IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11625

PREDICTING OVARIAN CANCER USING MACHINE LEARNING

A.V.D.N. Murthy[1], M. Sai Jahanavi[2], N. Vinay[3], N. Priyanka[4], M.V.S. Nitish Varma[5], P. Manikanta[6]

[1]Assoc Professor, Lendi Institute Of Engineering & Technology, Jonnada, A.P., India.

[2,3,4,5,6]Student, B.Tech (CSE), Lendi Institute Of Engineering & Technology, Jonnada, A.P., India.

Abstract: In India, ovarian cancer is the third most frequent malignancy. Every year, it affects over a lakh people. In 2018, 295,414 women were diagnosed with ovarian cancer, and 184,799 women died from the disease globally, according to statistics. in their life. At some time in their life, one out of every 78 women will develop ovarian cancer. Because early-stage tumors are often asymptomatic, the vast majority of ovarian cancer patients are diagnosed with advanced disease. As a result, long-term survival appears to be improbable. To find out if you have this cancer, consult a doctor or travel to a diagnostic center, which can take time. A few statistical-based approaches to dealing with this problem are currently being explored, and they have become part of a partial answer to some extent. In the healthcare industry, machine learning has made a wide range of tools, methodologies, and frameworks available. Machine learning is the most effective connectionist technique for predicting cancer outcomes because it can identify and recognize patterns in complex datasets. Intending to lower mortality rates, this project provides a set of classification-based machine learning algorithms for cancer detection and prevention. Our goal is to create a simple predictive model that also performs well. It's done using classification techniques including Decision Tree (DT), Logistic Regression (LR), and Support Vector Machine (SVM) (SVM). The accuracy of each categorization method's conclusions is compared.

keywords: Ovarian cancer, Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM).

INTRODUCTION:

Ovarian cancer is cancer that starts in or grows on the ovary. It promotes the development of abnormal cells that can infiltrate and spread throughout the body.

There may be no or simply a few unclear symptoms when this process begins. The symptoms of cancer become more obvious as the disease develops. Bloating, pelvic pain, stomach swelling, constipation, and loss of appetite are some of the symptoms. The stomach lining, lymph nodes, lungs, and liver are just the areas where cancer can spread. Ovarian cancer (OC) is the eighth most frequent disease in women, according to the Centers for Disease Control and Prevention (CDC). Most ovarian cancer patients are discovered with advanced illness because early-stage tumors are generally asymptomatic, resulting in a low long-term survival rate. Even though ovarian cancers are chemo-sensitive (a feature of cancer cells that describes the strength of the tumor's reaction to a given anti-cancer drug.)5-year recurrence rates in people with the advanced disease vary from 60% to 80%, despite initial success against platinum first-line chemotherapy combination treatment. As a result, a lot of work has gone into developing new approaches for forecasting disease progression and prognosis in this type of cancer.

1. CLASSIFICATION

In Machine Learning, Classification algorithms are used when the output variable is categorical, which means there are twoclasses such as Yes-No, Male-Female, True-false, etc. Classification Techniques are further divided into different types of models.

Some of them are:-

1. Logistic Regression
2. Naive Bayes Classifier
3. Nearest Neighbor
4. Support Vector Machines
5. Decision Trees
6. Random forest

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022
DOI: 10.17148/IJARCCE.2022.11625

2. LITERATURE SURVEY

Ming Yang Lu (2020):
This was thoroughly investigated, and it was discovered that we are using a dataset of 300 patients with 49 variables including demographics, blood routine, and cancer markers(tumors). The ovarian cancer risk of malignancy algorithm was developed using a decision tree (DT), logistic regression (LR), and an ovarian cancer risk of malignancy algorithm (ROMA)

Shiva Pirhadi (2021):
In this case, a decision tree is employed to make predictions. After that, the model is trained and the data is tested. The objectives are being met. They came to this conclusion based on protein levels in healthy and unhealthy samples. Methods such as the t-test, entropy, imperialist competitive algorithm, and KNN are used in this case.

Ahmedin Jemal DVM (2018):
The American Cancer Society forecasts the current range of cancer incidence and fatalities based on data from the National Center for Health Statistics in 2015. The United States is expected to see 1,735,350 new cancer cases and 609,640 cancer deaths in the year 2018.

David A.Fishman (2002):
Over the last ten years, the incidence of ovarian cancer has steadily increased, with an overall lifetime risk of 1.8 percent. After the age of 40, the occurrence of ovarian cancer rapidly increases, with the mean age of occurrence at 60. Surgical intervention is frequently required for patients with stage 1 disease. They have a 5-year survival rate of about 90%.

Richard G.Moore (2010):
In this study, we compared the Risk of Malignancy Index (RMI) and the Risk of Ovarian Malignancy Algorithm (ROMA) to predict epithelial ovarian cancer (EOC) in women.

3. EXISTING SYSTEM:

Ovarian cancer is now anticipated by the usage of entropy, imperialist aggressive algorithms, and KNN.

Limitations

- We need to continually decide the values of K inside the contemporary system, which may be complex at times.
- The calculation value is excessive due to the fact we're calculating the space among statistics factors for all education samples.
- Model has low efficiency.

4. PROPOSED SYSTEM

The primary goal of this research is to predict the disease at an early stage so that mortality rates can be reduced. Here, the system is made up of three techniques: decision tree, logistic regression, and support vector machine.

Steps to build a model:

1. First, the required ovarian dataset is obtained.
2. All of the libraries required for model preparation are being imported.
3. After that, the dataset is imported into the model for preprocessing.
4. The data is sent to be preprocessed. We calculate the mean which is used to cope with missing data.
5. Data encoding is now complete. Encoding is the process of converting data or a specified sequence of characters, symbols, alphabets, and so on into a specific format for secure data transmission.
6. The dataset has now been divided into 70:30 training and testing data.
7. The dataset's independent variables are within a defined range using a technique known as feature scaling.
8. The model is now being trained using the training dataset. The essential features are taken.
9. Using testing data, we then test the model to determine whether it is working properly or not. The accuracy is then calculated using the precision, recall, and F1 score metrics.
10. Finally, the comparison of three algorithms is done to compute the accuracy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11625

5. DATA DESCRIPTION

The primary goal of this study is to analyze multiple classification algorithms and apply them to a dataset to determine which model provides the most accuracy and precision while reducing false positives.
We collected an Ovarian cancer dataset with 49 attributes of 349 patients for the study. Given in our dataset, consists of 49 variables which include histology, demographic, blood routine test, general chemistry, and tumor markers. For the variables in the dataset, most have a low missing value rate (less than 7% missing), except for Neutrophil Ratio (percent) and CA74-4, which is lacking 69 percent of the time (100 percent complete in training and 80 percent complete in testing). For the variables that have a low missing value rate, they were imputed by using their mean.

6.SYSTEM DESIGN & IMPLEMENTATION

6.1 System Architecture

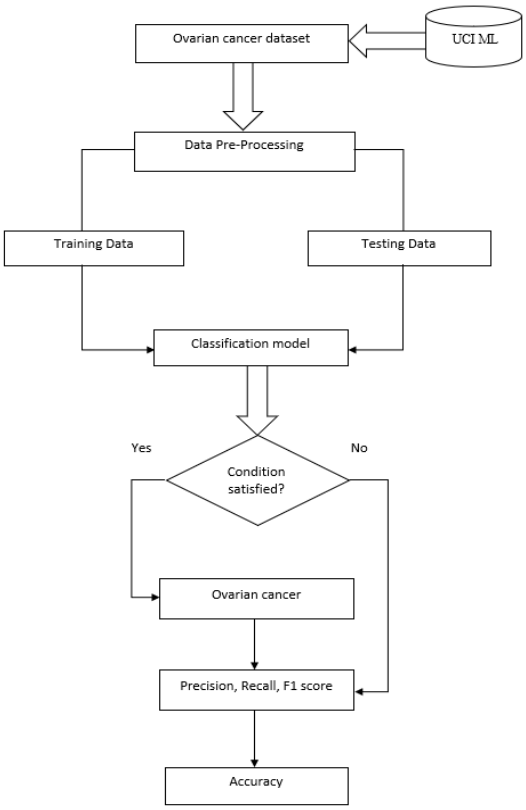

Fig 1: System Architecture

6.2 Data Preprocessing

Data Preprocessing is the first step in the machine learning process that must be completed before the process can be started. Raw and unfiltered data is transformed and converted into a more acceptable and intelligible format via data preprocessing.
Data preprocessing is a technique for transforming unclean data into a usable set. In other words, anytime data is acquired from various sources, it is obtained in raw format, which makes analysis impossible. Data Preprocessing consists of the following steps:

1. Libraries to Import
2. The dataset is being imported.
3. Verifying that no values are missing
4. Encoding Categorical Data
5. Splitting the Dataset into a Training set and Test Set
6. Scaling of Features

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11625

Decision Tree:

A non-parametric supervised learning method is used for classification and regression. Decision trees are used for a variety of reasons that are simple to understand and interpret. Based on the built tree, they can create if-then rules to explain the underlying relationships between biomarkers and ovarian cancer. Because our dataset contains both numerical and categorical biomarkers, these require minimal data preprocessing and can handle both numerical and categorical data. Each leaf node in the decision tree corresponds to a class label of the target variable, and each internal node represents a feature. The decision tree, which is a binary tree, was created using an Iterative Dichotomiser 3. (ID3). Each attribute's information gain is calculated, and the attribute with the highest information gain is designated as the root node. Label the attribute as a root node, and the attribute's possible values are represented as arcs.

The steps of ID3 are described as follows:

1. Calculate the entropy values for the dataset.
2. For each attribute/feature.
   - 2.1 Calculate entropy for all its categorical values.
   - 2.2 Calculate the information gain for each feature.

3 Find the feature with maximum information gain.

4 The technique is repeated until the desired tree is obtained.

Logistic regression:

Logistic regression is an easy and broadly used supervised system studying a set of rules for -magnificence classification. It is a statistical approach for reading a hard and fast of facts that carries one or greater unbiased variables that affect the results. A dichotomous variable turned into used to assess the outcome. It refers to the fact that the most practical training is feasible. The aim variable is expressed in nature and follows the Bernoulli distribution in this special case of linear regression.

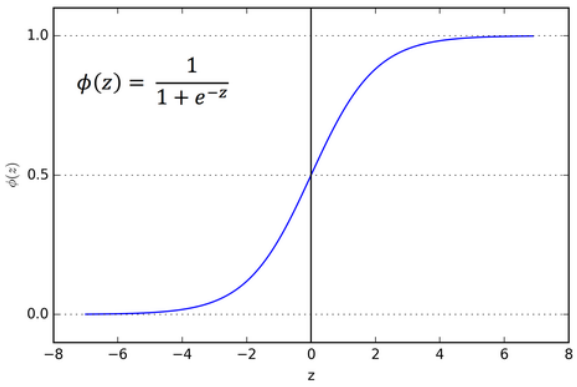

Fig 2: Sigmoid Function

Linear Regression Equation

$y = \beta 0+\beta 1X1+.....+\beta nXn$

Where, y is the dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:

$$\mathbf{P} = \frac{1}{1 + e^{-y}}$$

Apply Sigmoid function on linear regression:

$$\mathbf{P} = \frac{1}{1 + e^{(-\beta 0 + \beta 1 x 1 + \ldots\ldots\ldots \beta n X n)}}$$

Sigmoid function

The sigmoid function, also known as the logistic function, produces an 'S-shaped curve that may convert any real-valued number to a value between 0 and 1.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11625

When the curve reaches positive infinity, y predicted becomes 1, and when it reaches negative infinity, y predicted becomes 0. If the output of the sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it as 0 or NO.

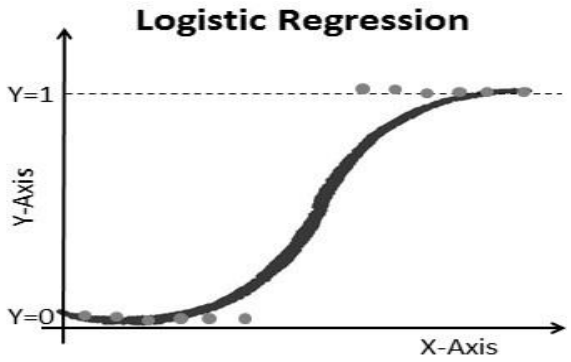

Fig 3: Logistic Regression

So, unlike linear regression, we get an 'S-shaped curve in logistic regression.

Support vector machine:

A Support Vector Machine (SVM) is a binary linear classification that uses an explicit decision boundary to reduce generalization error. It's a versatile Machine Learning model that can conduct linear and nonlinear classification regression as well as outlier detection. SVM is useful for categorizing small or medium-sized complex datasets. The purpose of the SVM algorithm is to find a hyperplane that distinguishes between data points in an N-dimensional space. The number of features determines the hyperplane's size. The hyperplane is essentially a line if there are just two input features. The hyperplane converts into a two-dimensional plane when the number of input features approaches three. It gets impossible to imagine when the number of elements exceeds three.

Consider two independent variables, x1, and x2, as well as one dependent variable, a blue or red circle.

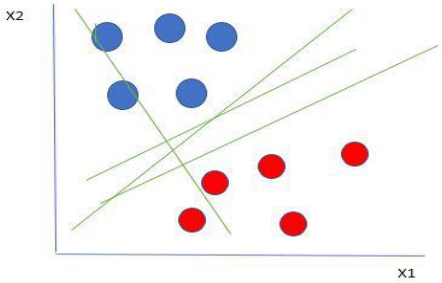

Fig 4: Classifying data

Multiple lines (our hyperplane here is a line because we're just examining two input features, x1, and x2) separate our data points or classify them into red and blue circles.

Selecting the best hyper-plane:

The hyperplane that represents the greatest separation or margin between the two classes is a viable choice as the best hyperplane.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11625

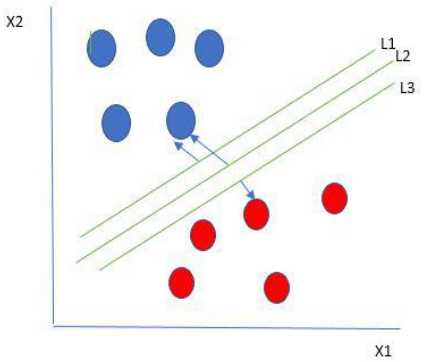

Fig 5: Describes SVM Hyperplane

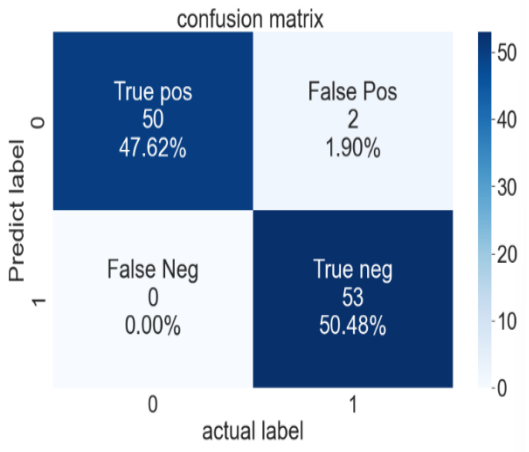

Fig 6: Confusion matrix of decision tree

RESULTS:

The decision tree predicts there are 50 True Positives (47.62%) and 53 True Negatives (50.48%). There are 2 false positives (1.90 percent) and 0 false negatives (0.00 percent). The sensitivity values for precision, recall, and F1 score is 96.3636, 1.0, 0.9815, and 0.9615384, respectively. The decision tree's accuracy is 98.0952.

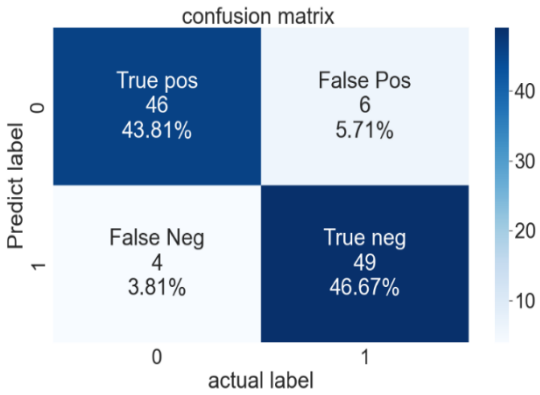

Fig 7: Confusion matrix of logistic regression

Logistic Regression has predicted 54 (51.43%) True positives and 45(42.86%) True Negatives. There are 3(2.86%) False positives and 3(2.86%) False Negatives. The precision, recall and F1 score, sensitivity values are obtained as 92.4528, 0.8909, 0.9074, 0.884615 respectively. The accuracy of the logistic regression 90.4762.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11625

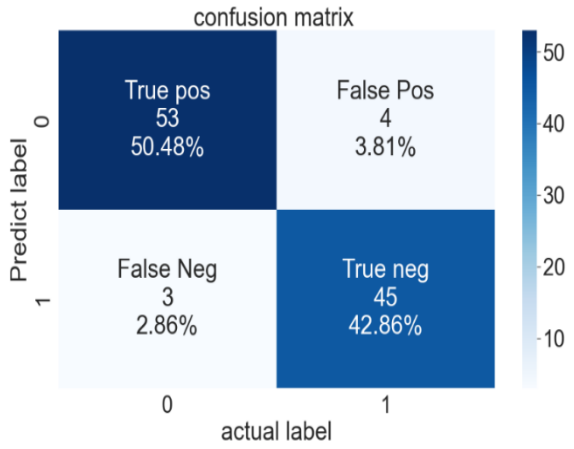

Fig 8: Confusion matrix of Support vector machine

SVM has predicted 53(50.48%) True positives and 45(42.86%) True Negatives. There are 4(3.81%) False positives and 3(2.86%) False Negatives. The precision, recall, and F1 score, sensitivity values were obtained as 94.3396, 0.8929, 0.9174, and 0.884615384 respectively. The accuracy of the support vector machine is 91.4286.

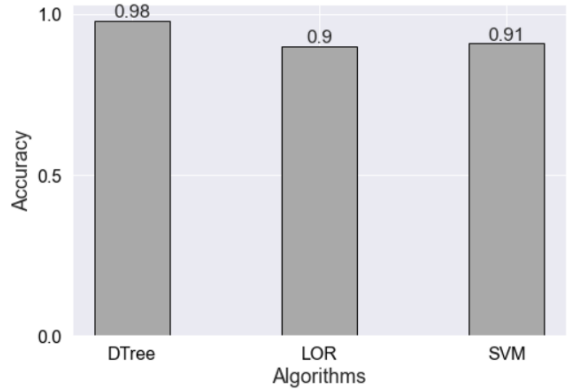

Fig 9: Graphical representation of accuracy

Algorithm	Accuracy	
Decision tree	0.98	(98%)
Logistic regression	0.9	(90%)
Support Vector machine	0.91	(91%)

The accuracy of the model is working well with the decision tree as compared to logistic regression and support vector machine.

7. CONCLUSION

Predicting the course of ovarian cancer is challenging because it is asymptomatic in the early stages. As a result, patients with ovarian cancer have a lower chance of survival. As a result, we presented a more accurate way of detecting the condition at an early stage. We used techniques such as decision tree (DT), logistic regression (LR), and support vector machine in this (SVM). We look into which of the three proposed techniques produces the best results.

REFERENCES:

1. Mingyang Lu, ZhenjiangFan, et al. “Using machine learning to predict ovarian cancer.”. Available at International Journal of Medical Informatics. Vol 141(2020):104195.
2. R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2018, CA: a cancer journal for clinicians 68 (1) (2018) 7–30.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11625

3. D.A. Fishman, K. Bozorgi, The scientific basis of early detection of epithelial ovarian cancer: the National Ovarian Cancer Early Detection Program (NOCEDP), Cancer treatment and research 107 (2002) 3–28.
4. R.G. Moore, M. Jabre-Raughley, A.K. Brown, K.M. Robison, M.C. Miller, W.J. Allard, R.J. Kurman, R.C. Bast, S.J. Skates, Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass, American Journal of Obstetrics and Gynecology 203 (3) (2010) 228.e221-228.e226.
5. Pirhadi S, Maghooli K, Moteghaed NY, Garshasbi M, Mousavirad SJ. Biomarker discovery by the imperialist competitive algorithm in mass spectrometry data for ovarian cancer prediction. J Med Signals Sens 2021;11:108-19
6.https://www.datacamp.com/community/tutorials/understanding-logistic-regressionpython - Introduction to logistic regression, datacamp.
7.https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logisticregression/-sigmoidfunction, analyticsvidhya.
8.https://www.datacamp.com/community/tutorials/svm-classification-scikit-learnpython