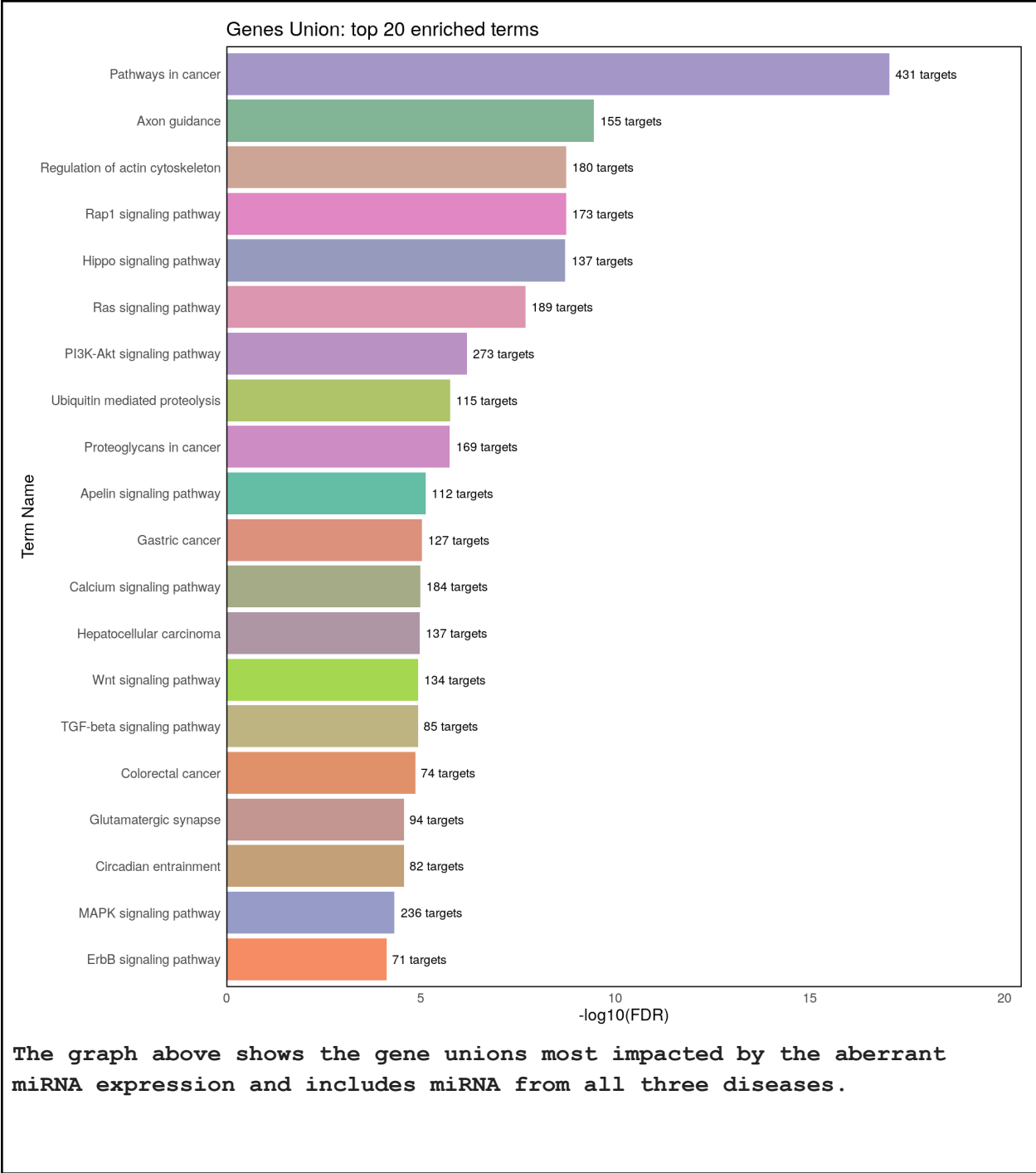
miRNA access key	miRNA name from <a href="https://mirbase.org/">https://mirbase.org/</a>
MIMAT0025857	hsa-miR-892c-5p
MIMAT0026479	hsa-miR-152-5p
MIMAT0002882	hsa-miR-510-5p
MIMAT0018361	hsa-miR-3945
MIMAT0004515	hsa-miR-29b-2-5p
MIMAT0004920	hsa-miR-541-3p
MIMAT0026765	hsa-miR-1537-5p
MIMAT0004697	hsa-miR-151a-5p
MIMAT0018982	hsa-miR-4460
MIMAT0018988	hsa-miR-4464
MIMAT0000425	hsa-miR-130a-3p

Dr. Parvati

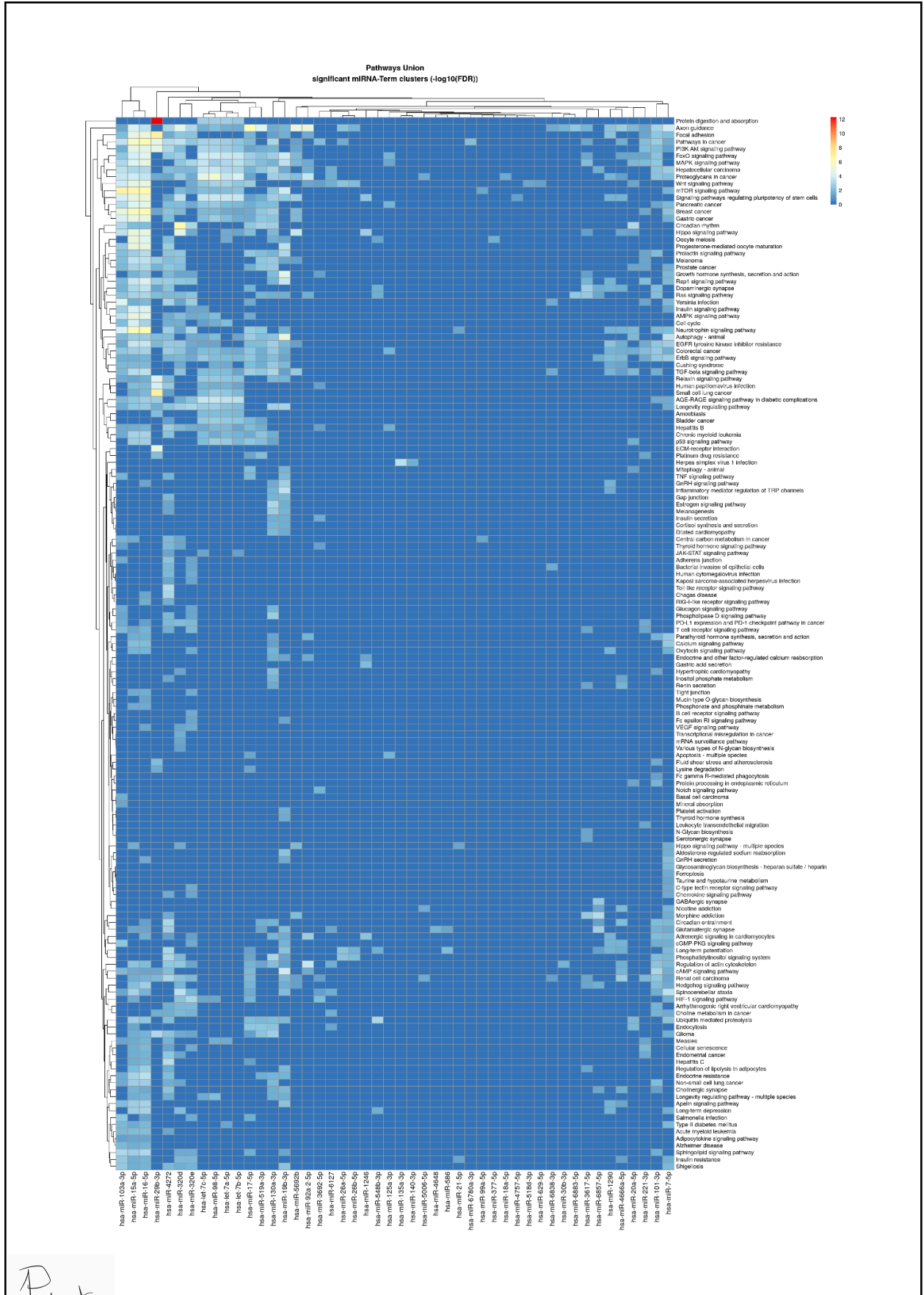


MIMAT0002830	hsa-miR-520f-3p
MIMAT0005886	hsa-miR-1297
MIMAT0000646	hsa-miR-155-5p
MIMAT0019926	hsa-miR-4772-5p
MIMAT0002833	hsa-miR-520a-5p
MIMAT0018078	hsa-miR-3658
MIMAT0004490	hsa-miR-19a-5p
MIMAT0003253	hsa-miR-587
MIMAT0000252	hsa-miR-7-5p

The DIANNE TOOLS software was used to model the pathways regulated by the aberrant expression of the select miRNA. The software is connected to the KEGG database and only accepts 200 inputs, so I inserted roughly 66 miRNA for each disease, incorporating those identified in this study as well as those found in previous works.



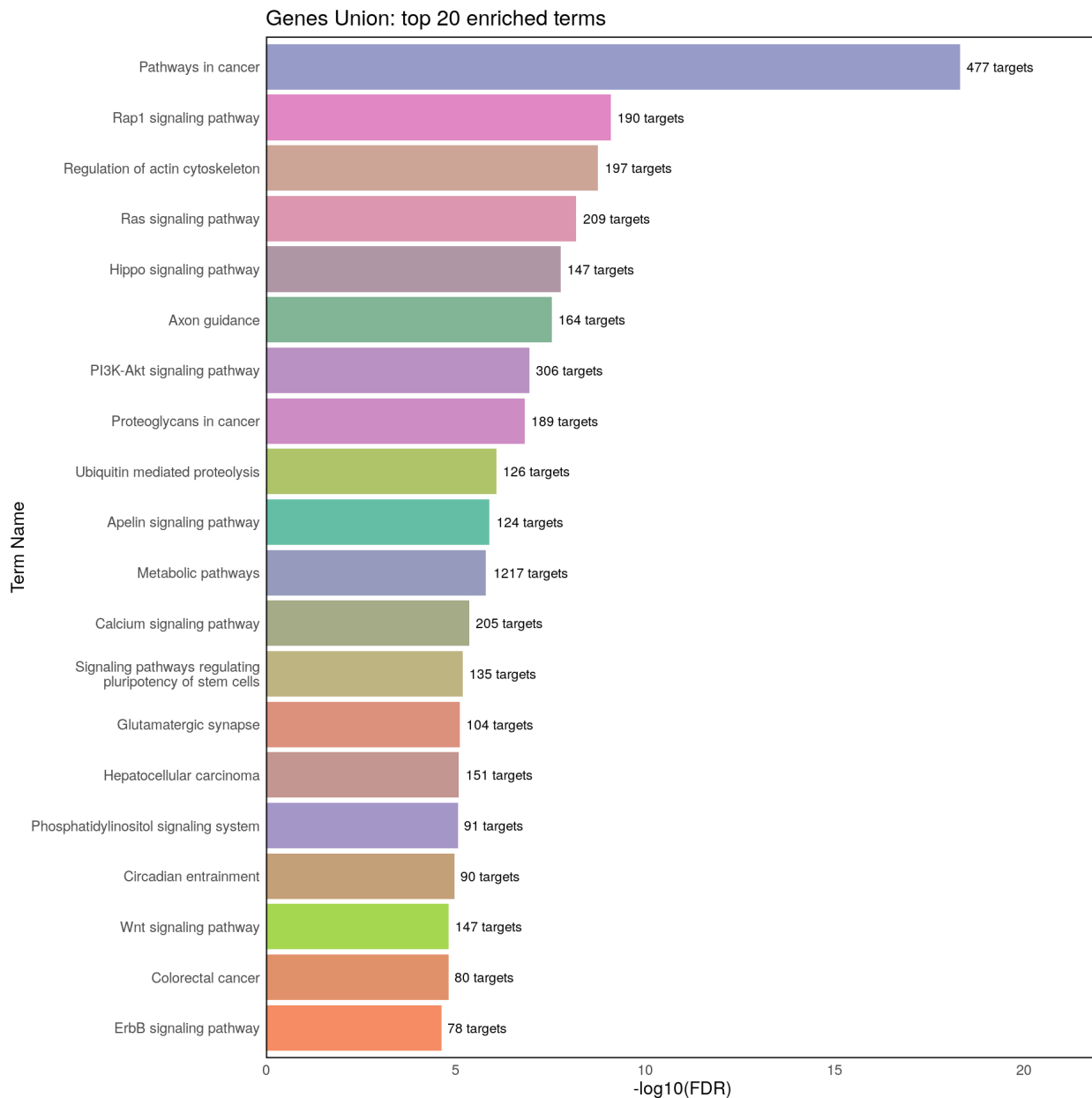
PD  
Palak



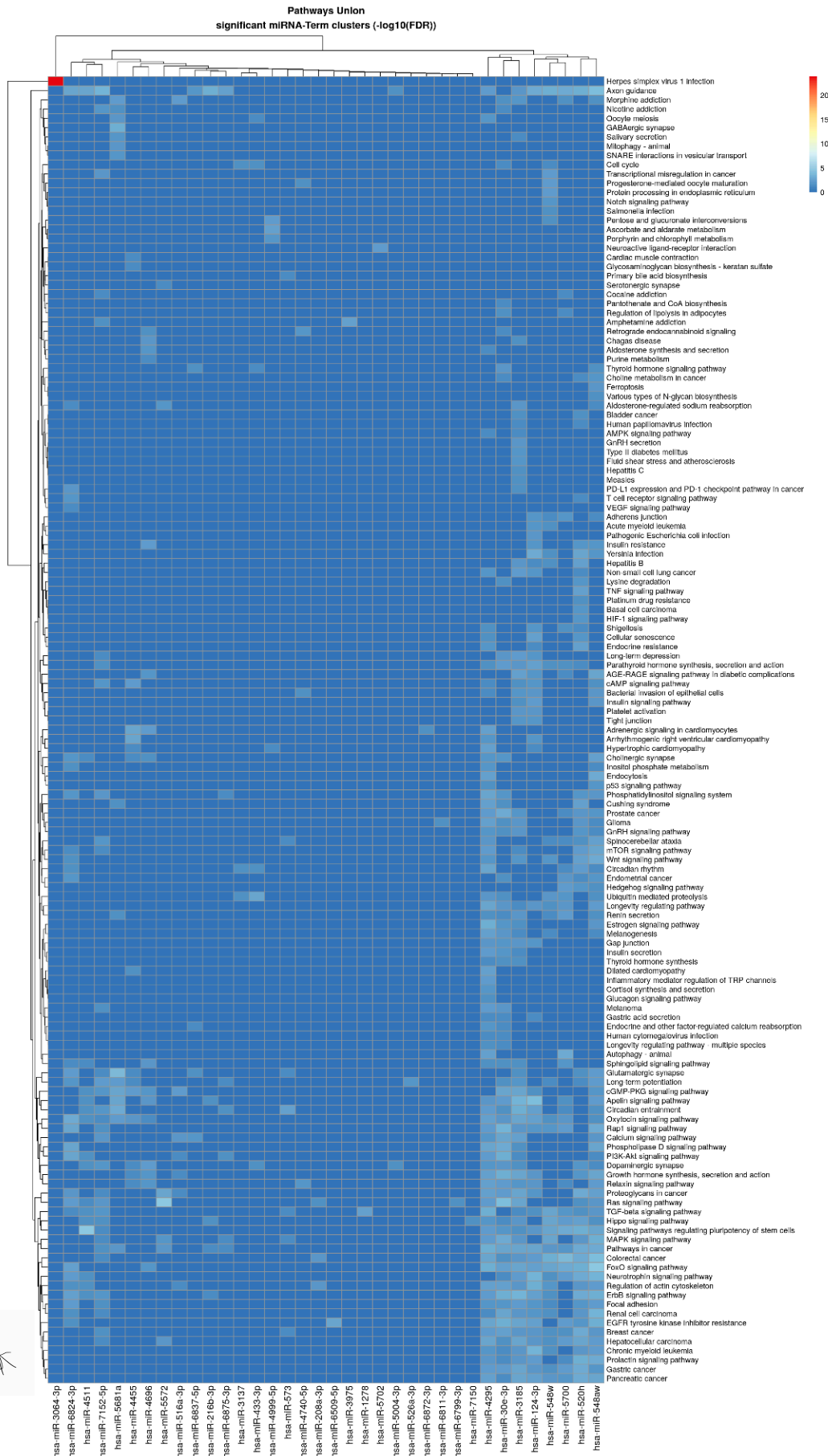
P. Talak

This shows the pathway intersection and their significance levels for each miRNA and the targeted pathway.

Endometriosis pathway - I chose to model only the endometriosis miRNA because I was curious to see if the ranking of the gene unions would differ. Interesting, a very similar range of pathways are affected in endometriosis as those found in breast cancer and ovarian cancer, indicating a shared pathology and increased risk for one having one disease lead to the development of the other.



PD  
Talak

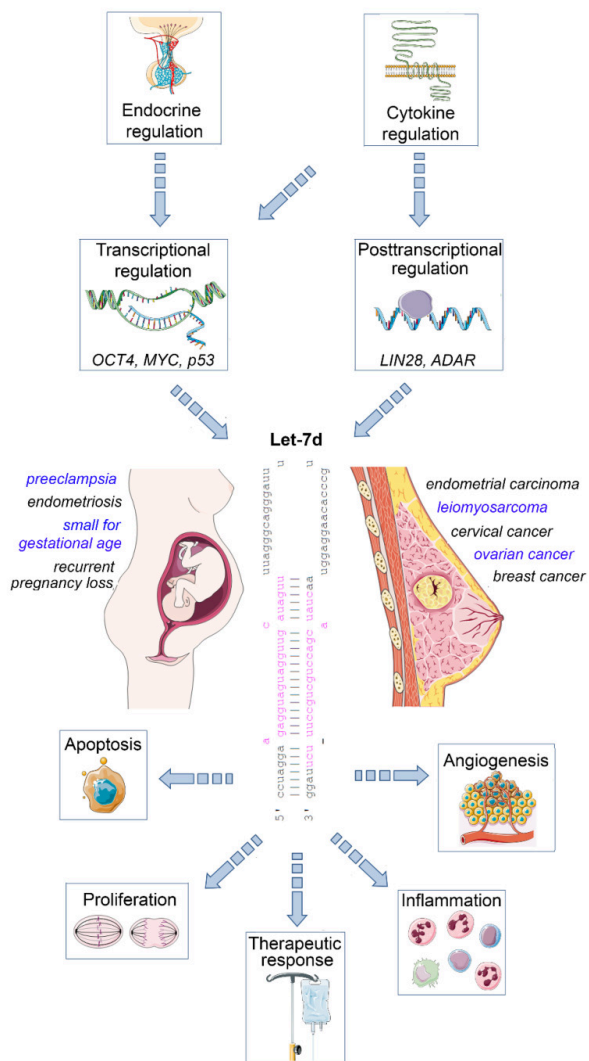


PD  
Tarak

**Biological significance:**

miRNAs that played a key role in the classification in each disease impacted the following pathways:

- Pathways in Cancer
  - Axon guidance → angiogenesis
  - RAP1 signaling pathway → tumor cell migration and invasion
  - Regulation of actin cytoskeleton → migration and mobility of cells
  - Ras signaling
  - Hippo signaling pathway
- Therapeutics could target dysregulated genes in each patient’s unique pathway
  - Diagnostic could seek for dysregulation for a specific miRNA



(De Santis, et. al 2021)

Patank

↑  
P  
Tarak