IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Ovarain Cancer Prediction in Early Stage Using Machine Learning Approaches

DOI:10.48047/IJFANS/V11/I12/186

Dr. K. Lohitha Lakshmi[1], Associate Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
P. Hima Chandana[2], **P. Hema Sri**[3], **N. Nitish Kumar**[4], **N. Hemanth**[5]
[2,3,4,5]UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
lohita.kanchi@gmail.com[1], himachandanap@gmail.com[2],
hemaperumalla20@gmail.com[3], nitishkumarnavuluri@gmail.com[4],
nallamothuhemanth50@gmail.com[5]

Abstract

Ovarian cancer is a disorder of ovarian cell growth that is triggered by series of acquired mutations affecting a single cell or its clonal progeny. It is purposeless prey on host and virtually autonomous. It is usually diagnosed at a late stage because of poor sensitivity of screening test. There are still no effective cures for this illness. Still early detection might lower the mortality rate. Our project's major goal is to conduct predictive analytics for early detection by using machine learning models and statistical techniques on clinical data collected from specific patients. Mutual information testing is crucial in statistical analysis for identifying indicative biomarkers. A collection of machine learning models, such as the Random Forest (RF), Extreme Gradient Boosting Machine (XGBoost), Logistic Regression (LR), Gradient Boosting Machine (GBM), and Light Gradient Boosting Machine (LGBM)are utilized in the classificationof ovarian tumors as benign or malignant. By using proposed system, it can significantly identify the class of benign and malignant patients. The data collected is analyzed and pre-processed before it is used for model training and testing.

Keywords: Biomarkers, Mutual Information, Ovarian Cancer, Predictive Models, Risk Factors.

1. Introduction

Ovarian cancer is an abnormal mass of ovarian tissue whose growth outpaces and is not coordinated with the normal tissue[10]. It also continues to grow excessively even after the initial stimulus that caused the alteration has stopped. Although most of the cases are typically seen in elderly women, ovarian cancer is also being diagnosed in young women. Due to inadequate screening methods and a lack of distinctive symptoms, more than 70% of ovarian cancer patients receive a diagnosis of the disease in a late stage. Trans Vaginal Ultra Scan (TVS) and Cancer Antigen 125 (CA125) Blood Test are the current screening methods available to detect ovarian cancer. Usually elderly women present with pain abdominal distention. In some cases it is associated with bloating. Age, certain genetic mutations, and a family history of the disease are all risk factors for ovarian cancer. There are various studies evaluating the effectiveness of the biomarkers that distinguish between benign tumour and ovarian cancer. Researchers have been investigating the use of machine

learning algorithms to identify ovarian cancer which are derived from many data sources, including clinical data, genetic data also medical imaging, in recent years. The ability to predict disease progression and the diagnosis of cancer is a great potential of machine learning algorithms with new approaches. Machine learning is widely accepted technique to understand the prognosis of the disease. In order to develop predictive analytics for early detection, this study used mutual information to identify pertinent features and machine learning models.

2. Literature Survey

Martuza et al. made work to apply machine learning models to the clinical data which is considered from 349 patients[1]. 49 features are present which are sub divided into three groups. The results obtained from Random Forest(RF),Gradient Boosting Machine(GBM),Light Gradient Boosting Machine (LGBM) classifiers showed high level of accuracy of 88 percent.

Arcuda et al. employed serum proteome profile data to drive wavelet feature selection in machine learning methods.

Patrick et al. made efforts to predict ovarian cancer using regularized logistic regression[2]. 349 medical records from the “Third Affiliated Hospital of Soochow University” make up the data set used in this study. A logistic regression model is constructed with a Least Absolute Shrinkage and Selection Operator (LASSO) regularization penalty. Accuracy of 90.6% is obtained using Logistic regression.

SuthamerthiElavarasu et al. made a review on Machine Learning Applications in Ovarian Cancer Prediction[3]. Three principle approaches are stated for the selection of features, namely integrated, filters and envelopes approach. The survey concludes that by the analysis of various studies for predicting the outcomes of disease, the machine learning techniques and classification algorithms provides useful tools.

Viji Vinod et al. considered the TCGA ovarian cancer database of gene expression values. Machine learning model SVM showed the highest accuracy for recurrence and survival predictions.Yang et al. showed that the decision tree combined with Support Vector Machine-Synthetic Minority Over sampling Technique (SVMSMOTE) showed the top PPV (Positive Predictive Value) with 0.9041[4].

Munetoshi et al. had established that the histo-pathological identification of ovarian cancer may be predicted using artificial intelligence from preoperative assessment [5].

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Lu et al. decided to evaluate 49 features from patient characteristics to build machine learning models. For performing modeling procedure they had used the combination of method named Minimum Redundancy Maximum Relevance (MRMR) feature selection, ReliefF feature selection also decision tree analysis. Finally they found that a prediction accuracy of 92.1% through decision tree approach using HE4 and carcino-embryonicantigen(CEA)[7].

Many studies have used CA125 as a marker of ovarian tumor[8]. To differentiate benign and malignant tumor In 1990, Jacobs et al. used the following factors- age, ultrasound status, clinical feature, menstrual history ,CA-125 levels as features to distinguish between benign and malignant ovarian tumors. Their experiment yielded a sensitivity of 81% and specificity of 75%.

He has used the data set consisting of 202 patients information from preoperative examinations. Highest accuracy of 80% was obtained using XGboost machine learning model.

R. Kasture et al. had used the histopathological images for the prediction of ovarian cancer[9]. Deep Learning method DCNN is used for training and evaluation[10]. Results of which highest accuracy of 91% was obtained from the KK-net model.

3. Problem Identification

Ovarian cancer is challenging to diagnose at an early stage because of non-specificity of the signs and symptoms. The problem with ovarian cancer is that it is often not recognized until it is advanced, making it more difficult to treat and less likely to have a successful out come. As the incidence of ovarian cancer is increasing day by day, early diagnosis can reduce the morbidity and mortality.

The significance of late-stage ovarian cancer diagnos is highlights the importance of early detection and the need for improved screening and diagnostic tools to improve patient outcomes[6]. Often there are challenges for developing treatment for all ovarian cancer patients such as heterogeneity of ovarian cancer and also side effects of the treatment. [12-20]

4. Methodology

Screening strategies that are available are Trans-vaginal Sonography(TVS) and CA125, but neither is specific enough to identify cancer when used alone. As a result we proposed a solution for early prediction of ovarian cancer. For early prediction identification of bio

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

markers plays a significant role. To identify the significant bio markers from the data set that helps in prediction we implemented a feature selection technique mutual info.

4.1 Data Set

The samples from patients with benign and malignant ovarian tumour used in this investigation were taken from a clinically tested raw data setwas gathered by the "Third Affiliated Hospital of Sooc how University" during the period of July 2017 to July 2018. A total of 349 patients samples are present in the present in the data set out of them there are 171 people with ovarian cancer and 178 people with benign tumours.

Figure 1 List of attributes[1]

4.2 Data Scaling

We have applied data scaling, often referred to as feature scaling, to transform the values of the many attributes in our data set on a single scale. Many of automated learning techniques are sensitive to the scale of input data and may not operate as intended if the data is not scaled properly. Data scaling is done by standardizing the data by subtracting the mean from actual value and dividing with the standard deviation.

$$\text{Scaled Value} = (\text{actual value} - \text{mean})/\text{deviation} \quad (1)$$

As a part of statistical analysis we have considered mutual information for feature selection.

4.3 Feature Selection

The process of selecting the pertinent characteristics from a huge data collection that contribute most to the prediction is known as feature selection[7].The fundamental objective of feature selection is to lessen model complexity and training time.

4.4 Mutual Information

A statistical entity known as mutual information measures the degree of dependency between two variables. In this context, mutual information is used to identify which

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

features or biomarkers are more strongly associated with the presence of ovarian cancer. Mutual information between two variables X and Y can be calculated as

$$MI(X,Y) = H(X)+H(Y)-H(X|Y) \quad (2)$$

Where H(X) is the entropy of X, H(Y) is the entropy of Y and H(X|Y) is the joint entropy.

$$H(X) = -\sum p(X) \log p(X) \quad (3)$$

$$H(Y) = -\sum p(Y) \log p(Y) \quad (4)$$

The Machine Learning models plays a significant role in prediction or classification. In this Study the Models implemented are Random Forest, Logistic Regression, Gradient Boosting Machine, Light Gradient Boosting Machine, Extreme Gradient Boosting and Voting Classifier.

5. Implementation

To implement the proposed methodology we have used python programming language. The popular python library Scikit Learn provides wide range of algorithms for data scaling, feature selection and model training.

5.1 Random Forest:

An ensemble learning technique, random forests or random decision forests build a large number of decision trees during the training phase. It is applied to both classification and regression models. Each decision tree in this model is created by selecting a subset of features and data points. A majority vote or average is used to evaluate the output.

5.2 Logistic Regression:

A statistical technique known as logistic regression is primarily used to predict the connection between a binary dependent variable and one or more independent variables for binary classification[2]. Sigmoid Function was calculated for the prediction of output class label.

5.3 Gradient Boosting Machine:

It is a boosting method which combines several weak predictors to form a strong predictor by optimizing a loss function. A Loss Function is calculated as the difference in predicted value and actual value. Each decision tree is constructed so as to minimize this error.

5.4Light Gradient Boosting Machine(LGBM):

It is a framework designed to increase model efficiency and reduce memory usage. It works on the basis of two techniques - A. Gradient based one side sampling (GOSS). B. Exclusive

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Feature bundling or EFB.GOSS is used to select the features that have high gradient which means features which has more predictive ability of class label where as EFB eliminates the features that are mutually exclusive.

5.5Extreme Gradient Boosting(XGBoost):

It is an execution of gradient boosted decision tree in which weights plays an important role. Unlike all other boosting models XGboost has built parallel processing which makes it train the large datasets in less time.

5.6 Voting Classifier:

The ensemble voting classifier based its output prediction on the projected class with the highest probability after being trained on a variety of models.

Steps involved in the implementation:

1. Importing Library Functions
2. Loading of Data
   data = load_data()
3. Train and Test Split
   80% of the data is used for training and 20 % for testing.
4. Datascaling is done by the process of Standardization.
5. Feature Selection using mutual info.
6. Defining all the models.
7. Training the models with the features obtained.

The Evaluation metrics considered are:

Accuracy: The percentage of accurate predictions among all other predictions is known as accuracy.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

Precision: when a positive outcome was anticipated, precision is the percentage of true positives.

$$\text{Precision} = TP / (TP + FP) \quad (6)$$

Recall: Among all positive instances in the data set the predictions that are predicted as true positive.

$$\text{Recall} = TP / (TP + FN) \quad (7)$$

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

F1 Score: The harmonic mean of recall and precision can be used to get the F1 Score.

$$F1\ score = 2 * (Precision * Recall) / (Precision + Recall) \quad (8)$$

TP - True Positive,

TN - True Negatives,

FP - False Positives,

FN - False Negatives.

6. Analysis Of Result

As per our research, computing time required to train a model increases as more and more features are included. So, by the feature selection process, we have not considered the features that have little or no impact on the prediction of ovarian cancer. These characteristics may be crucial for predicting other malignancies but not in ovarian cancer. Since only the best features were taken into account through k best features function, the predictive power of the model has not dropped. Moreover, fewer attributes aid in speeding up model computation. While voting classifier demonstrated accuracy of 90% for 15 features, random forest demonstrated accuracy of 91%.The three metrics by which the random forest model outperforms all other models are precision, recall and F1 score.

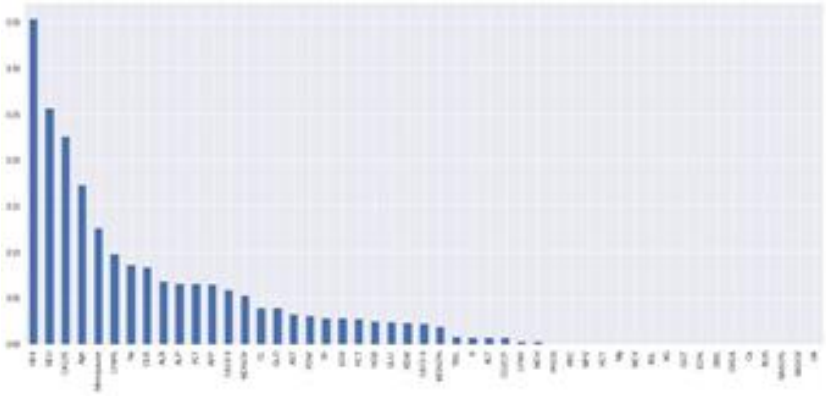

Figure 2: Mutual Information Values

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

k=number of features	Model	Accuracy	Precision	Recall	F1_score
49	RF	0.88	0.87	0.92	0.89
	LR	0.82	0.86	0.81	0.83
	XGB	0.88	0.89	0.89	0.89
	GBM	0.87	0.89	0.89	0.89
	LGBM	0.88	0.87	0.92	0.89
	Voting	0.88	0.89	0.90	0.89
25	RF	0.90	0.89	0.92	0.90
	LR	0.84	0.82	0.89	0.86
	XGB	0.90	0.91	0.89	0.90
	GBM	0.85	0.82	0.89	0.86
	LGBM	0.87	0.85	0.92	0.88
	Voting	0.88	0.89	0.88	0.89
15	RF	0.91	0.90	0.94	0.92
	LR	0.90	0.86	0.97	0.91
	XGB	0.85	0.86	0.86	0.86
	GBM	0.87	0.86	0.86	0.86
	LGBM	0.90	0.87	0.94	0.91
	Voting	0.90	0.89	0.92	0.90

Table 1 Accuracy and evaluation metrics

7. Conclusion.

To conclude, in order to save computing time and improve model performance, relevant biomarkers must be identified using feature selection. These discovered biomarkers aid in the early detection of ovarian cancer.

8. Limitations & Future Scope

Early stage detection of ovarian cancer is difficult due to lack of symptoms, low incidence rate, variations in tumor types, accessibility of early stage detection test etc. Our data set is limited to only 349 patients, if larger data set is available the models will be trained on more number of samples which might increase the performance of models.

Our project could be developed in the future to distinguish between various ovarian cancer sub-types. The clinical data set was used to anticipate the incidence of ovarian cancer; moving forward, the project can be expanded to include image data set.

9. References

[1] Md. Martuza Ahamad, Sakifa Aktar, Md. Jamal Uddin, Tasina Rahman, Salem, Samer AI-Ashhab, AKM Azad, and Mohammad Ali Moni "Early Stage Detection Of Ovarian

Cancer Based on Clinical Data Using Machine Learning Approaches", Journal Of Personalized Medicine, Vol. 12, July 2022.

[2] Anna F.Han and Patrick Emedom-Nnamdi "Predicting Ovarian Cancer Using Regularized Logistic Regression", July 2021.

[3] SuthamerthiElavarasu, Viji Vinod and Elavarasan Elangovan "Machine Learning Applications in Ovarian Cancer Prediction: A Review", International Journal of Pure and Applied Mathematics, Vol. 117, Issue 20, 2017.

[4] Xiaoyan Yang, Matlob Khushi and kamaran Shaukat, "Biomarker CA125 Feature Engineering and Class Imbalance Learning Improves Ovarian Cancer Prediction", IEEE Access 2020.

[5] Munetoshi Akazawa and Kazunori Hashimoto, "Artificial Intelligence in Ovaian Cancer Diagnosis", Anticancer Research 40, 4795-4800, 2020.

[6] Sreeja Sarojini, Ayala Tamir, Heeiin Lim, Shihona Li, Shifana Zhang, Andre Goy, Andrew Pecora and Stephen "Early Detection Biomarkers for Ovarian Cancer", Jounal Of Oncology, Volume 2012.

[7] Mingyang Lu, Zhenjiang Fan, Bin Xu, Lujun Chen, Xiao Zheng, Jundong Li, Qi Mi and Jingting Jiang "Using Machine Learning to predict Ovarian Cancer", Journal od Medical Informatics, 1386-5056, May 2020.

[8] VincentDochez, Helene Caillon,Edouard Vaucel, Jerome Dimet, Norbert Winer and Guillaume Ducarme "Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA a review", Issue 10, 2019.

[9] Kokila. R. Kasture and Praven N.Matte, "Prediction and Classification of Ovarian Cancer using Enhanced Deep Convolutional Layer", Volume 70, Issue 3, 310-318, March 2022.

[10] Mansi Mathur, Vikas Jindal and Gitanjali Wadhwa, "Detecting Malignancy of Ovarian Tumour Using Convolutional Layer: A Review", IEEE Access, 2020.

[11] Robert C.Bast, Zhen Lu, Chase Young Han, Karen H. Lu, Karen S. Anderson , Charles W. Drescher and Steven J. Stakes "Biomarkers and Strategies for Early Detection Of Ovarian Cancer", Departments of Experimental Therapeutics and Gynecologic Oncology, Texas University, October 2020.

[12] Sri Hari Nallamala, et al., "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[13] Sri Hari Nallamala, et.al., "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[14] Sri Hari Nallamala, et.al, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer

Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[15] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

[16] Sri Hari Nallamala, et.al, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[17] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., "Accurate Hand Gesture Recognition using CNN and RNN Approaches" International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[18] Sri Hari Nallamala, et al., "A Review on 'Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management'", Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[19] Nallamala, S.H., et al., "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems", IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[20] Nallamala, S.H., Mishra, P., Koneru, S.V., "Breast cancer detection using machine learning approaches", International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.