A Comparative Study to Predict Ovarian Cancer

Binadie Wickramasinghe[1] and Gloria Regisford[2]

[1]Bridgeland High School, Cypress, TX ; [2]Prairie View A&M University, Prairie View, TX

Ovarian cancer is considered the fifth most common cancer type among females in the United States. Furthermore, ovarian cancer accounts for 25% of all gynecologic cancers, and usually, this cancer is diagnosed at a late stage. A patient can live at least five years longer if ovarian cancer is diagnosed early. Therefore, the early diagnosis of ovarian cancer is essential. This study aims to classify ovarian cancer using biomarkers such as ovarian cancer antigen (CA125), tumor necrosis factor alpha (TNF-α), interleukin 6 (IL-6), human epididymis protein 4 (HE4), and anti-TP53 antibodies. Rising or persistent CA125 blood levels provide a highly specific biomarker for epithelial ovarian cancer, but not an optimally sensitive biomarker. Addition of HE4, CA 72.4, anti-TP53 autoantibodies and other biomarkers can increase sensitivity for detecting early stage or recurrent disease. It also uses three data-classifying models called Decision Tree (DT), kth Nearest Neighbor (kNN), and Logistic Regression (LR) to compare their performances. We computed various model performances, such as accuracies, precision, and recall values. Based on the findings, the LR model shows the highest performance compared to the other two models. Furthermore, it records 87% accuracy and 99% recall in classifying ovarian cancer.

Introduction

Ovarian cancer has become a global health issue. Out of a total of 1,806,590 all novel cancer cases and 606,520 cancer-related deaths recorded in 2020 in the US, ovarian cancer deaths appear as the main cause of deaths related to female reproductive tract. (Matsas et. al., 2023). Furthermore, ovarian cancer is the fifth most frequent cause of death in women and is the leading cause of death out of all gynecological cancers (PDQ, 2023).

The high mortality rate of ovarian cancer is due to the fact that this cancer is typically diagnosed at a late stage (Matulonis et al., 2016). Furthermore, the already existing screening tests do not provide a considerable predicting value in diagnosing the disease. The trans-vaginal ultrasound and cancer antigen 125 (CA125) are being used as screening tests for ovarian cancer. However, they are not good enough in early detection of the disease, and consequently, do not lead to significant reductions in morbidity and mortality (Dochez et. al., 2019). The standard treatment of care includes surgery and chemotherapy. However, there is often a high rate of recurrence followed by the standard treatment (Giampaolino et. al., 2020). Studies have reported that if relapses occur, ovarian cancer becomes less curable (Yang et. al., 2017). Due to all these factors, early detection is vital in enhancing the efficacy of the treatment and thus extending the lifespan of these patients.

Machine learning is a subset of artificial intelligence that attempts to identify the relationships among variables instead of forcing instructions on the data. With the availability of large number of datasets, the application of machine learning algorithms has become widely popular in various domains. Computer vision, speech recognition, natural language processing are some of such domains of the applications of machine learning (Ghassemi et al., 2020). Due to machine learning's capability to extract hidden features and relationships using the dataset, the healthcare sector benefited from the applications of machine learning. Some of the areas in the healthcare domain include medical diagnosis (Mahoto et al., 2023, Kononenko, 2001), disease prediction (Uddin et al, 2019), clinical research (Doan et. al., 2023), drug development (Réda et al., 2020), electronic health records (Weissler et al., 2021), healthcare improvement (Habehh and Gohel, 2021), and personalized healthcare (Nguyen et al., 2021).

Cancer research is imperative compared to other studies due to the impact of cancer on society. Machine learning is playing a major role related to cancer research. Unfortunately, literature indicates an asymmetry of the distribution of the studies using machine learning and different types of cancers. Though ovarian cancer has a significant impact on women, the number of studies to predict ovarian cancer research is not noteworthy. Therefore, this study aims to develop three distinct ovarian cancer prediction models using machine learning and compares the performance of each model to identify the best, most effective model.

Literature Findings

A glance at the literature reveals that Machine Learning techniques have been used in studies related to various types of cancer types. Based on the findings, 36.1% of the studies are related to breast cancer, 11.1% are lung cancer, 8.3% are ovarian cancer, 5.6% are colorectal, and 5.6% are related to multiple cancer types. Figure 1 summarizes the percentages of the above types of studies. Furthermore, according to table 1, the literature indicates the use of various types of data ranging from actual clinical data, biopsies, RNA, and miRNAs.

Most of the prior research activities are focused on cancer classification, cancer prediction, early diagnosis, or risk assessments. In addition, researchers have used a vast variety of approaches to conduct their studies. Figure 2 summarizes the types of ML techniques they have used in the literature. According to figure 2, out of the above prior studies, 22.1% have used Support Vector Machines (SVM), 11.5% have used Decision Trees (DT), 10.6% have used Random Forest, (RF), and 7.1% have used Linear Regression (LR).
According to figure 3, it is shown that prior research has used various sizes of samples and features sets. Some research studies have used large quantities of data points (instances) over 10,000, while the majority have used less than 5000 data points. The average number of data points the prior research have used is approximately 3842, while the average number of features is 541.

Data Set:

This dataset was collected from Kaggle online data repository (Mi et al., 2020) representing 335 instances and 49 features. The class variable in this dataset is TYPE, representing whether the subject does not have ovarian cancer (0) or has the cancer (1). In the dataset, 51% of the subjects had ovarian cancer while 49% did not have cancer, therefore the dataset was very closely balanced.

Data Preparation:

There were some missing values in the data, and they were replaced by the mean values due to the distribution of the data are approximately symmetric. There were two instances with the AFP value recorded as >1200 without the numerical value, therefore those values were removed from the analysis. Two instances from CA 125 values were removed as their values were recorded as >5000. Finally, four instances were recorded

Table 1: Machine learning and prior cancer research

Reference	**Year**	**Type of Cancer**	**Aim of the study**	**Type of data used**
Achilonu et al. (2021)	2021	Colorectal	Predict recurrence of cancer	Clinical and molecular
Agossou et al. (2021)	2021	Breast	Classify tumors	miRNAs
Akazawa et al. (2020)	2020	Ovarian	Prediction	RNA
Al Mudawi et al. (2022)	2022	Cervical	Early prediction of cancer	-
Alabi et al. (2020)	2020	Tongue	Predict survival	Cancer patients
Alafeef et al. (2020)	2020	Breast	Classify various cancer cell types	Tumor data
Alkhathlan& Saudagar, (2022).	2021	Breast	Predict & classify	Microarray data
Anil Kumar et al. (2022)	2022	Lung	Classification based on their symptoms	Microscopic biopsy
Braz et al. (2022)	2022	Oral cavity	Identification of cancer	Cancer patients
Chiu et al. (2021)	2021	-	Cancer dependency	
Ding et al. (2019)	2019	Pan-cancer	Identification of potential biomarkers	Patients
Doppalapudi et al. (2021)	2021	Lung	Survival prediction	Hyperspectral imaging
Famitha et al. (2022)	2022	Multiple	Prediction	Cancer patients
Fatima et al. (2021)	2021	Breast	Classification	Cancer Cells
Finkelstein et al. (2022)	2022	Prostrate	Early prediction of cancer	microRNAs
Kouznetsova et al. (2021)	2021	Oral	Recognition of oral cancer vs periodontitis	Patients
Li et al. (2022)	2022	Gastric	Early diagnosis	Cancer Registry
Li, X., Dai, A., Tran, R., & Wang, J. (2023).	2023	Breast and ovarian	-	Gene Expression
Lu et al. (2020)	2020	Ovarian	-	FNA biopsies
Lynch et al. (2018)	2018	Lung	Cancer prediction	
Maray et al. (2022)	2022	Breast	Cancer prediction	Clinical data
Masud et al. (2021)	2021	Lung and Colorectal	Classification	Wisconsin (UCI-breast cancer)
Mourad et al. (2020)	2020	Thyroid	Classify patients	Breast biopsies
Nissim et al. (2021)	2021	Any	Detection and classification of untreated cancer cells	Cultured breast cancer cell lines
Poirion et al. (2021)	2021	Liver and breast	Prediction	Patient data
Rasool et al. (2022)	2022	Breast	Cancer diagnostic	Patient data
Redjdal et al. (2022)	2022	Breast	-	Hospital records
Sharma and Mehra (2020)	2020	Breast	Classification	Cancer tissue images
Taghizadeh et al. (2022)	2022	Breast	Diagnose breast cancer	Optical path delay profile of the cell
Ting et al. (2020)	2020	Colorectal	Risk factor identification	Online
Toprak (2018)	2018	Breast	Cancer classification	Patient data
Urbanos et al. (2021)	2021	Brain	Prediction	Biopsy data
Zhang et al. (2022)	2022	Breast	Cancer classification	Online Lung Cancer
Zhao et al. (2022)	2022	-	Cancer prediction	Patient data
Zheng, S., Wu, Y., Donnelly, E. D., & Strauss, J. B. (2023).	2023	Endometrial	Risk assessment	Genomics data

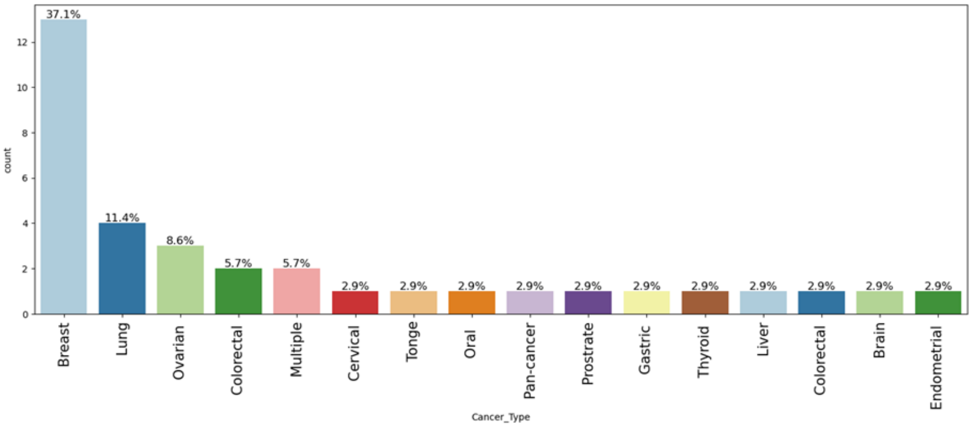


Figure 1: Prior research work in cancer research domain

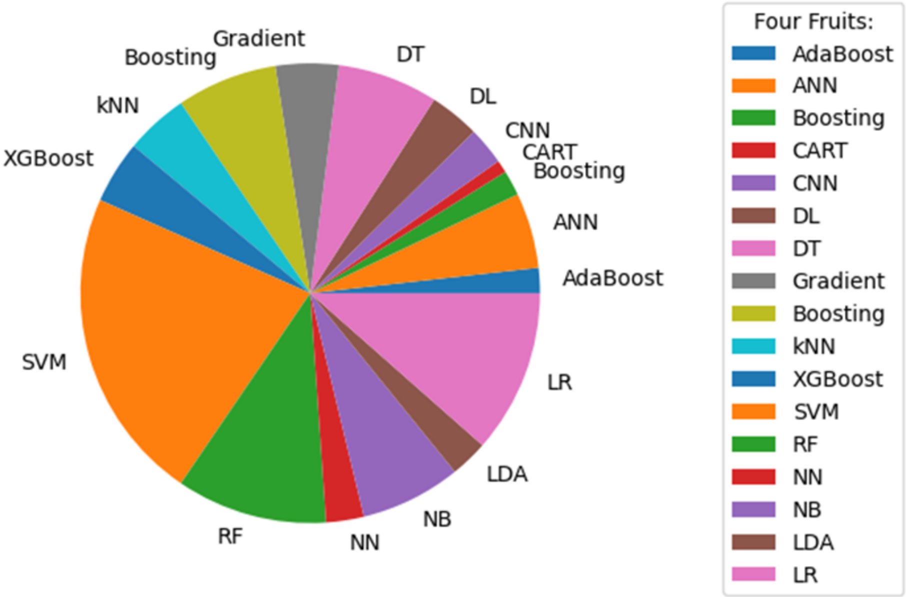


Figure 2: Machine learning models used in prior cancer research

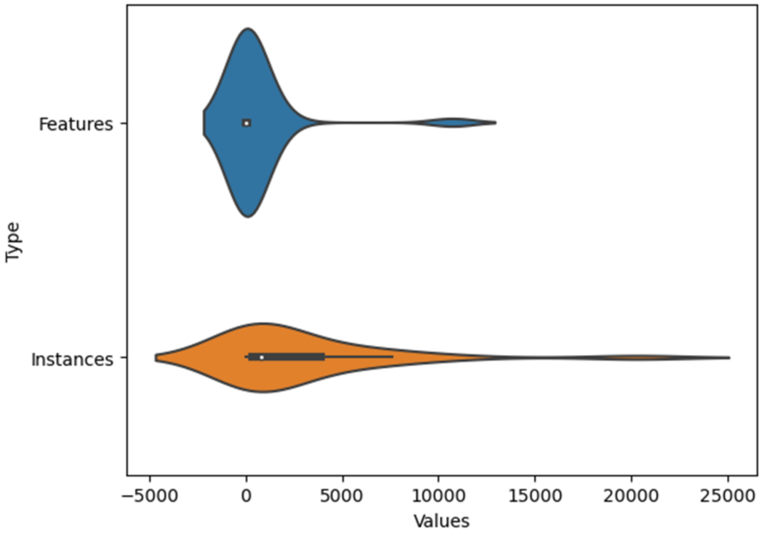


Figure 3: Number of features vs sample size in prior research

as CA19-9 >10000 and six were as <0.6, therefore they were removed from the analysis. Hence, the last number of instances for the rest of analysis was 335.

Feature Extraction

As there are 47 features and only 335 instances, a feature selection technique was applied to select a sample of features from the 47 to analyze the final data. The Chi-square test was used for selecting the most suitable features for this study based on the Chi-square scores.

$$\chi^2 = \frac{(\text{Observed frequency}-\text{Expected frequency})^2}{\text{Expected frequency}}$$

where observed frequency is the number of observations of the selected class and the expected number of observations of the selected class if there is no relationship between the selected feature and the target variable of whether the patient has a ovarian cancer or not.

We selected only features with significant values (i.e., $p<0.05$) which indicate that the selected feature significantly impacts the class variable. Based on the above criteria, we selected the Carbohydrate Antigen 125 (CA125) , Human Epididymis Protein 4 (HE4), Platelet Count (PLT), Carbohydrate Antigen 19-9 (CA19-9), Age, Alpha-Fetoprotein (AFP), Alkaline Phosphatase (ALP), Carcinoembryonic Antigen (CEA), Carbohydrate Antigen 72-4 (CA72-4), Lymphocyte Ratio (LYM%), Neutrophil Ratio (NEU), Aspartate Aminotransferase (AST), Menopause, Albumin (ALB), Gama Glutamyl transferase (GGT), Total Bilirubin (TBIL), Hemoglobin (HGB), Indirect Bilirubin (IBIL), Urie Acid (UA), Lymphocyte Count (LYM#), and Platelet Distribution Width (PDW).for the final analysis. Descriptive statistics of the selected features can be seen below.

Methods

Methodology
In this study, we used three machine learning techniques namely, decision tree, kth nearest neighbors, and logistic regression. A short introduction to each of these techniques is provided below.

Decision Tree (DT)

Decision Tree is a non-parametric and supervised machine learning technique. DT can be used both in classification tasks and regression applications (CART). DT possesses a hierarchical tree structure comprising of the root node, internal nodes, and the leaf nodes. Each level of nodes is connected by branches. Applications of DT related to cancer prediction can be seen in various prior research work. Venkatesan and Velmurugan (2015) used Classification and Regression Trees (CART) to classify breast cancer. In another study, Sathiyanarayanan et al (2019) also used DT approach to detect cancer patients. Naik and Patel (2014) used DT to detect brain tumors and classify brain cancer. These authors compared their performances with the Naïve Bayes approach and find the DT approach outperforms the other to classify brain cancer. As Lewis (2000) shows, CART has advantages over other classification techniques. As CART does not make parametric assumptions, it can handle data with even skewed distributions. Another advantage is the ability to handle classification in the presence of missing values. In this study, DT considers all the subject features and selects the most appropriate variable to select as the root node that splits the rest of the attributes in an optimal way. Recursively, the next best attribute is selected from the rest of attributes from the dataset to form the decision tree.

kth Nearest Neighbors (kNN)

K Nearest Neighbor (kNN) is considered as one of the simplest, but effective classification algorithms. kNN attempts to assign the related class (whether the subject has a cancer or doesn't) of a given subject (point) by calculating the proximity from the novel subject to each of the subjects in the dataset. The decision of being a cancer subject or not is decided based on the majority of the subjects' class out of *k* subjects. In this distance calculation, kNN uses different distance metrics including Euclidian, Manhattan, Mahalanobis, and Chebyshev distances, though the Euclidian being the most frequently used one. Among the many advantages of kNN are simplicity, versatility, and higher accuracy, while some of the disadvantages include larger prediction time for larger dataset and the sensitivity of the prediction accuracy for the scale of data. When using kNN, the value of k plays an important role. When selecting the appropriate k, identification of both training error and the validation of kNN model is important (Moldagulova and Sulaiman, 2017). Application of kNN in cancer research can be found in the literature related to various canver types including Lung cancer (Anil Kumar, et al., 2022), Breast (Fatima, et al., 2020; Taghizadeh et. al., 2022), Oral cavity (Braz et al., 2022), and Cervical (Al Mudawi at. al., 2022).

Logistic Regression (LR)

Logistic regression estimates the probability of a subject having cancer based on the given features of the subject. In this study, the dependent variable (*Y*) is of binary nature (0-the subject does not have cancer, 1- the subject has cancer). With logistic regression a logit transformation is applied on the odds, that is the ration of the probability of the patient having cancer (*p*) and the probability that the subject does not have cancer (*1-p*). This is commonly known as the log odd and is presented by the following equation.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

, here *X,* the predictor represents is the set of factors related to each subject.

Here, Y is the outcome that the subject has cancer (Y=1), and the subject does not have cancer (Y=0). $\beta 1$ represents the regression coefficient, the change in the logarithm of the odd ratio of the event with a 1-unit change in the predictor X.

Checking the accuracy of the models

In this study two criteria were used to measure the effectiveness of each of the above models. These two criteria are based on the confusion matrix shown in figure 4. The notations are defined as follows:

TP = True Positive,

i.e., the model predicts that the subject has cancer when the subject actually has cancer

FN = False Negative,

i.e., the model predicts that the subject does not have cancer when the subject actually has cancer

FP = False Positive,

i.e., the model predicts that the subject has cancer when the subject actually does not have cancer

TN = True Negative,

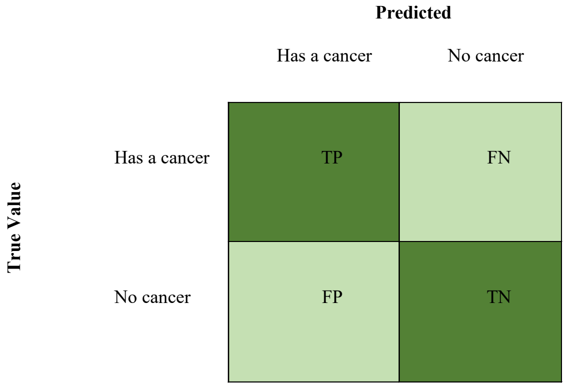
i.e., the model predicts that the subject does not have cancer when the subject actually does not have cancer

Then, the recall and accuracy are .lo09-[-[[-- based as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FP}$$

Both accuracy and recall values take values from 0 (or 0%) to 1 (100%). Higher values of the above indicate the goodness of the model.
In this study, we trained each of the machine learning models using randomly selected subsets of data (training set) and the model is tested from the remainder subset (testing set) of the dataset. Furthermore, we varied the training dataset and measured both accuracy and recall values for each case.



		Predicted	
		Has a cancer	No cancer
True Value	Has a cancer	TP	FN
	No cancer	FP	TN

Figure 4: Confusion Matrix

Results and Discussion

Table 2 shows the descriptive statistics of the selected features of subjects we selected for this study. Based on this data, Age (-0.52) and Menopause (-0.47) indicate the lowest negative correlation with ovarian cancer, while ALB (0.35) and LYM% (0.31) show the highest positive correlation. Figure 5 illustrates the distribution of ALB between the cancer subjects (1) and non-cancer (0) subjects. According to Figure 5, though both distributions of ALB for those with cancer and non-cancer are left skewed, the mean ALB of those who do not have ovarian cancer is shifted to the left compared to the other distribution.

These selected models were trained using a randomly selected subset of the data. The accuracy of the models was measured using the remainder subset of the dataset. To identify the ideal training and testing data ratio, we varied this ratio and measured the performance for each case. Table 3 shows the performances of each model with the training and testing sets of data with proportions of 20%:80%, 30%:70%, and 40%:60% training to testing datasets.

According to the obtained observations, the performances of each model with the training sets are better than those of the respective testing sets. We want to select the model that performs well with the testing dataset. Also, we do not want our model to perform well with the training data and

Table 2: Descriptive Statistics of the selected features

Features	Meaning	Mean	Stdev	Min	25%	50%	75%	Max
CA125	Carbohydrate antigen 125	314.76	698.07	4.48	20.51	49.14	279.30	4468.00
HE4	human epididymis protein 4	183.88	374.29	16.71	42.39	54.46	164.55	3537.60
PLT	platelet count	254.04	94.90	74.00	200.50	235.00	291.00	868.00
CA19-9	Carbohydrate antigen 19-9	35.65	65.84	0.60	8.43	15.11	35.65	566.10
AG	Age	45.04	15.21	15.00	32.50	45.00	57.00	83.00
ALP	Alanine aminotransferase	77.04	44.48	26.00	60.00	71.00	85.50	763.00
AFP	Alkaline phosphatase	4.48	27.72	0.61	1.67	2.41	3.69	508.00
CEA	Carcinoembryonic antigen	2.71	8.85	0.20	0.84	1.38	2.27	138.80
CA72-4	Carbohydrate antigen 72-4	8.46	11.58	0.20	8.46	8.46	8.46	131.60
LYM%	lymphocyte ratio	26.23	10.25	3.90	19.15	26.60	33.00	51.60
NEU	neutrophil ratio	66.27	9.94	37.20	61.75	66.27	70.95	92.00
AST	Aspartate aminotransferase	19.06	8.09	7.00	14.00	18.00	22.00	78.00
MNP	Menopause	0.34	0.48	0.00	0.00	0.00	1.00	1.00
ALB	albumin	41.19	5.55	22.00	38.50	42.00	45.15	51.50
GGT	Gama glutamyl transferase	21.34	18.15	4.00	12.00	16.00	23.00	176.00
TBIL	total bilirubin	9.10	4.15	2.50	6.30	8.60	10.70	38.30
HGB	hemoglobin	125.21	15.71	61.80	118.00	127.00	135.00	189.00
IBIL	Indirect bilirubin	5.99	2.96	1.00	4.00	5.50	7.25	28.40
UA	uric acid	243.05	68.83	96.00	199.70	234.10	275.95	632.00
LYM#	Lymphocyte count	1.56	0.56	0.35	1.19	1.51	1.87	3.49
PDW	Platelet distribution width	14.36	3.00	8.80	11.90	13.80	16.80	22.80

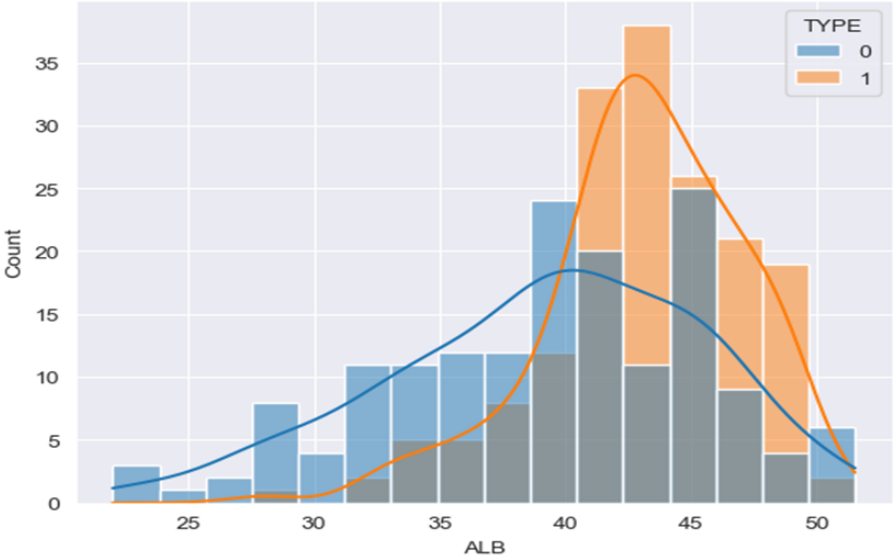


Figure 5: Distribution of ALB between cancer and non-cancer subjects

Table 3: Training data, testing data, accuracy, and recall values

Model	Training: Testing	Training Accuracy	Testing Accuracy	Training Recall	Testing Recall
Logistic	20%:80%	91%	87%	96%	97%
	30%:70%	93%	86%	96%	98%
	40%:60%	94%	87%	97%	99%
Decision Tree	20%:80%	100%	78%	100%	79%
	30%:70%	100%	77%	100%	77%
	40%:60%	100%	83%	100%	78%
kNN	20%:80%	87%	78%	93%	91%
	30%:70%	86%	77%	93%	96%
	40%:60%	87%	78%	93%	96%

poorly with the testing data. Based on Figure 6, the logistic regression model showcases 91% accuracy when predicting ovarian cancer, with 80% of the data selected randomly from the dataset as the training data. When this model predicts ovarian cancer patients with the remaining 20% of testing data, the accuracy drops slightly to 87%. With the same training and testing datasets, logistic regression indicates a recall value of 96% for the training and 97% for the testing data. When 70% of data is randomly selected for the training data and 30% for the testing data, the logistic regression model shows 93% and 86% accuracy with training and testing data. Similarly, for the 60% to 40% training to testing data, this model indicates 94% to 87% accuracies and 96% to 97% recall values with training to testing data.

When the decision tree model is considered, for all three ratios of training to testing samples, both accuracies and recall values for training data are recorded at 100%. Then, with testing data, the accuracy values drop to 78%, 77%, and 83% with 80%:20%, 70%:30%, and 60%:40% training-to-testing ratios. For the same case, testing recall values drop to 79%, 77%, and 78%, respectively.

With the kNN model, for the 80%:20%, 70%:30%, and 60%:40% sample ratios, training accuracies are 87%, 86%, and 87%, and testing accuracies are 78%, 77%, and 78%, respectively. For each ratio of training samples, kNN record 93% recall values. Testing recall values for this model are 91%, 96%, and 96%.

Logistic regression records the highest testing accuracies and recall values by considering the above experimental findings. In addition, logistic regression shows a minimal difference in the performance between training and testing data compared to the rest of the models. The decision tree records the highest training accuracies and recall values, but the testing performances dropped significantly compared to the training values. kNN indicates better performances than the decision tree but not to the level of a logistic regression model. Furthermore, with logistic regression, the accuracy increases with the increment of the sample size of the training data and shows the optimal with 40%: 60% of training to testing data sets.

These selected models were trained using a randomly selected subset of the data. The accuracy of the models was measured using the remainder subset of the dataset. To identify the ideal training and testing data ratio, we varied this ratio and measured the performance for each case. Table 3 shows the performances of each model with the training and testing sets of data with proportions of 20%:80%, 30%:70%, and 40%:60% training to testing datasets.

According to the obtained observations, the performances of each model with the training sets are better than those of the respective testing sets. We want to select the model that performs well with the testing dataset. Also, we do not want our model to perform well with the training data and poorly with the testing data. Based on Figure 6, the logistic regression model showcases 91% accuracy when predicting ovarian cancer, with 80% of the data selected randomly from the dataset as the training data. When this model predicts ovarian cancer patients with the remaining 20% of testing data, the accuracy drops slightly to 87%. With the same training and testing datasets, logistic regression indicates a recall value of 96% for the training and 97% for the testing data. When 70% of data is randomly selected for the training data and 30% for the testing data, the logistic regression model shows 93% and 86% accuracy with training and testing data. Similarly, for the 60% to 40% training to testing data, this model indicates 94% to 87% accuracies and 96% to 97% recall values with training to testing data.

When the decision tree model is considered, for all three ratios of training to testing samples, both accuracies and recall values for training data are recorded at 100%. Then, with testing data, the accuracy values drop to 78%, 77%, and 83% with 80%:20%, 70%:30%, and 60%:40% training-to-testing ratios. For the same case, testing recall values drop to 79%, 77%, and 78%, respectively.

With the kNN model, for the 80%:20%, 70%:30%, and 60%:40% sample ratios, training accuracies are 87%, 86%, and 87%, and testing accuracies are 78%, 77%, and 78%, respectively. For each ratio of training samples, kNN record 93% recall values. Testing recall values for this model are 91%, 96%, and 96%.

Logistic regression records the highest testing accuracies and recall values by considering the above experimental findings. In addition, logistic regression shows a minimal difference in the performance between training and testing data compared to the rest of the models. The decision tree records the highest training accuracies and recall values, but the testing performances dropped significantly compared to the training values. kNN indicates better performances than the decision tree but not to the level of a logistic regression model. Furthermore, with logistic regression, the accuracy increases with the increment of the sample size of the training data and shows the optimal with 40%: 60% of training to testing data sets.

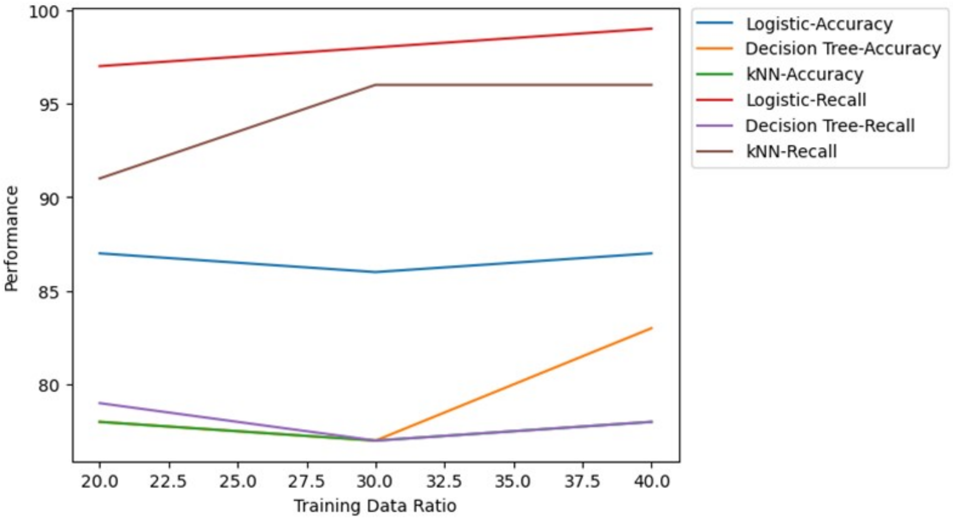


Figure 6: Performance of each model

Conclusion

Ovarian cancer has created a global health matter resulting in one fourth of all gynecological cancers. Therefore, studies conducted to prevent this cancer have a significant importance. This study aimed to classify whether subjects have ovarian cancer or not based on their selected features. Using a publicly available dataset consisting of subjects with ovarian cancer and some without cancer, and their 335 related variables, 49 features are selected for this study. Three models, Decision tree, kth Nearest Neighbor, and Logistic Regression models were trained by selecting a random sample of training datasets and testing the accuracy of the model using the remainder of the dataset. Two performance indicators, accuracy and recall values were calculated for each model with varied training to testing sample ratios. Based on experimental findings, logistic regression model outperformed the other two models recording 78% accuracy and 99% recall values to predict ovarian cancer patients. Kth Nearest Neighbor method perform better than the Decision Tree model, though not to the same level as Logistic Regression to predict the ovarian cancer in this study.

According to this study and other previously conducted studies, it is clear that not all models perform equally well across all the performance indicators and all datasets. Therefore, it is essential to implement different models and measure the classification performance for a given dataset before selecting the best performer.

The authors identify some of the limitations of this study and would like to correct them in a similar future study. This study utilized only three machine learning models, though other competitive models exist. It would be good to compare the performances with other models to compare the performances. Moreover, we used only two performance indicators, namely accuracy and recall values, though other available indicators based on the confusion matrix exist. Even with these identified imitations, the model we developed can be effectively used in predicting ovarian cancer in the future with the availability of the feature set of new patients.

References

Achilonu, O. J., Fabian, J., Bebington, B., Singh, E., Nimako, G., Eijkemans, M. J. C., & Musenge, E. (2021). Predicting colorectal cancer recurrence and patient survival using supervised machine learning approach: a South African population-based study. *Frontiers in Public Health*, *9*, 694306.

Agossou, C., Atchadé, M. N., Djibril, A. M., & Kurisheva, S. V. (2022). Mathematical modeling and machine learning for public health decision-making: the case of breast cancer in Benin. *Mathematical Biosciences and Engineering*, *19*(2), 1697-1720.

Akazawa, M., & Hashimoto, K. (2020). Artificial intelligence in ovarian cancer diagnosis. *Anticancer research*, *40*(8), 4795-4800.

Alabi, R. O., Mäkitie, A. A., Pirinen, M., Elmusrati, M., Leivo, I., & Almangush, A. (2021). Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. *International journal of medical informatics*, *145*, 104313.

Alafeef, M., Srivastava, I., & Pan, D. (2020). Machine learning for precision breast cancer diagnosis and prediction of the nanoparticle cellular internalization. *ACS sensors*, *5*(6), 1689-1698.

Alkhathlan, L., & Saudagar, A. K. J. (2022). Predicting and Classifying Breast Cancer Using Machine Learning. *Journal of Computational Biology*, *29*(6), 497-514.

Al Mudawi, N., & Alazeb, A. (2022). A model for predicting cervical cancer using machine learning algorithms. *Sensors*, *22*(11), 4132.

Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B. P., Mohanavel, V., ... & Asfaw, A. K. (2022). Lung cancer prediction from text datasets using machine learning. *BioMed Research International*, *2022*.

Board, P. A. T. E. (2023). Ovarian Epithelial, Fallopian Tube, and Primary Peritoneal Cancer Treatment (PDQ®). In *PDQ cancer information summaries [internet]*. National Cancer Institute (US).

Braz, D. C., Neto, M. P., Shimizu, F. M., Sá, A. C., Lima, R. S., Gobbi, A. L., ... & Oliveira Jr, O. N. (2022). Using machine learning and an electronic tongue for discriminating saliva samples from oral cavity cancer patients and healthy individuals. *Talanta*, *243*, 123327.

Chiu, Y. C., Zheng, S., Wang, L. J., Iskra, B. S., Rao, M. K., Houghton, P. J., & Chen, Y. (2021). Predicting and characterizing a cancer dependency map of tumors with deep learning. *Science Advances*, *7*(34), eabh1275.

Colombo, N., Van Gorp, T., Parma, G., Amant, F., Gatta, G., Sessa, C., & Vergote, I. (2006). Ovarian cancer. *Critical reviews in oncology/hematology*, *60*(2), 159-179.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34 (2), 113-127.

Ding, W., Chen, G., & Shi, T. (2019). Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. *Epigenetics*, *14* (1), 67-80.

Doan, L. M. T., Angione, C., & Occhipinti, A. (2023). Machine Learning Methods for Survival Analysis with Clinical and Transcriptomics Data of Breast Cancer. *Methods in molecular biology (Clifton, N.J.)*, *2553*, 325–393. https://doi.org/10.1007/978-1-0716-2617-7_16

Dochez, V., Caillon, H., Vaucel, E., Dimet, J., Winer, N., & Ducarme, G. (2019). Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review. *Journal of ovarian research*, *12*(1), 28. https://doi.org/10.1186/s13048-019-0503-7

Doppalapudi, S., Qiu, R. G., & Badr, Y. (2021). Lung cancer survival period prediction and understanding: Deep learning approaches. *International Journal of Medical Informatics*, *148*, 104371.

Famitha, S., & Moorthi, M. (2022). Intelligent and novel multi-type cancer prediction model using optimized ensemble learning. *Computer Methods in Biomechanics and Biomedical Engineering*, *25*(16), 1879-1903.

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360-150376.

Finkelstein, J., Cui, W., Martin, T. C., & Parsons, R. (2022). Machine Learning Approaches for Early Prostate Cancer Prediction Based on Healthcare Utilization Patterns. *Informatics and Technology in Clinical Care and Public Health*, *289*, 65.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, *2020*, 191.

Giampaolino, P., Foreste, V., Della Corte, L., Di Filippo, C., Iorio, G., & Bifulco, G. (2020). Role of biomarkers for early detection of ovarian cancer recurrence. *Gland surgery*, *9*(4), 1102–1111. https://doi.org/10.21037/gs-20-544

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics*, *22*(4), 291.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109.

Kouznetsova, V. L., Li, J., Romm, E., & Tsigelny, I. F. (2021). Finding distinctions between oral cancer and periodontitis using saliva metabolites and machine learning. *Oral diseases*, *27*(3), 484-493.

Lewis, R. J. (2000, May). An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California* (Vol. 14). San Francisco, CA, USA: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

Li, C., Liu, S., Zhang, Q., Wan, D., Shen, R., Wang, Z., ... & Hu, B. (2023). Combining Raman spectroscopy and machine learning to assist early diagnosis of gastric cancer. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *287*, 122049.

Li, X., Dai, A., Tran, R., & Wang, J. (2023). Identifying miRNA biomarkers for breast cancer and ovarian cancer: a text mining perspective. *Breast Cancer Research and Treatment*, 1-10.

Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J., ... & Jiang, J. (2020). Using machine learning to predict ovarian cancer. *International journal of medical informatics*, *141*, 104195.

Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International journal of medical informatics*, *108*, 1–8. https://doi.org/10.1016/j.ijmedinf.2017.09.013

Mahoto, N. A., Shaikh, A., Sulaiman, A., Al Reshan, M. S., Rajab, A., & Rajab, K. (2023). A machine learning based data modeling for medical diagnosis. Biomedical Signal Processing and Control, 81, 104481.

Maray, M., Alghamdi, M., & Alazzam, M. B. (2022). Diagnosing cancer using IOT and machine learning methods. *Computational Intelligence and Neuroscience*, *2022*.

Masud, M., Sikder, N., Nahid, A. A., Bairagi, A. K., & AlZain, M. A. (2021). A machine learning approach to diagnosing lung and colon cancer using a deep learning -based classification framework. *Sensors*, *21*(3), 748.

Matsas, A., Stefanoudakis, D., Troupis, T., Kontzoglou, K., Eleftheriades, M., Christopoulos, P., ... & Iliopoulos, D. C. (2023). Tumor markers and their diagnostic significance in ovarian cancer. Life, 13(8), 1689.

Matulonis, U. A., Sood, A. K., Fallowfield, L., Howitt, B. E., Sehouli, J., & Karlan, B. Y. (2016). Ovarian cancer. *Nature reviews. Disease primers*, *2*, 16061. https://doi.org/10.1038/nrdp.2016.61

Mi, Q., Jingting, Z., Ty, F., Zhenjiang, L., Jundong, X., Bin, C., ... & Xiao, L. (2020). Data for: Using Machine Learning To Predict Ovarian Cancer. *Mendeley Data, Version*, *11*.

Modugno, F., & Edwards, R. P. (2012). Ovarian cancer: prevention, detection and treatment of the disease and its recurrence. molecular mechanisms and personalized medicine meeting report. *International journal of gynecological cancer: official journal of the International Gynecological Cancer Society*, *22*(8), S45.

Moldagulova, A., & Sulaiman, R. B. (2017, May). Using KNN algorithm for classification of textual documents. In *2017 8th international conference on information technology (ICIT)* (pp. 665-671). IEEE.

Mourad, M., Moubayed, S., Dezube, A., Mourad, Y., Park, K., Torreblanca-Zanca, A., ... & Wang, J. (2020). Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis. *Scientific reports*, *10*(1), 5176.

Naik, J., & Patel, S. (2014). Tumor detection and classification using decision tree in brain MRI. *International journal of computer science and network security (ijcsns)*, *14*(6), 87.

Nguyen, M., Corbin, C. K., Eulalio, T., Ostberg, N. P., Machiraju, G., Marafino, B. J., ... & Chen, J. H. (2021). Developing machine learning models to personalize care levels among emergency room patients for hospital admission. *Journal of the American Medical Informatics Association*, *28*(11), 2423-2432.

Nissim, N., Dudaie, M., Barnea, I., & Shaked, N. T. (2021). Real-time stain-free classification of cancer cells and blood cells using interferometric phase microscopy and machine learning. *Cytometry Part A*, *99*(5), 511-523.

Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., & Garmire, L. X. (2021). DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome medicine*, *13*(1), 1-15.

Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. *International journal of environmental research and public health*, *19*(6), 3211.

Réda, C., Kaufmann, E., & Delahaye-Duriez, A. (2020). Machine learning applications in drug development. *Computational and structural biotechnology journal*, *18*, 241-252.

Redjdal, A., Bouaud, J., Gligorov, J., & Seroussi, B. (2022). Using Machine Learning and Deep Learning Methods to Predict the Complexity of Breast Cancer Cases. *Studies in health technology and informatics*, *294*, 78–82. https://doi.org/10.3233/SHTI220400

Sathiyanarayanan, P., Pavithra, S., Saranya, M. S., & Makeswari, M. (2019, March). Identification of breast cancer using the decision tree algorithm. In *2019 IEEE International conference on system, computation, automation and networking (ICSCAN)* (pp. 1-6). IEEE.

Sharma, S., & Mehra, R. (2020). Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. *Journal of digital imaging*, *33*, 632-654.

Shayan, Z., Mezerji, N. M. G., Shayan, L., & Naseri, P. (2016). Prediction of depression in cancer patients with different classification criteria, linear discriminant analysis versus logistic regression. Global journal of health Science, 8(7), 41.

Taghizadeh, E., Heydarheydari, S., Saberi, A., JafarpoorNesheli, S., & Rezaeijo, S. M. (2022). Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC bioinformatics*, *23*(1), 1-9.

Ting, W. C., Lu, Y. C. A., Ho, W. C., Cheewakriangkrai, C., Chang, H. R., & Lin, C. L. (2020). Machine learning in prediction of second primary cancer and recurrence in colorectal cancer. *International Journal of Medical Sciences*, *17*(3), 280.

Toprak, A. (2018). Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer. *Medical science monitor: international medical journal of experimental and clinical research*, *24*, 6537.

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), 1-16.

Urbanos, G., Martín, A., Vázquez, G., Villanueva, M., Villa, M., Jimenez-Roldan, L., ... & Sanz, C. (2021). Supervised machine learning methods and hyperspectral imaging techniques jointly applied for brain cancer classification. *Sensors*, *21*(11), 3827.

Venkatesan, E. V., & Velmurugan, T. (2015). Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology*, *8*(29), 1-8.

Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., ... & Ghassemi, M. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, *22*(1), 1-15.

Zhang, L., Li, C., Peng, D., Yi, X., He, S., Liu, F., ... & Huang, X. (2022). Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *264*, 120300.

Zhao, L., Xie, S., Zhou, B., Shen, C., Li, L., Pi, W., ... & Luo, H. (2022). Machine Learning Algorithms Identify Clinical Subtypes and Cancer in Anti-TIF1γ+ Myositis: A Longitudinal Study of 87 Patients. *Frontiers in Immunology*, *13*, 802499.

Zheng, S., Wu, Y., Donnelly, E. D., & Strauss, J. B. (2023). A cost-effective, machine learning-based new unified risk-classification score (NU-CATS) for patients with endometrial cancer. *Gynecologic Oncology*, *175*, 97-106.

Yang, W. L., Lu, Z., & Bast, R. C., Jr (2017). The role of biomarkers in the management of epithelial ovarian cancer. *Expert review of molecular diagnostics*, *17*(6), 577–591. https://doi.org/10.1080/14737159.2017.1326820