Using Artificial Intelligence to Predict Ovarian Cancer

Nour Mlaiki[1]

[1] *Les Elites High School, Sousse, Tunisia*

Abstract— Ovarian cancer is the 5th leading cause of death among women. [2] It is often diagnosed in its advanced stages, making it difficult to cure. This study aims to use artificial intelligence to predict the likelihood of developing ovarian cancer by analyzing certain biomarkers to aid in early diagnosis. The data included 234 patients with 40 diagnosed with ovarian cancer. I have trained the AI model on different classification methods, including Logistic Regression , Forest Tree Classifier, Decision Tree and Neural Networks in which the latter performed best. The values for the highest accuracy and area under the ROC curve (AUC) for segregating between positive and negative cases were both 0.98 .

Keywords— Ovarian Cancer, Biomarkers, Machine Learning, Neural Networks

I. Introduction

Ovarian cancer is very common amongst women all around the world. However, it remains one of the most deadly, with a 60% death rate in 2020.[1]The reason for this dangerously high percentage lies in the tardiness of diagnosis, as the cancer is only detected in advanced stages when it's difficult to treat and cure. In fact, there are many factors that hinder the onset detection of this cancer including but not limited to: the limited number of specialized professionals and that it requires expensive, invasive methods such as biopsies that are often inaccessible in rural, resource-limited areas. My goal is to develop a non-invasive, time and money-saving method to detect the cancer thus facilitating treatment and potentially saving lives. Multiple genetic biomarkers have been linked to ovarian cancer; their presence could indicate a high risk of developing ovarian cancer. I have trained various AI models using different supervised classification methods to predict ovarian cancer using numerical data: the levels of these biomarkers in the bloodstream. These methods include Decision Trees, Random Forest Classifiers, Logistic Regression as well as Neural Network. To assess the AI model's performance, I used the following metrics: Accuracy, Precision, Recall, F1 Score and AUC. I employed multiple metrics because the consequences of making a wrong diagnosis could be deadly. Therefore, I needed to prove future collaborators the reliability of the model and its readiness to be deployed in healthcare facilities.

Contact data: Nour Mlaiki, nmlaiki2004@gmail.com

II. Background

- In Wang et al's paper : "Evaluation of a convolutional neural network for ovarian tumor differentiation based on magnetic resonance imaging" [5] , they discuss how accurate their AI model was at diagnosing ovarian cancer in comparison with junior and senior radiologists' performance. Their highest performing AI model performed better than the radiologists, as well as served as a confirmation tool for them when diagnosing. These results encourage further studies in the implementation of AI in diagnosing ovarian cancer and served as inspiration for this study.
- In Xie et al's paper : "Early lung cancer diagnostic biomarker discovery by machine learning methods "[7], they demonstrate that Neural Networks are efficient at analyzing genetic biomarkers. Therefore, I was encouraged to use this method in my study. Moreover, the researchers explain the importance of False Negatives in medical analysis which inspired me to further look into it in the study.
- Sherafatian discusses in his paper : "Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping " [6] the use of Random Forest classifiers to analyze genetic biomarkers. I wanted to implement this method in my study to assess the AI model's performance with varied techniques.

III. Dataset

I retrieved data published by Mi et al's paper "Using Machine Learning to Predict Ovarian Cancer"[4] . This data included 234 patients and 49 biomarkers. An important metric to con-

sider in any supervised classification problem is the split in positive (patients with ovarian cancer) and negative examples in the dataset. The perfect split in examples is 50/50. However, this split doesn't correspond to real life examples seen in hospitals. This data is close to ideal as it has 89 positive examples and 146 negative examples as visualized by the graph below.

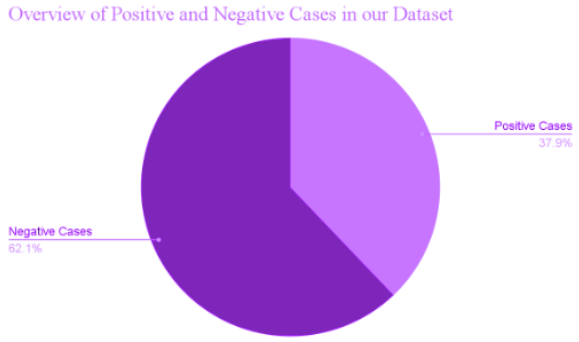


Fig. 1: Overview of Positive and Negative cases in the Dataset

An important feature in the dataset is age. As indicated by the boxplot, the data includes a large range of ages, in which the youngest patient is 15, the oldest patient is 83 and the median is 47. This feature is highly advantageous as it trains the AI model on a broad range of ages that resembles what a doctor would be exposed to on field. Hence reducing risk of bias and making the AI model reliable to be deployed in hospitals.

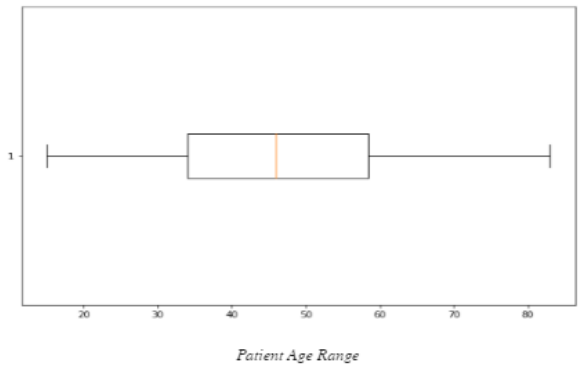


Fig. 2: Patient Age Range

The AI analyzed 49 genetic biomarkers, including but not limited to proteins, platelet volume and distribution, hemoglobin levels and red blood cells count and distribution. Some notable biomarkers are displayed in the table below.

TABLE 1: NOTABLE GENETIC BIOMARKERS AND THEIR ROLES

Biomarker	Role
CA72-4	Protein found on the surface of tumors
CA125	Protein found to be high in early ovarian cancer stages
CA19-9	Tumor marker
GGT	Has been associated with cancer risk and prognosis

To train the AI model, the data was split into a training set and a testing set which it would be evaluated upon. The data source provided testing data, which I have merged with the training data to create a new training dataset composed of 349 data. I wanted to train the AI model on the largest amount of data available in order to get the best results. After merging the data, we resplit it using the train_split_function, in which 15% of the data was the test data set and 85% was the training data. I chose a large training dataset in order to expose the AI model to as much data as possible in order for it to learn and make meaningful patterns.

IV. METHODOLOGY

a. Data Preprocessing

To train the AI model, I used different supervised classification methods. I used these methods specifically because they return a "yes or no" answer, which is the best fit for our problem. Considering that the features consisted of numerical data, with each feature having a different range than the other,it imposed a risk of bias. In other words, the AI model could overlook a feature that had a small value , less than 1 per example, because it was exposed to values that were greater than 100, thus affecting its performance. Therefore, I needed to normalize the data, meaning to transform the numerical data into values ranging between 0 and 1. To do so, I ran the following equation:

$$\frac{Value - Min}{Value - Max} = NewValue$$

With min being the minimum value in the dataset, and max being the maximum value in the dataset.

b. Logistic Regression

Logistic Regression is a statistical analysis method used to predict an outcome, in this case, whether a patient has ovarian cancer, based on multiple criteria including age and biomarkers. To improve the model's performance, we can choose how many times the AI model runs the model by tuning a parameter known as "Maximum Iteration", the performance of the AI model is proportional to this parameter. I will touch on this in the results section.

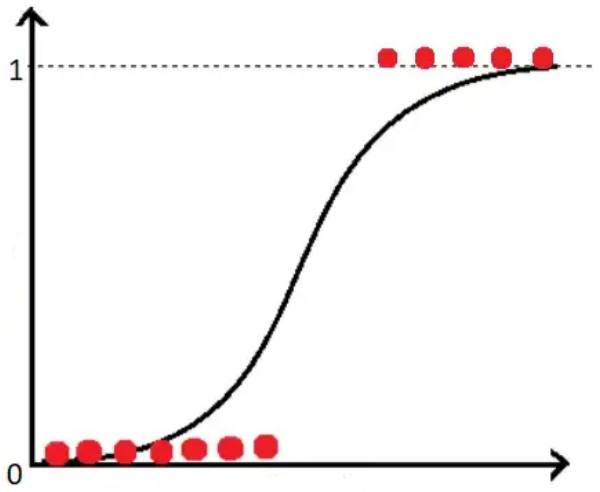


Fig. 3: Visualization of Logistic Regression [3]

c. Decision Tree Model

Decision trees consist of nodes and branches. Nodes are features that are recursively evaluated by the model to make a data follow a particular path called a branch. This path ends up at a leaf node, which is the final node that outputs the prediction, either 0 or 1. When evaluating a feature, decision tree models use different metrics. In this study, I looked at two metrics : Gini and Entropy.

Gini measures the likelihood of incorrectly identifying a randomly chosen element. It is modeled by the following equation:

$$\text{Gini} = \sum j^2$$

When evaluating features using the Decision Tree method, the AI model relies on features with a low Gini index. Entropy can be defined as a measure of the certainty of a split. It helps the AI model determine the best feature to split as it runs the model. It is modeled by the following equation ; where x is a variable that can take different values and p(x) is its probability.

$$H(x) = -\sum_{i=1}^{N} p(x_i)\, log_2 p(x_i)$$

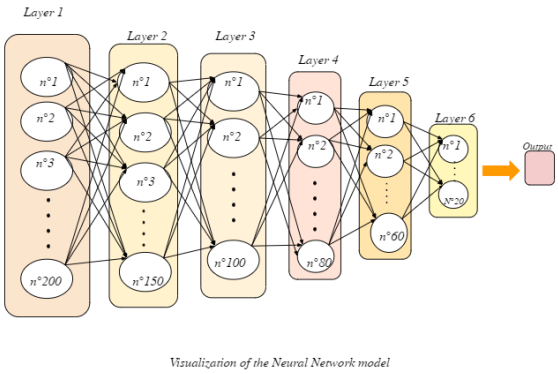
Other metrics that stayed fixed for both models are "random state", "max depth" and "min samples leaf". Random state controls the randomness involved in splitting the dataset into branches, which I have set to 100. Max depth is defined as the number of splits a tree would make before reaching a final prediction. In other words, it is the distance between the first node to the final leaf which in this study was fixed at 5. Min leaf samples are the minimum number of samples required to be at a leaf node, set to 10.

d. Random Forest

Random Forest Classifiers consist of an ensemble of decision trees that operate as a committee. The output provided by the Random forest model is the most common output given by the decision trees. Similarly to the Decision Tree Model, I have trained the AI model on Gini based models, and Entropy based models. The AI model also took into consideration other metrics that remained unchanged for both models, including "n jobs", "random state", and " n estimators". "N jobs" is defined as the number of decision trees that run parallel to each other , which I set to 10. "n estimators" indicates the number of decision trees, which was set to 10. Finally, "random state" was set to 5.

e. Neural Networks : Multi-layer Perceptron classifier

Neural Networks are a machine learning algorithm that consist of multiple layers, each containing nodes where the outcome of a prediction is fed to a new input recurrently. In this research, I used 6 layers, containing 250, 150, 100, 80, 60 and 20 nodes respectively. Other metrics used by the AI model to evaluate features include "learning rate", "solver" and maximum iterations. "Learning rate" manages how quickly the model adapts to the problem, which I've set to 0.001. "Solver" and "Max iterations" varied in order to study how well the AI model responded to each metric. The findings will be detailed in the results section.



Visualization of the Neural Network model

Fig. 4: Visualization of a Neural Network model containing 6 layers and their corresponding neurons.

V. Results and Discussion

a. Results

To evaluate the AI model's performance, I classified its outputs into 4 main categories: True Positives, False Positives, True Negatives and False Negatives. True Positives indicate that the model correctly predicted positive cases, whereas False Positives are incorrectly identified normal cases. True Negatives are correctly predicted negative cases, whereas False Negatives are incorrectly identified abnormal cases.

When training the AI model using logistic regression, I visualized how the AI's accuracy varied when changing the "max iteration" parameter. I tested it from 1 to 200. Accord-

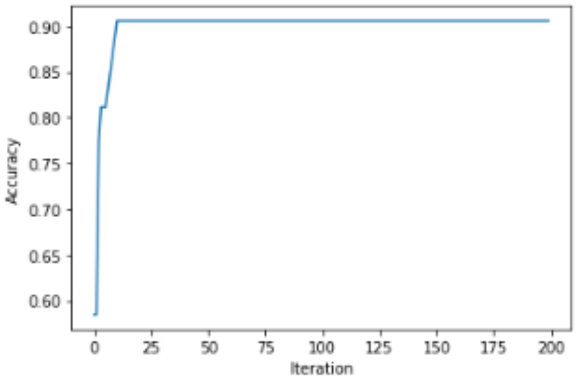


Fig. 5: The Correlation between maximum iterations and Accuracy

ing to the graph, the AI model's accuracy improved to reach a maximum of 90

Similarly to how I trained the AI model using logistic regression, I visualized the AI model's accuracy whilst varying the "max iteration" parameter for the Neural Networks method, testing it from 1 to 200. According to the graph, the AI model's accuracy improved, reaching a maximum of 98% when the parameter was set to 176 iterations.

Moreover, I trained the Neural Network based AI model using 3 different "solvers": "Adam, "Sgd" and "lbfgs". For

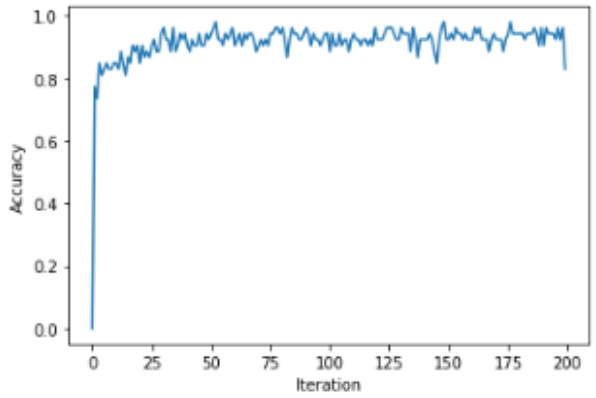


Fig. 6: The Correlation between maximum iterations and Accuracy

"Adam" and "Sgd", the "max_iter" parameter functions differently. In fact, instead of running the program for the indicated amount, it uses an epochs approach. In other words, it is the number of times that the model looks at all of the data. The Adam model performed best, followed by the lbfgs at 96% accuracy, and Sgd at 84% accuracy.

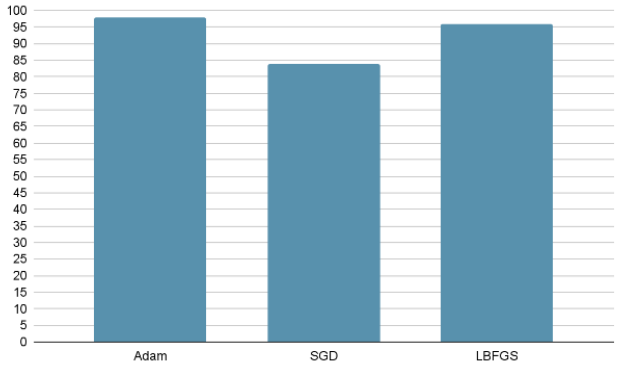


Fig. 7: Evaluation of the model's accuracy when varying solvers

To evaluate the AI model's performance, I tested it on 4 metrics: Accuracy, Precision, Recall and F1 score.

$$\text{Accuracy} = \frac{\text{TrueNegatives} + \text{TruePositives}}{\text{Total}}$$

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The table above indicates the performance of the Logistic Regression, Gini based Decision Tree, Entropy based Decision Tree, Gini based Random Forest, Entropy based Random Forest and Neural Networks AI models respectively on the 4 indicated metrics. The Neural Networks based AI model performed best. To further assess its performance, I visualized its predictions through a confusion matrix and a ROC Curve. The Area Under Curve (AUC) was equal to 0.98.

TABLE 2: EVALUATION OF THE AI MODELS' PERFORMANCES

"	Log Reg	GDT	EDT	GRF	ERF	NNs
Accuracy	90%	90%	92%	86%	84%	98%
Precision	95%	95%	95%	91%	100%	100%
Recall	84%	84%	88%	80%	69%	96%
F1 Score	89%	89%	91%	85%	81%	98%

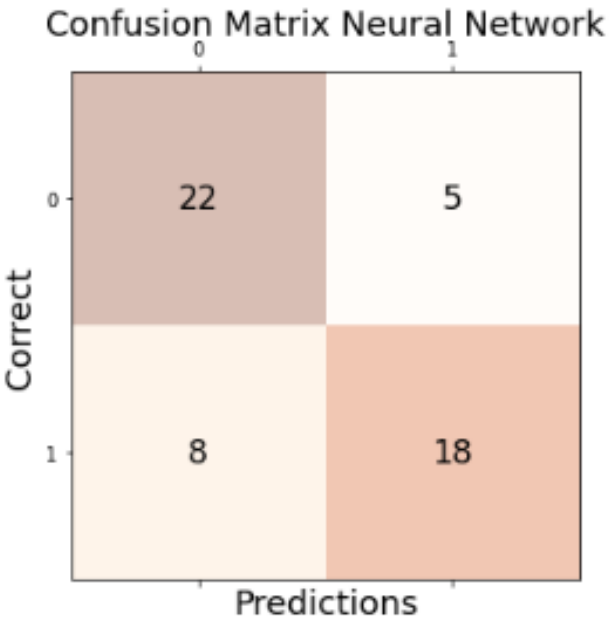


Fig. 8: Neural Network Confusion Matrix

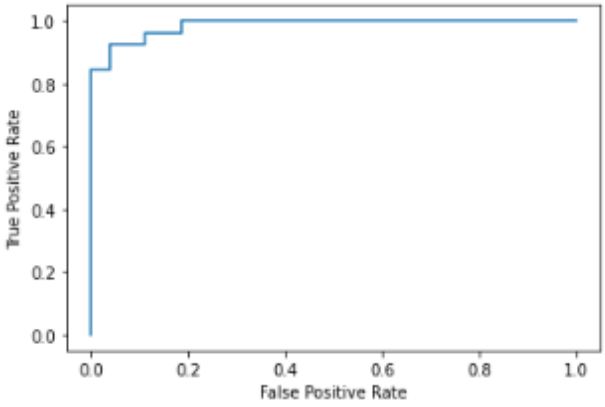


Fig. 9: Neural Network based model ROC Curve

b. Discussion

The implementation of AI in the healthcare field comes with risks, most importantly, the risk of predicting a False Negative. In other words, it could be life-threatening to entirely depend on an AI model which incorrectly diagnosed a patient to be unlikely to have ovarian cancer. This is why I tested the model on recall, which indicates the percentage of positive cases correctly identified. The higher the recall, the better the model is fit to perform in hospitals. The NN model had the highest recall rate of 96%, whilst the Entropy based Random Forest model had the lowest rate of 69%. Predicting False Positives are also risky as they could lead the patient to spend more on invasive treatments yet they remain far less dangerous than False Negatives.

c. Future Works

In order to implement this model in healthcare facilities, I need to test the AI on a larger dataset to improve accuracy and recall scores. Furthermore, I need to collect data from different countries to eliminate the possibilities of having geographical factors interfere with the AI's accuracy and to be able to generalize the model for women all around the world. This could be done by collaborating with healthcare providers, cancer research laboratories and hospitals.

VI. Conclusion

In this study it was proven that AI could be trained and deployed in the healthcare field, in particular in detecting ovarian cancer. I experimented with different classification models (Logistic Regression, Neural Networks, Decision Trees and Random Forest Classifiers) to determine the best method fit to train the AI model. The Neural Network based AI model performed best, with a 98% accuracy and 96% recall. This demonstrates that AI is highly efficient at making reliable predictions thus aiding doctors in detecting ovarian cancer in its early stages.

VII. Acknowledgments

I would like to acknowledge and thank my mentor Oscar O'Rahilly for helping me throughout this study with data retrieval, training the AI, writing the paper and for his encouraging words throughout this journey.

VIII. GitHub Code

The code is public and can be found in the following github link. https://github.com/NourMlaiki/AI_Ovarian

References

[1] Centers for Disease Control and Prevention.

[2] World Cancer Research Fund International. *Ovarian Cancer Statistcs*. Tomado de `https://www.wcrf.org/cancer-trends/ovarian-cancer-statistics/s`. 2020.

[3] Medium. *The math behind Logistic Regression*. Tomado de ://medium.com/analytics-vidhya/the-math-behind-logistic-regression-c2f04ca27bca. 2020.

[4] Jingting; Znati Ty; Fan Zhenjiang; Li-Jundong; Xu Bin; Chen Lujun; Zheng Xiao; Lu Mingyang Mi Qi; Jiang. "Data for: USING MACHINE LEARNING TO PREDICT OVARIAN CANCER." In: *Mendeley Data* (2020).

[5] Iris K. Lee Rong Hu-Subhanik Purkayastha Ian Pan Thomas Yi Thi My Linh Tran-Shaolei Lu Tao Liu Ken Chang Raymond Y. Huang Paul J.Zhang-Zishu Zhang Enhua Xiao Jing Wu Harrison X. Bai Robin Wang Yeyu Cai. "Evaluation of a convolutional neural network for ovarian tumor differentiation based on magnetic resonance imaging". In: *European Society of Radiology* (2020).

[6] Masih Sherafatian. "Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping." In: *Department of Molecular Genetics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran* (2018).

[7] Run-Ze Li Yu-Wei Wang Xin Qian Chang Chan Zhi-Fang Yu Xing-Xing Fan Hu-Dan Pan Chun Xie Qi-Biao Wu Pei-Yu Yan Liang Liu Yi-Jun Tang Xiao-Jun Yao Mei-Fang Wang Elaine Lai-Han Leung Ying Xie Wei-Yu Menga. "Early lung cancer diagnostic biomarker discovery by machine learning methods". In: *ELSEVIER* (2021).