Identification of a Panel of Biomarkers for the Early Detection of Ovarian Cancer

Riya Davar[1] and Madhuri Yalamanchili[#]

[1]Texas Academy of Mathematics and Science, Denton, TX, USA
[#]Advisor

ABSTRACT

According to the CDC, in the United States, Ovarian Cancer is the second most prevalent form of gynecologic cancer and is the fifth leading cause of mortality in women. The only reliable method to screen for this cancer is TVS (transvaginal sonography), which is both invasive and costly. The goal of this project was to use the mRMR (Maximum Relevance Minimum Redundancy) Feature Selection Algorithm to select a panel of biomarkers from the Ovarian Cancer dataset and create a non-invasive and inexpensive software tool that could help validate the panel and assist with the early detection of Ovarian Cancer, with a reasonable level of sensitivity. This project uses an ovarian cancer dataset with 49 features. The mRMR filter method [9, 10, 12]of feature selection eliminates the redundant features while keeping the relevant features that impact the target class. This project accomplished the final goal of creating a working web application that asks a clinician to provide a few basic blood test results and generates a prediction. The machine learning model [7] used by the application is Random Forest Machine Learning model which is created with the K best features picked by the mRMR algorithm and is successfully utilized to predict the disease and treatment targets thus helping with reducing the mortality rate from ovarian cancer. This project used the Random Forest Classifier model machine learning model. It has been shown to work well with smaller datasets (as with this project s dataset) and had a sensitivity score of 0.96.

Background

Treatment is most effective when Ovarian Cancer is discovered in its **early stages** - in fact, only 20% of all cases are found early, meaning in stage I or stage II; if the cancer is caught in stage III or higher, the survival rate can be as low as 28%." (National Ovarian Cancer Coalition, 2022). While there are known risk factors like menstrual age, genes, family history, obesity, hormone replacement therapy, and endometriosis that can increase the chances of Ovarian Cancer, there is no convenient method to catch it in the early stages. [16,17]

The CA-125 blood test is one of the most common ovarian cancer screening tests, as many ovarian cancer patients do have a high level of CA-125. The problem with using CA-125 as a screening test is that it can be caused by illnesses other than cancer. Previous research done shows CA-125, CEA and HE4 [11] as some of the important biomarkers used for screening of ovarian cancer. However, for practical purposes, a solution that can predict the onset of the disease using the biomarkers commonly available at a physician's or a gynecologist's office would tremendously help with catching the disease at an early stage resulting in a successful clinical treatment outcome.

Method

This project has two phases: In the first phase, the project determines the best biomarkers using the mRMR [9, 10, 12] algorithm. The project's second phase deduces an optimal set from the list of selected biomarkers from Phase I.

Phase I

Feature Selection Method

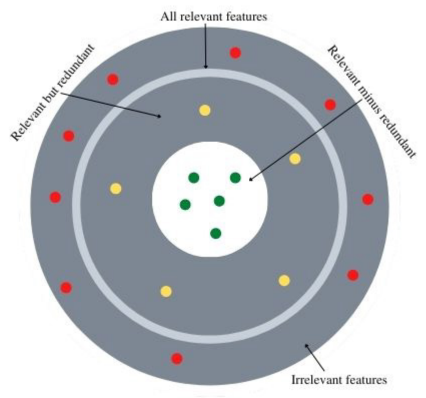


Figure 1. Features distribution in a dataset in terms of relevance and redundancy [10]

This project aims to assist clinicians in identifying ovarian risks early on, reducing ovarian malignancy-related complications and mortality. Feature selection is the process of selecting a subset of features in the dataset that the machine learning model can use. The feature selection process helps determine which features affect the model's performance, improve the model's generalization, and increase the computation speed. **Figure 1** shows the white circle with the filtered features that are relevant and are not redundant. Using such features selected by the mRMR algorithm, the Machine Learning model will predict whether or not the patient should undergo advanced screening for Ovarian Cancer at an early stage.

The Ovarian Cancer data is available at Mendeley Data,[13], an online cloud-based repository that provides datasets for various research projects. The raw dataset with 349 rows and 51 columns has an even distribution of the target column values, with approximately half of the target values being positive. The initial value for the number of features to be selected (K) was 40 for the mRMR algorithm.

Equation 1: $score_i(f) = \frac{relevance(f\,|target)}{redundancy(f\,|\,features\ selected\ until\ i\ -\ 1)}$ The mRMR selection process in simple terms: f is the feature, the numerator being the relevance between the feature f with the target variable and denominator being the redundancy among the features calculated using Pearson correlation.[9, 10]

Equation 2: $score_i(f) = \frac{F(f,target)}{\sum_{s \in features\ selected\ until\ i\text{-}1} |corr(f,s)|/(i-1)}$ The F-statistic between the feature and the target variable is used to calculate the importance of a feature f at the i^{th} iteration (relevance). The average correlation (*corr*) between the feature and all the features that have been selected in earlier iterations is used to calculate the redundancy.[9, 10]

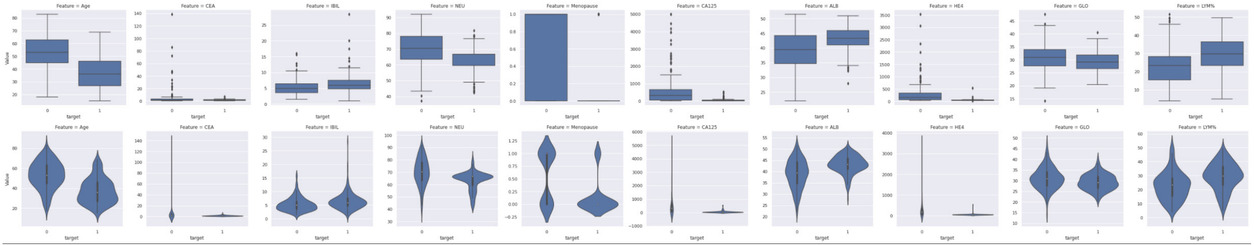
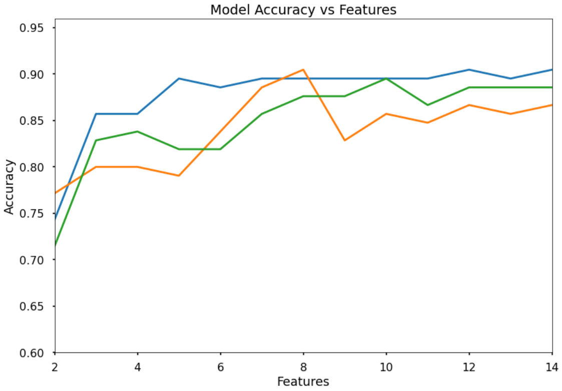


Figure 2. The box plot and violin graph for the 10 optimal features as picked by mRMR feature selection process

Phase II



phase selects a subset of features from the previously selected 40 features by mRMR algorithm. This is This phase selects a subset of features from the previously selected 40 features by the mRMR algorithm. In order to select the minimum optimal panel of biomarkers that can accurately classify input, the three models, namely the RandomForest, MLPClassifier[1, 2, 15], and XGBoost, are used. Various metric comparisons from the three machine learning models resulted in the final set of ten biomarkers. **Figure 3** shows that two out of the three models peak at ten features, and the accuracy plateaus after that for the rest of the features selected in Phase I. The Random Forest model seems to outperform the other two models, although XGBoost is close enough. Random Forest also has the best AUC and score.

The confusion matrix for RandomForest and XGBoost classifiers, overall performance graph, and ROC curve analysis

Figure 3. Accuracy plot with different number of features

for all three classifiers are shown in **Figure 4, Figure 5**, and **Figure 6**.

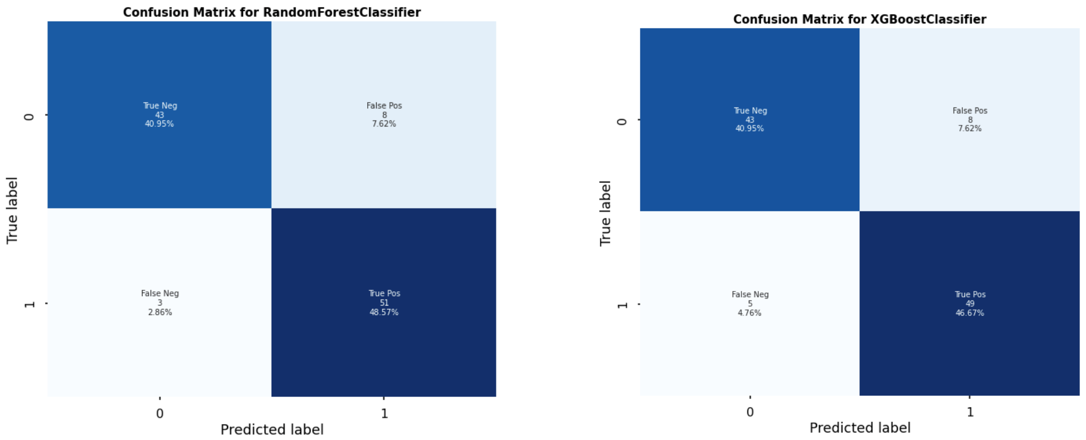


Table 1. Evaluation Metrics for models using different feature sets as selected by mRMR algorithm

Feature Sets *(Defined in Table2)*	**Accuracy**			**AUC**		
	Random Forest	**MLP Classifier**	**XGBoost**	**Random Forest**	**MLP Classifier**	**XGBoost**

Figure 4. Confusion matrix for **RandomForestClassifier** and **XGBoostClassifier** using the 10 optimal features

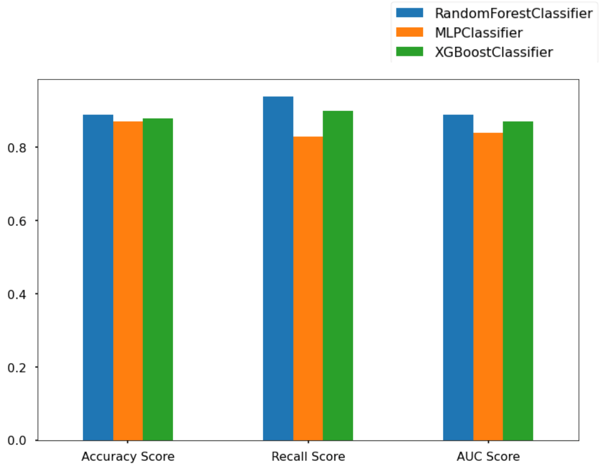


Figure 5. Metrics for **RandomForestClassifier**, **MLPClassifier** and **XGBoostClassifier** using the 10 optimal features

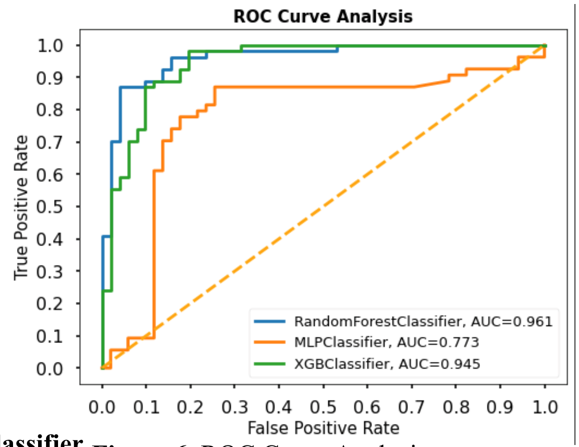


Figure 6. ROC Curve Analysis

1	0.85	0.69	0.81	0.93	0.87	0.92
2	0.86	0.69	0.81	0.95	0.90	0.93
3	0.88	0.72	0.85	0.95	0.89	0.92
4	0.88	0.59	0.87	**0.96**	0.89	0.93
5	**0.90**	0.59	0.89	**0.96**	0.92	0.94
6	0.89	0.79	0.89	**0.96**	0.88	0.94
7	0.89	0.81	0.86	**0.96**	0.91	0.93

Table 2. Features sets

Features Set#	**Combination of features from the top 20 as identified by mRMR**
1	*Age, Carcinoembryonic Antigen, Indirect Bilirubin, Neutrophil Ratio, Menopause, Carbohydrate Antigen 125*
2	*Age, Carcinoembryonic Antigen, Indirect Bilirubin, Neutrophil Ratio, Menopause, Carbohydrate Antigen 125, Albumin*
3	*Age, Carcinoembryonic Antigen, Indirect Bilirubin, Neutrophil Ratio, Menopause, Carbohydrate Antigen 125, Albumin, Human Epididymis Protein 4*
4	*Age, Carcinoembryonic Antigen, Indirect Bilirubin, Neutrophil Ratio, Menopause, Carbohydrate Antigen 125, Albumin, Human Epididymis Protein 4, Globulin ,*
5	*Age, Carcinoembryonic Antigen, Indirect Bilirubin, Neutrophil Ratio, Menopause, Carbohydrate Antigen 125, Albumin, Human Epididymis Protein 4, Globulin , Lymphocyte Ratio*
6	*Age, Carcinoembryonic Antigen, Indirect Bilirubin, Neutrophil Ratio, Menopause, Carbohydrate Antigen 125, Albumin, Human Epididymis Protein 4, Globulin , Lymphocyte Ratio, Aspartate Aminotransferase*
7	*Age, Carcinoembryonic Antigen, Indirect Bilirubin, Neutrophil Ratio, Menopause, Carbohydrate Antigen 125, Albumin, Human Epididymis Protein 4, Globulin , Lymphocyte Ratio, Aspartate Aminotransferase, Platelet Count*

Results

This research project uses the mRMR (Maximum Relevance Minimum Redundancy) feature selection algorithm - an algorithm used to find the minimal-optimal subset of features” [10] - to find a set of ten biomarkers that have the most impact on the target variable (outcome). The Machine Learning models (RandomForest, XGBoost, & MLP Classifier) are then trained and tested on the scrubbed data (with these ten features) to make predictions about a given woman s chances of having Ovarian Cancer (based on their vitals). The results of this project, indicating that the RandomForest Machine Learning model performed the best, are shown in **Table 1**. The recall score or sensitivity depicts how well an ML model can identify true positives (Wikipedia, 2021). Random Forest had a recall score of 0.94, while MLPClassifier had 0.83, and XGBoost had 0.9.

Discussion

The AUC score refers to the area under the ROC curve (also included in the diagrams above) [4, 14]. An AUC score of 1 indicates that the model can easily distinguish between the two classes (in this case, 0 - for not having Ovarian Cancer - or 1 - for having Ovarian Cancer).

An AUC score between 0.5 and 1 indicates an overlap between the two classes, but the model is still able to distinguish between the two relatively well. For example, if the AUC score is 0.7, there will be a slight overlap between the two classes, and there is a 70% chance that the model can distinguish between them. If the AUC score is exactly 0.5, then the model cannot distinguish between classes, as there is only a 50% chance. However, the worst-case scenario is if the model has an AUC score of 0, which indicates that the model is switching classes. As shown in **Figure 5** above, the project results reveal that all models distinguished between classes remarkably well, with Random Forest s AUC score at 0.96, MLP Classifier s AUC score at 0.92, and XGBoost s AUC score at 0.94. **Figure 4** above also includes a confusion matrix - a table that is often used to describe the performance of a classification model" (Data Table, 2014). In a confusion matrix, the top left container includes all the true positive values; top right container includes all the false positive values; bottom left container includes all the false negative values; and the bottom right includes all the true negative values. When comparing the confusion matrices of all of the Machine Learning algorithms, it is discovered that out of 105 predictions, Random Forest made 92 accurate predictions, MLPClassifier made 89 accurate predictions, and XGBoost made 92 accurate predictions. Based on these evaluation metrics, the RandomForest Learning model produces the overall best results for the given dataset.

Conclusion

Among all the Machine Learning models trained and tested, **RandomForest** proved to be the best trained model with the selected features of the dataset with an average **accuracy score** of **90%**, **recall score** of **0.94**, and **AUC score** of **0.96**

The Heroku web application prepared using Jinja2 template, and Flask web server[5] is published and provides an interface to test the biomarker panel in clinical settings. The test results collected can be used further to tune the model and the product's usability. The link to the testing web application is http://ovarify.herokuapp.com.

Project Extension for General Usability

One of the features used by the trained model is HE4 - which is generally unavailable as a part of the basic lab test results. For this research and application to be practically useful in clinical settings, an additional model was created that only uses nine out of the ten best features, leaving out HE4. The accuracy and sensitivity score of the new model was slightly lower when compared to the original model with ten features. To accommodate for clinicians who may not have this biomarker value available in the blood test, the HE4 biomarker field on the app is optional, and when it isn't filled out, the application uses the ML model that was trained using the nine features, including all the selected features except HE4.

Future Improvements

As the model in this project uses a smaller dataset, it can be trained and improved further when used in actual clinical settings with more data alongside a qualified professional.

Acknowledgments

I would like to thank Dr. Madhuri Yalamchilli (MD - Oncologist in Binghamton, New York) for her guidance on this project.

References

[1] "1.17. Neural Network Models (supervised)." *Scikit-learn*, scikit-learn.org/stable/modules/neural_networks_supervised.html. Accessed 17 Jan. 2022.

[2] "12 Types of Neural Networks Activation Functions: How to Choose?" *V7 - AI Data Platform for ML Teams*, 17 Jan. 2022, www.v7labs.com/blog/neural-networks-activation-functions.

[3] "Advantages of Tree-Based Modeling." Summit | Quantitative Consulting and Data Analytics, www.summitllc.us/blog/advantages-of-tree-based-modeling.

[4] "Understanding the AUC-ROC Curve in Machine Learning Classification." Analytics India Magazine, 7 Oct. 2021, analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/#:~:text=ROC%20curve%2C%20also%20known%20as,sensitivity%20of%20the%20classifier%20model.

[5] "Complete Guide on Model Deployment with Flask and Heroku." *Medium*, 1 Jan. 2022, towardsdatascience.com/complete-guide-on-model-deployment-with-flask-and-heroku-98c87554a6b9.

[6] "How to Use StandardScaler and MinMaxScaler Transforms in Python." *Machine Learning Mastery*, 27 Aug. 2020, machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/.

[7] "Machine Learning: What It is and Why It Matters." www.sas.com/en_us/insights/analytics/machine-learning.html.

[8] Malik, Farhad. "What Are Hidden Layers?" *Medium*, 20 May 2019, medium.com/fintechexplained/what-are-hidden-layers-4f54f7328263.

[9] Catà Villà, M. (2014, June). *FEATURE SELECTION METHODS FOR PREDICTING PRECLINICAL STAGE IN ALZHEIMER S DISEASE.* https://imatge.upc.edu/web/sites/default/files/pub/xCata16.pdf

[10] Mazzanti, S. (2022, February 15). *MRMR" explained exactly how you wished someone explained to you.* Medium. https://towardsdatascience.com/mrmr-explained-exactly-how-you-wished-someone-explained-to-you-9cf4ed27458b

[11] Song, H., Yang, E., Kim, J., Park, C., Kyung, M., & Kim, Y. (2018). Best serum biomarker combination for ovarian cancer classification. *BioMedical Engineering OnLine*, *17*(S2). https://doi.org/10.1186/s12938-018-0581-6

[12] “Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform." *ArXiv.org E-Print Archive*, arxiv.org/pdf/1908.05376.pdf.

[13] *Mendeley Data*, data.mendeley.com/.

[14] Narkhede, Sarang. "Understanding AUC - ROC Curve." *Medium*, 15 June 2021, towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[15] “NN - Multi-layer Perceptron Classifier (MLPClassifier)." *Michael Fuchs Python*, 3 Feb. 2021, michael-fuchs-python.netlify.app/2021/02/03/nn-multi-layer-perceptron-classifier-mlpclassifier/.

[16] “Ovarian Cancer - Symptoms and Causes." *Mayo Clinic*, 25 July 2019, www.mayoclinic.org/diseases-conditions/ovarian-cancer/symptoms-causes/syc-20375941.

[17] “Types and Stages." *Ovarian.org*, 18 June 2021, ovarian.org/about-ovarian-cancer/types-and-stages/?utm_term=&utm_campaign=Dynamic+Search+Ads+2021&utm_source=adwords&utm_medium=ppc&hsa_acc=9835623983&hsa_cam=15289195583&hsa_grp=130033718819&hsa_ad=562233134588&hsa_src=g&hsa_tgt=dsa-437115340933&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gclid=CjwKCAiAn5uOBhADEiwA_pZwcFUeMcpbvRsb6T4OfXm5Nl7kkO8ESDtUJZPEc4RS8YdzF-hlvvU_oxoCLE8QAvD_BwE.