Capstone

Andrea Hannah
MMC
School of Applied Computational Sciences
handrea22@email.mmc.edu

Abstract— Identifying biomarkers that predict when patients are at risk for ovarian cancer, the fifth deadliest cancer for women, is the key to improved survival rates. Ovarian Cancer is a group of diseases that originate in the ovaries or in the fallopian tubes or peritoneum. Ovaries produce the female hormone and reproductive eggs. Ovarian cancer is best treated at its earliest stages before becoming aggressive. Therefore, early screening and diagnosis is key to successfully treating the disease, curing, or entering remission. Using the heatmap method of exploratory data analysis to assess correlations and using existing literature, the biomarkers of importance identified include Age, Menopause, Human Epididymis Protein 4 (HE4), Alkaline Phosphatase (ALP), and Calcium. Using Scatterplot visualizations, each contributing variable is assessed for comparison with elevated CA125 levels. All variables of interest, except HE4, correspond with elevated CA125 levels and would be biomarkers to pay closer attention to in ovarian cancer screening.

Keywords—Ovarian Cancer; Gynecological Health; Cancer Antigen; Big Data Visualizations; Gynecology; HE4; ALP; Ca; CA125

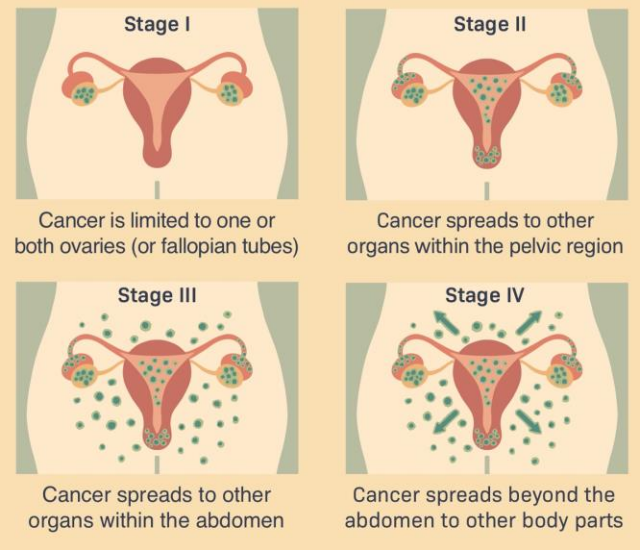

Figure 1

1 INTRODUCTION

The goal of this study is to identify biomarkers and other factors that effectively predict Ovarian Cancer to improve early diagnosis screening. Ovarian Cancer is a group of diseases that originate in the ovaries or in the fallopian tubes or peritoneum. Ovaries produce estrogen and progesterone hormones in females, as well as reproductive eggs. Ovarian Cancer is best treated at its earliest stages before becoming aggressive, as it is the fifth deadliest cancer to women. Therefore, early screening and diagnosis is key to successfully treating the disease, curing, or entering remission [1]. Presently, Ovarian Cancer is screened through two methods:

1. TVUS (transvaginal ultrasound, a test using sound waves to look at the uterus, fallopian tubes, and ovaries [2]. It can find a mass, but it cannot determine if a mass is malignant or benign. When it is used for screening, most of the masses found are not cancer.

2. The CA-125 blood test measures the amount of a protein called CA-125 in the blood [3]. Many women with ovarian cancer have high levels of CA-125. This test can be useful as a tumor marker to determine if treatment is working in women known to have Ovarian Cancer, because a high CA-125 level, which is above 35 U/ml, often goes down if treatment is working. The normal CA-125 range is 0-35 U/ml.

The CA-125 blood test measures the amount of the protein, Cancer Antigen – 125 in the blood. Many women with ovarian cancer do, in fact, have high levels of CA-125. This test can be useful as a tumor marker to help guide treatment in women known to have Ovarian Cancer, because a high level will often decrease if cancer treatment is working. Also, checking CA-125 levels has not been found to be as useful as a screening test for Ovarian Cancer because high levels of CA-125 is more often caused by common conditions such as endometriosis and pelvic inflammatory disease, and not everyone who has Ovarian Cancer has a high CA-125 level. Therefore, other biomarkers may perform better than CA-125 when screening for Ovarian Cancer.

This study seeks to determine the best predictive biomarkers in a dataset for women by using machine learning to help predict Ovarian Cancer risk and as a result, substantiate a case to continue study meant to increase early detection for the improvement of survivor outcomes in marginalized populations. To accomplish these objectives, the study hypothesis that multi-biomarker prediction is more effective than using CA-125 alone to predict Ovarian Cancer in women must be found acceptable given the model results.

2 LITERATURE REVIEW

Using CA-125 levels as a measure has not been found to be the most useful method to screen for Ovarian Cancer, since high

levels of CA-125 can be associated with other common conditions besides cancer, such as endometriosis and pelvic inflammatory disease. Additionally, low CA-125 levels in Ovarian Cancer patients are not uncommon. In addition to metabolomics, total metabolites are found in a biological sample of cell, tissue, or organism, and in this case blood or serum. Some researchers believe proteomics, expressed proteins collected at a given time from the sample found in blood and serum tests, serve as a non-invasive approach to diagnose and monitor Ovarian Cancer. Therefore, other biomarkers should be used in conjunction with or in addition to CA-125 to improve prediction of ovarian cancer with machine learning algorithms.

2.1 Biomarkers

Farinella et al. [4] claims that biomarkers such as age or menopausal state result in only weak improvements in model accuracy, making it necessary to identify novel proteomic or protein-based molecular signatures to develop new predictive algorithms for High Grade Serous Ovarian Cancer (HGSOC), one of the most aggressive forms of Ovarian Cancer, diagnosis. Using an open source HGSOC proteomic data set to develop the decision support system or DSS, the researchers deployed a double step feature selection and decision tree that resulted in them being able to distinguish between tumor and non-tumor patient based on the differentiation between three proteins, Topoisomerase 1 (TOP1), Protein Disulfide Isomerase Family A Member 4 (PDIA4) and Osteoglycin (OGN). The resulting DSS performed at 98.2% accuracy. The most promising proteins to monitor in clinical practice for HGSOC were PDIA4 and OGN, while TOP1 tends to be overexpressed in Epithelial Ovarian Cancer or EOC patients. Additionally, Farinelli et. al finds that seral biomarkers CA-125 and HE4 did not correlate with a tumor phenotype.

2.2 Machine Learning Models of Interest

The Lu et al. [5] study was based on 349 Chinese patients to predict Ovarian Cancer (OC) or Benign Ovarian Tumors (BOT). For one part of the data, Minimum Redundancy – Maximum Relevance (MRMR) feature selection method was applied on the 235 patients' data (89 BOT and 146 OC) to select the most relevant features. For the second part of the data, a Decision Tree model was constructed and tested on the rest of the 114 patients (89 BOT and 25 OC). The decision tree model results were compared with the predictions produced by using the risk of ovarian malignancy algorithm (ROMA) and logistic regression model. Eight notable features were selected by MRMR in comparison to two identified as the top features by the decision tree model: human epididymis protein 4 (HE4) and carcinoembryonic+ antigen (CEA), which is a valuable marker for Ovarian Cancer prediction in patients with low HE4. Using their respective features, the decision tree model yielded better prediction results than risk of ovarian malignancy algorithm (ROMA) and logistic regression model.

These studies demonstrate machine learning can be used to identify ovarian cancer and its stages. However, researchers in the Taleb et. al study posit that most modern research studies on ovarian cancer use a single classification model, leading to poor performance in diagnosis. More sophisticated Machine learning algorithms, such as Support vector machine (SVM) and K-Nearest Neighbor (KNN) are employed in the Taleb et. al [6] study. SVM outperformed KNN in both training and validation performance and achieved an accuracy of 98.1% and 97.16% for training and validation, respectively. If used in medical diagnosis systems, the proposed model can significantly improve the accuracy of ovarian cancer detection leading to effective treatment and an increase in patient survival rates.

2.3 SVM Cross-Validation

An older study by Guan et al. [7] examines how to detect ovarian cancer using metabolomic liquid chromatography/mass spectrometry data by support vector machines. There were three evaluation processes for the SVM performance models. The processes of leave-one-out-cross-validation, 12-fold-cross-validation, and 52-20-split-validation were used to examine the SVM model performance of differentiating control vs. disease serum samples and exhibited over 90\% accuracy. This was a substantially lower performance accuracy than found with the SVM and KNN models in the Taleb study's 97.16% and 98.1% for the same models [6].

2.4 Alternatives with Deep Learning

Ziyambe et. al. [8] takes a single modal data approach by proposing a a novel convolutional neural network (CNN) algorithm for predicting and diagnosing ovarian cancer by training the algorithm on a histopathological image dataset. This CNN model achieved an accuracy of 94% overall with 95.12% of cancerous cases identified and 93.02% of healthy cells classified accurately. Ghoniem et al. [9] research expands beyond the use of single modality data with their deep learning models by using a multimodal fusion framework joins gene modality with histopathological image modality. Accuracy scores ranged between 95-99% outperforming the Ziyambe [8] single modal data approach. This study found that better performance with the highest performance accuracy and lowest classification error rates were yielded from multi-modal data. The researchers set up a deep feature extraction at work for each modality including a predictive ant-lion optimize long short term memory (LSTMM) model to process gene longitudinal data and a predictive ant-lion optimized convolutional neural network (CNN) model to process histopathology images.

3 DATASET AND FEATURES

The original Ovarian Cancer Dataset has 349 observations. The data is multivariate with 51 variables including Patient ID, Age, Menopause (Y/N), and Type, as well as 47 biomarkers found in blood or serum samples [12]. Patient ID has no algorithmic value and was removed from the cleaned dataset, leaving 50 potential variables of algorithmic value. The original dataset includes the following biomarker features:

Index	Variable	Description	Index	Variable	Description
1	AFP	Alpha-fetoprotein	26	HE4	Human Epididymis Protein 4
2	AG	Anion Gap	27	HGB	Hemoglobin
3	Age	Age	28	IBIL	Indirect Bilirubin
4	ALB	Albumin	29	K	Potassium or Kalium
5	ALP	Alkaline Phosphatase	30	LYM#	Lymphocyte Count
6	ALT	Alanine Aminotransferase	31	LYM%	Lymphocyte Ratio
7	AST	Aspartate Aminotransferase	32	MCH	Mean Corpuscular Hemoglobin
8	BASO#	Basophil Cell Count	33	MCV	Mean Corpuscular Volume
9	BASO%	Basophil Cell Ratio	34	Menopause	State of Menopause
10	BUN	Blood Urea Nitrogen	35	Mg	Magnesium
11	Ca	Calcium	36	MONO#	Mononuclear Cell Count
12	CA125	Carbohydrate Antigen 125 or Cancer Antigen 125	37	MONO%	Monocyte Ratio
13	CA19-9	Carbohydrate antigen 19-9 or Cancer Antigen 19-9	38	MPV	Mean Platelet Volume
14	CA72-4	Carbohydrate Antigen 72-4 or Cancer Antigen 72-4 (removed from cleaned dataset)	39	Na	Sodium or Natrium
15	CEA	Carcinoembryonic Antigen	40	NEU	Neutrophil Ratio
16	CL	Chlorine	41	PATIENT ID	Patient ID (removed from cleaned dataset)
17	CO2CP	Carbon Dioxide-Combining Power	42	PCT	Thrombocytocrit or Plateletcrit
18	CREA	Creatinine	43	PDW	Platelet Distribution Width
19	DBIL	Direct Bilirubin	44	PHOS	Phosphorus
20	EO#	Eosinophil Count	45	PLT	Platelet Count
21	EO%	Eosinophil Ratio	46	RBC	Red Blood Cell Count
22	GGT	Gamma-Glutamyl Transferase	47	RDW	Red Blood Cell Distribution Width
23	GLO	Globulin	48	TBIL	Total Bilirubin
24	GLU	Glucose	49	TP	Total Protein
25	HCT	Hematocrit	50	TYPE	Benign or Malignant
			51	UA	Uric Acid

Table 1

The cleaned dataset will leave us with 235 patient observations and 49 features, because CA72-4 was the only blood or serum-based biomarker subsequently removed, decreasing the variable count from 50 to 49.

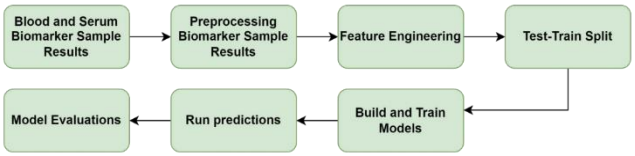

Figure 2

The data was already clean, but as part of preprocessing, risk levels were applied to CA-125, where measures above 35 U/ml (units per milliliter) were considered higher risk as unhealthy or indicative of cancer or some other condition and levels below that measure, were considered healthy or lower risk. Additionally, to begin to understand the cleaned dataset, we created a Heatmap and scatterplots to explore the correlations of the variables before proceeding with the next step in the project pipeline of feature engineering as shown in Figure 2. The heatmap shows which variables have correlations.

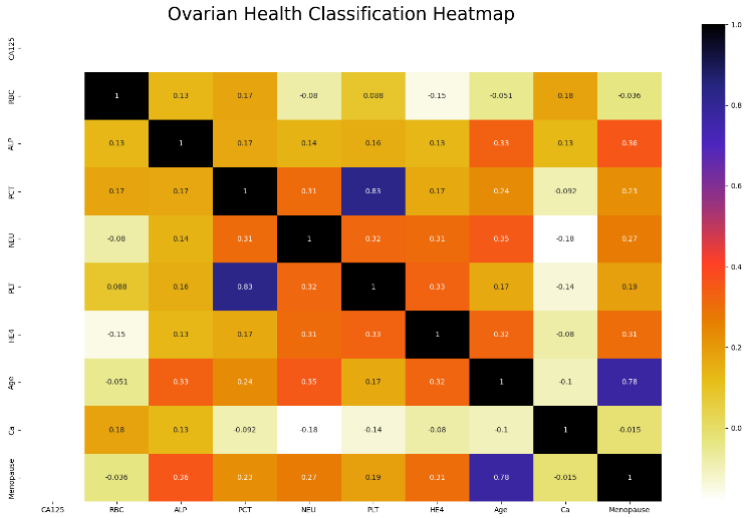

Figure 3

With any research using this dataset, we would drop the highly correlated features shown in this heatmap from the model. PCT and PLT were dropped.

Through exploratory data analysis, a relationship between PCT (Thrombocytocrit or Plateletcrit) and PLT (Platelet Count) is visible. Age and Menopause share an expected correlation. PCT and PLT are measures of essentially the same thing through different methods, so these variables will be dropped.

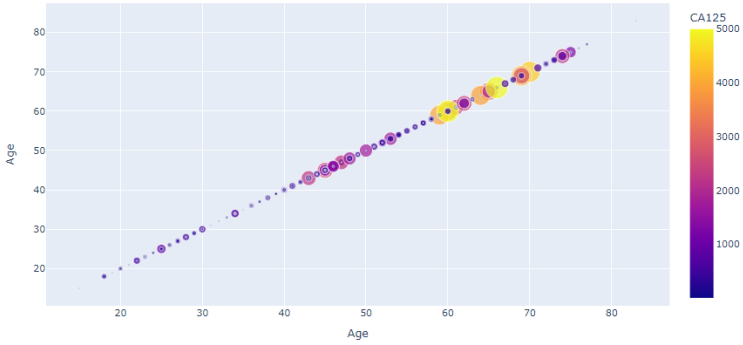

Figure 4

It appears that elevated CA125 levels show at age 59 for this dataset. The prevailing notion that ovarian cancer shows up for women 63 years or older is dangerous and it is potentially a factor in late screening and diagnosis.

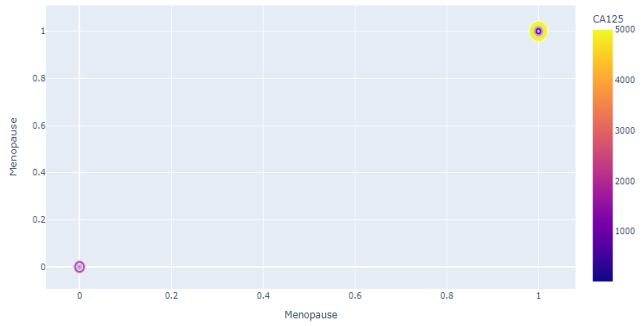

Figure 5

When looking at the relationship between Menopause and CA125, there is a relationship of higher CA125 levels with the presentation of menopause. However, there may be some outliers, as low CA125 values are also represented with the presence of menopause and vice versa when the onset of menopause has not begun.

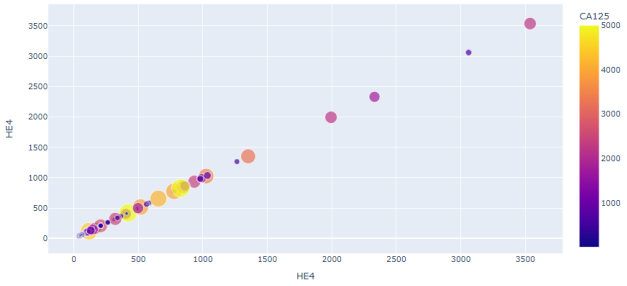

Figure 6

Finally, Human Epididymis Protein 4 (HE4) has an inverse relationship with CA125 other than expected. High levels of HE4 are not associated with higher levels of CA125 as expected. Rather, lower levels would have a stronger connection if not for the fact that there are so few scatter plot points as HE4 increases, so that would be disputable and hopefully, proved otherwise with a larger dataset. Scientists have stated the initial promise for HE4 in early Ovarian Cancer detection, but this dataset cannot support this claim at its present size.

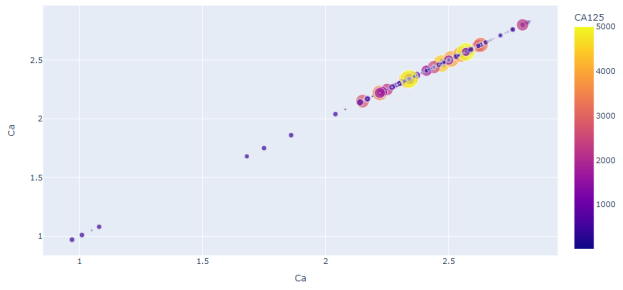

Figure 7

In examining the relationship between Ca (Calcium) with CA125, notice the elevation in CA125 corresponds with an increase in Ca.

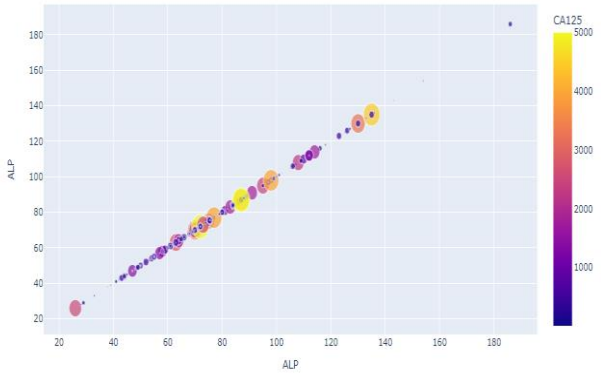

Figure 8

In examining the correlation between ALP (Alkaline Phosphatase) with CA125, notice the elevation in CA125 corresponds with an increase in ALP.

4 BRIEF VISUALIZATION METHODS RESULTS & FEATURE ENGINEERING

The Heatmap visualization shows which variables are correlated with one another. As a result, PCT and PLT were removed from the variables of interest. The Scatterplot visualization is used to show the relationship between variables. Using this method, the visualization results yielded:

1. Higher Ca levels are associated with higher CA125 levels of risk.
2. Higher ALP levels are associated with higher CA125 levels of risk.
3. Higher HE4 levels are not supported as related with higher CA125 levels of risk with this dataset. Instead, lower HE4 levels are.
4. As expected, menopause and age correlate with higher CA125 levels, but will these associations and those above be present in our mRMR feature selections for our target variable, Ovarian Cancer status?

To ensure the top ten features related to the presence of cancer were used, the feature selection tool of mRMR was helpful. Minimum redundancy maximum relevance (mRMR) feature selection is an algorithm that is commonly used to accurately identify or narrow down relevant features in a data set by selecting A subset of features that have the most correlation with the output and the least correlation between each other.

The mRMR algorithm was used to narrow down the 49 biomarker features to the top 10 selections: 'Age', 'CREA', 'LYM%', 'AST', 'CA125', 'PDW', 'Menopause', 'NEU', 'CEA', and 'LYM#'. From this, we see that 'ALP', 'HE4', and 'Ca' are

not in the top features defining the X-variable(s) for the first set of predictive models. In the second set of chosen features, some of which are supported by the literature, defining the X-variable(s) for the predictive models, we include: 'CA125', 'RBC', 'ALP', 'PCT', 'NEU', 'PLT', 'HE4', 'Age', 'Ca', 'Menopause'. Finally, 'CA125' alone serves as the third chosen feature selection, defining the X-variable for the final set of predictive models.

5 METHODS

Using the top biomarkers and features selected by mRMR, machine learning algorithms can help with prediction of Ovarian Cancer. The SVM, KNN, and Decision Trees Classifiers, supervised machine learning algorithms, which use labeled data sets to train algorithms that classify data or predict outcomes accurately, were deployed.

SVM or Support Vector Machine is a linear classification and regression algorithm that creates a line or a hyperplane between data of two classes. The SVM takes the Ovarian Cancer dataset as an input and the output produces a line that separates where the points are closest to the line from both classes when possible. Then the distance between the line and support vectors are computed as margin, which are maximized for the optimal hyperplane. This creates a decision boundary, where the separation of the two classes is at its widest.

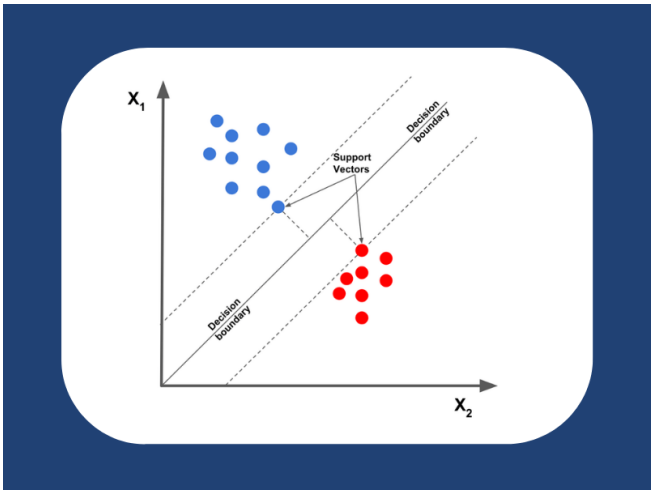

Figure 9

KNN or K Nearest Neighbor is a simpler machine learning algorithm, which classifies data points based on its similarity with how its neighbor or earlier stored data points are classified. KNN works well for labeled, small datasets. The K represents the number of nearest neighbors used to classify new data points. Adjusting the value of K is a function of parameter tuning; using the square root of the target testing data, K could be identified as 9.

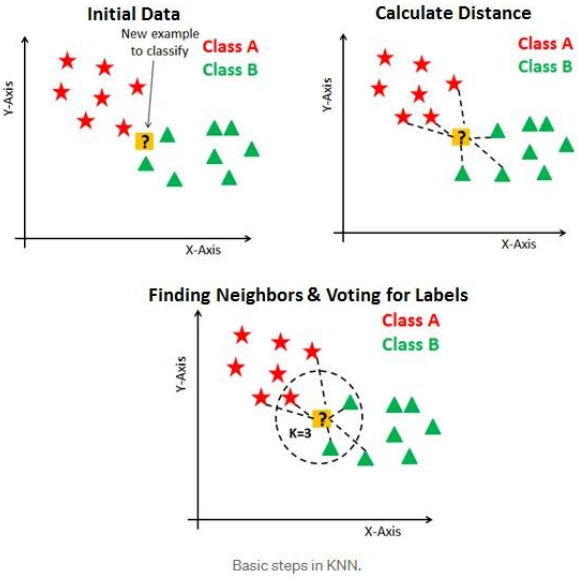

Figure 10

Decision trees use a tree-like flowchart model of decisions and the corresponding possible consequences. A decision tree consists of three types of nodes:

1. Root Node – represents the feature all other nodes split from

2. Decision Nodes – represent a test on a feature or attribute for a decision to be made on. As the tree depth increases, the loss entropy should decrease, and the information gain should increase until we end up with a pure leaf or end node

3. Leaf/End Nodes – represents the outcome with reduced uncertainty

It is advantageous to use decision trees because they are easily interpretable and understandable after a brief explanation. However, a disadvantage is that model calculation can become complex if many values are uncertain. Since our data has no missing values, there is little to no uncertainty to impact decision trees as a selected model. Accuracy of the decision tree model is increased when the depth increases.

Machine Learning Models Comparison:

Support Vector Machine

- Lower computation power needed
- High accuracy
- Works for classification and regression problems, but very popular with classification

K Nearest Neighbor

- Ease of interpretation
- Low calculation time

- Works for classification and regression problems, and can have very effective predictive power

Decision Trees Classifier

- A single decision tree has faster computation
- Uses rules to predict from input that is a dataset with features
- May experience overfitting if maximum depth is reached

Note, the test-split is based on a 70/30 ratio for the dataset, and the evaluation of model performance will primarily focus on the Accuracy metric.

6 MODEL PERFORMANCE RESULTS

Table 2, below, shows the model performance results for the features engineered for the study.

Model	Accuracy	Precision	Recall	F-1 score	Rank
SVM 1	79%	69%	86%	76%	4
SVM 2	86%	75%	96%	74%	3
SVM 3	87%	83%	86%	84%	2
KNN 1	77%	69%	79%	73%	5
KNN 2	77%	71%	71%	71%	6
KNN 3	90%	80%	100%	89%	1
DT 1	69%	62%	54%	58%	9
DT 2	75%	69%	64%	67%	7
DT 3	73%	67%	64%	65%	8

Model Key by Features Extraction Method
1: mRMR selected features - 'Age', 'CREA', 'LYM%', 'AST', 'CA125', 'PDW', 'Menopause', 'NEU', 'CEA', and 'LYM#'
2: Literature supported features – 'CA125', 'RBC', 'ALP', 'PCT', 'NEU', 'PLT', 'HE4', 'Age', 'Ca', 'Menopause
3: mRMR selected features excluding 'CA125' - 'Age', 'CREA', 'LYM%', 'AST', 'Menopause', 'PDW', 'NEU', 'HE4', 'LYM#', 'PCT'

Table 2

This work produced 10 features beneficial in predicting Ovarian Cancer using mRMR in our first ser of biomarkers, and those include: 'Age', 'CREA', 'LYM%', 'AST', 'CA125', 'PDW', 'Menopause', 'NEU', 'CEA', and 'LYM#'. The best performing model with these selected biomarker features from the original biomarker selection is SVM at 79% overall accuracy and the worst performing was Decision Trees at 69% overall accuracy.

For the second set of biomarkers, SVM was the best performing predictive model using literature supported features: 'CA125', 'RBC', 'ALP', 'PCT', 'NEU', 'PLT', 'HE4', 'Age', 'Ca', 'Menopause'. SVM and Decision Trees model using these features performed better at 86% and 75% overall accuracy, than the mRMR selected features for their counterpart models.

For the third set of biomarkers, accuracy for mRMR selected features excluding CA-125 (['Age', 'CREA', 'LYM%', 'AST', 'Menopause', 'PDW', 'NEU', 'HE4', 'LYM#', 'PCT']) which ends up also excluding CEA and replaces the two variables with HE4 and PCT, shows SVM performance is 87%, KNN is 90%, and Decision Trees is 73%. KNN is the better performing model for variables selected by mRMR with targeted exclusion of CA-125.

Finally, model accuracy for CA-125 alone shows performance for SVM is 73%, KNN is 65%, and Decision Trees is 69%. SVM is the better performing model when the only variable is CA-125. SVM, KNN, and DT for this singular variable practiced in clinics are outperformed by their counterparts for mRMR selected features excluding CA-125 for accuracy.

Decision Trees only performed better with CA-125 included in the variable/feature selection from previous model runs but given that SVM and then KNN appear to be the more superior performing models throughout the entirety of the model runs for the most part, the mRMR selected features SVM and KNN models that exclude CA-125 are the best model and feature selection combination presented. Thus, the model results support the use of a multi-biomarker approach to perform early diagnosis of Ovarian Cancer rather than relying on CA-125 alone.

A larger dataset is required to focus on postmenopausal women as a subset of the data, and this could yield varied results in the mRMR feature selection process.

7 CONCLUSION AND FUTURE WORK

The study results allow acceptance of the hypothesis that multi-biomarker prediction is more effective than using CA-125 alone to predict Ovarian Cancer in women. Rather than solely relying on CA-125-only test for early diagnosis and screening, implementing the use of a multi-biomarker screener in clinical environments could help eliminate disparities in screening/diagnosis, treatment, and survival rates for Ovarian Cancer patients of targeted populations. Additional features could be used to screen patients, including the following:

- Racial or ethnic identification
- Number of pregnancies
- Number of c-sections
- Status of other gynecological conditions or reproductive system conditions
- Weight (BMI, obesity)
- Inherited factors
- Diet and nutrition (calcium intake, vitamin D)
- Lifestyle (sun exposure, exercise)
- Stress factors

These missing social and environmental factors and others are features which could improve research and help with eliminating disparities in clinical practice for those with known factors that contribute to poor health outcomes. Additional datasets with these attributes can be analyzed for future study. In addition to the above, increasing the number of observances, i.e. patients, in our dataset will help improve model performance in identifying the most important biomarkers. Given the importance of menopause as a milestone condition for women, a larger dataset could focus on postmenopausal women, as this study's dataset would have been left with only 90 observances if postmenopausal women were the primary focus.

A future study could also analyze and predict comorbidities in order to predict the composition of the wholistic care team of specialists (heart-renal-metabolic, etc.) needed that may be required for particular cancer patients to improve survival

rates by starting with a larger dataset requested from the National Cancer Institute [11], a similar research database, or create one in-house by adding a data sharing clause in patient forms.

A primary motivation of this work is to eliminate disparities in screening/diagnosis, treatment, and survival rates for ovarian cancer patients. Since the dataset is not as large, another motivation is to collect more data by establishing an in-house study at a historical African American and Black-serving institution, like Meharry, given that 5-year relative survival rate for ovarian cancer increased from 33% to 48% among non-Hispanic White women but decreased from 44% to 41% in African American women [12]. The study might best be served as a joint collaboration between Meharry institutes: the Center for Advanced Scientific Computing and Innovation, the Center of Women's Health, and eventually the Center of Health Policy, since there are disparities in ovarian screening and treatment costs under Medicaid. Alternatively, a study could be established at another institution sharing a mandate for improved equity in African American health.

REFERENCES

[1] ACS, "Key statistics for ovarian cancer," All About Cancer, 2023.

[2] ACS, "Can ovarian cancer be found early?," All About Cancer, 2023.

[3] C. T. C. of America, "CA-125 blood test," City of Hope, Nov 2021.

[4] F. Farinella, M. Merone, L. Bacco, A. Capirchio, M. Ciccozzi, and D. Caligiore, "Machine learning analysis of high-grade serous ovarian cancer proteomic dataset reveals novel candidate biomarkers," Scientific Reports, vol. 12, no. 1, p. 3041, 2022.

[5] M. Lu, Z. Fan, B. Xu, L. Chen, X. Zheng, J. Li, T. Znati, Q. Mi, and J. Jiang, "Using machine learning to predict ovarian cancer," International Journal of Medical Informatics, vol. 141, p. 104195, 2020.

[6] N. Taleb, S. Mehmood, M. Zubair, I. Naseer, B. Mago, and M. U. Nasir, "Ovary cancer diagnosing empowered with machine learning," in 2022 International Conference on Business Analytics for Technology and Security (ICBATS), pp. 1–6, IEEE, 2022.

[7] W. Guan, M. Zhou, C. Y. Hampton, B. B. Benigno, L. Walker, A. Gray, J. F. McDonald, and F. M. Fern´andez, "Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines," BMC bioinformatics, vol. 10, no. 1, pp. 1–15, 2009.

[8] B. Ziyambe, A. Yahya, T. Mushiri, M. U. Tariq, Q. Abbas, M. Babar, M. Albathan, M. Asim, A. Hussain, and S. Jabbar, "A deep learning framework for the prediction and diagnosis of ovarian cancer in pre- and post-menopausal women," Diagnostics, vol. 13, no. 10, 2023.

[9] R. M. Ghoniem, A. D. Algarni, B. Refky, and A. A. Ewees, "Multi-modal evolutionary deep learning model for ovarian cancer diagnosis," Symmetry, vol. 13, no. 4, 2021.

[10] Shahane, S. (2021, February 6). Predict ovarian cancer. Kaggle. Retrieved March 31, 2023, from https://www.kaggle.com/datasets/saurabhshahane/predict-ovarian-cancer

[11] February 23, 2023, January 20, 2023, & March 9, 2023. (n.d.). Ovarian cancer studies aim to reduce racial disparities. National Cancer Institute. Retrieved March 31, 2023, from https://www.cancer.gov/news-events/cancer-currents-blog/2020/ovarian-cancer-racial-disparities-studies

[12] Ovarian - datasets - PLCO - the cancer data access system. National Cancer Institute - Cancer Data Access System. (n.d.). Retrieved March 31, 2023, from https://cdas.cancer.gov/datasets/plco/23/

[13] Editorial Team. "What Are the Stages of Ovarian Cancer?" AdvancedOvarianCancer.Net, advancedovariancancer.net/diagnosis/stages. Accessed 2023. (Figure 1).

[14] Arora, Avi. "8 Unique Machine Learning Interview Questions on SVM." Analytics Arora, 30 June 2022, analyticsarora.com/8-unique-machine-learning-interview-questions-on-svm/. (Figure 9).

[15] R, Deepthi A. "KNN Visualization in Just 13 Lines of Code." Medium, Towards Data Science, 24 Sept. 2019, towardsdatascience.com/knn-visualization-in-just-13-lines-of-code-32820d72c6b6. (Figure 10).