OVARIAN CANCER DETECTION USING MACHINE LEARNING APPROACHES BASED ON CLINICAL DATA

*S.Mahalakshmi**

ABSTRACT

Ovarian cancer is one of the most predominant forms of cancer in women. Currently, there is still no effective medication therapy available to treat this fatal illness. On the other hand, early discovery may lengthen the patients' lives. This work's main objective is to do predictive analytics for early diagnosis by applying machine learning models and statistical techniques to clinical data collected from 549 patient people. In statistical analysis, classification models to separate patients with benign from malignant ovarian cancer are constructed using Student's Gradient Boosting Machine (XG Boost). Furthermore, the predictive analysis results indicate that the machine learning model can distinguish between malignant and benign patients with up to 91% accuracy. Machine learning detection may be crucial to the diagnosis of cancer since early-stage detection is typically not possible.

Keywords: Machine Learning, Ovarian Cancer, Clinical Data, Detection, Algorithms, Predictive Modeling, Data Analysis, Feature Extraction, Classification, Healthcare, Medical Research, Diagnostic Tools, Precision Medicine, Bio-informatics, Women's Health

I. INTRODUCTION

Ovarian cancer (OC), one of the more common cancers, is the 7th most shared disease in women and has a lifetime risk factor of 2.7% [1]. Although 2.5% of all cancers in women are ovarian cancers, only 5% of these cases are fatal due to poor survival rates. The absence of early symptoms and the late diagnosis are typically blamed for this [2]. Chemotherapy-sensitive ovarian tumours have inherent resistance to platinum/taxane therapy, with a 60–80% 5-year recurrence rate [3].

Department of Computer Science and Engineering
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
mahakoresh31@gmail.com
* Corresponding Author

Gynaecologists are typically in charge of figuring out whether a patient has developed malignant pelvic masses, which can be mistaken for tumours [4]. Tumour biomarkers such as have been used to distinguish benign non-gynecologic conditions from benign tumours, while other techniques, such as helical CT scanning and ultrasonography, have also been used to do so [4], are important factors in identifying female pelvic masses [4,5].

The effectiveness of those indicators in separating benign tumours from ovarian cancer has been determined in a few trials. Moore et al. found that ROMA had a increased susceptibility to ovarian epithelial cancer patient prediction than RMI in a study analysing the RMI and ROMA algorithms for predicting the incidence of epithelial ovarian cancer in 457 patients [6]. When Anton et al. examined the sensitivity of HE4, ROMA, RMI, and CA125 in 128 individuals; they found that HE4 had the uppermost compassion for determining if an ovarian tumour was malignant [7].

Furthermore, Zhang et al. created a multi-marker linear model using CA125, HE4, progesterone, and oestradiol to forecast the course of ovarian cancer [8]. Novel methodologies in machine learning algorithms hold great potential for forecasting the course of disease and diagnosing cancer. Alqudah et al. used a wavelet feature selection technique in their application of artificial intelligence algorithms [9]. In order to predict the size of the tumour, However, the models only managed to obtain an AUC score of less than 70% [10]. With the use of a 4 staged OC, old data, numerous primary treatment options, and information on chemotherapy regimens, Paik et al. were able to predict the cancer stages with an accuracy score of approximately 83% [11]. Newly, Akazawa et al. shown a The exploration based on machine learning using a variation of models, comprising SVM, Naive Bayes, XGBoost, LR, and RF.

Among other competing models, the XG Boost algorithm produced better typical enactment, with the best accuracy score of about 80% [12]. Nonetheless, the size of the set of characteristics had an impact on this investigation; that is, the accuracy decreased by approximately 60% as the number of features decreased. A further limitation of this work has been its small feature set just 16 distinct blood parameters. Oxygen consumption markers, samples of blood, and diagnostic tests for typical chemistry were the three types of biomarkers used by Lu et al. They demonstrated a low accuracy rating but a high validation accuracy score [5], indicating a possibility of over-fitting, a problem common to machine learning algorithms. Consequently, the current need for a robust framework that makes use of statistical methods and artificial intelligence to use biomarker features to stratify ovarian cancer patients.

It is evident that even with the numerous studies that have been done to diagnose ovarian cancer, there is still room for improvement because the accuracy ratings are insufficient. Furthermore, no study has used criteria like . Oxygen consumption markers, samples of blood, and diagnostic tests to separate the various aspects of the data. Thus, data separation is where we have started.

The previous study employed solely statistical methods; however, our data analysis approach combined statistical and machine learning techniques. For the benefit of clinicians and doctors, this approach expanded the scope of work and improved the reliability of real clinical testing. The following are the primary goals of our work:

- Using bio-markers to identify ovarian cancer in its early stages;
- Identify the associative and significant bio-markers by utilising machine learning models and statistical techniques.

II. MATERIALS AND METHODS

In this study, we used a surgically confirmed raw dataset which includes samples from patients struggling from both benign and malignant ovarian cancerous tumours. Next, the most important bio-markers were found by statistical methods, linked to cancer.

Furthermore, early-stage ovarian cancer detection was achieved through the use of machine learning classification models. Figure 1 shows a comprehensive graphic illustration of the workflow.

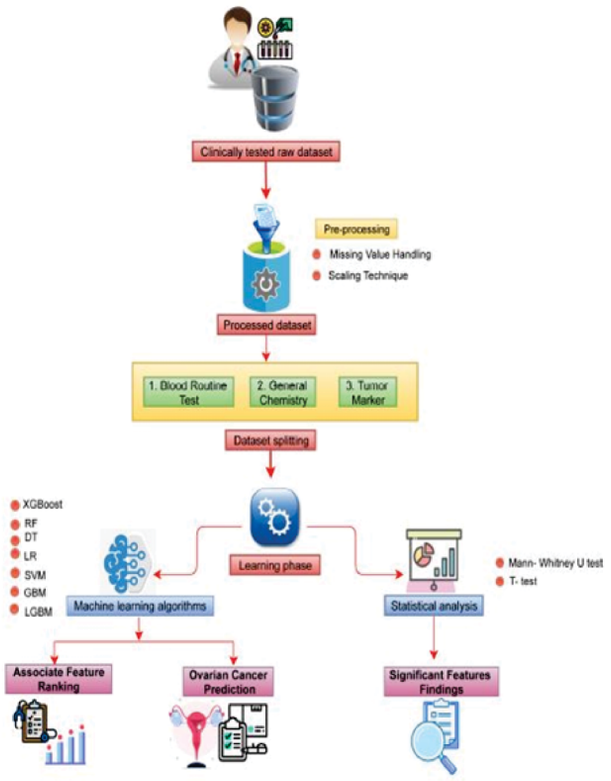

Figure 1 : Proposed System.

A. Data Processing

A number of pre-processing procedures were applied to the raw dataset, including data cleaning, scaling, data division, and missing value imputation. Our dataset contains information on 349 individual patients, and the mean values of each feature's existing values, which were used to impute the missing values, that also accounted for only approximately 7% of the total.

We have “Standardised” the data scaling using equation [13], which centres the values on the mean values and includes a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

Where μ is the mean and σ is the standard deviation.

A. Relationship And Effects Of The Characteristics On The Patients

In this study, patients with ovarian cancer were the case category, and patients with benign ovarian tumours were the control group. The Student's t-test and the Mann-Whitney U-test were then used in two statistical evaluations because they are helpful in determining the critical traits that can be used to distinguish among individuals who have harmless ovarian tumours and those with cancer of the ovary.

The p-values of the significant features were less than 0.05. The relationship between the continuous variable attributes was examined using the Student's t-test, in which features are retained, as they demonstrate a significant correlation (i.e., a p-value < 0.05); if not, they are removed [14]. When comparing two population means, the Mann-Whitney U-test is utilised without assuming that the data came from a standard distribution.

III. MACHINE LEARNING MODEL

The programming language Python (Python 3.7.13) has been utilised for all machine learning investigations. For basic data processing, we have used sklearn. To create all of the plots and figures, the programming languages R and Python's "ggplot2" were also used.

The 'feature importance' function was utilised for the XGB algorithm, the 'coefficient' method was applied for SVM and LR, and the 'feature importance()' function was called for the LGBM algorithm. An ensemble learning method called the Gradient Boosting Machine (GBM) maximises the reduction function [18], usually by utilising a deviant or increasing loss function to combine several weak learners into a robust one.

Adaboosting is used to control, it also handles the exponential loss function; the variance of the function of loss is handled by logistic regression. LGBM is enhanced by Light Gradient Boosting Machine (LGBM), which uses tree-based learning methods. In comparison to other models, it may be able to manage enormous volumes of data and operate at a great degree of correctness with constrained Power for computing, such as memory capacity and processing speed) [20]. The range of the learning was (0.005, 0.01).

Extreme gradient boosting (XGB) reduces loss by joining a new model using a gradient descent technique. A boosting tree provided by XGB offers a quick and accurate solution to a number of data science issues [18]. We completed all of the simulation tasks using the cloud platform Google Co-laboratory.

A. Evaluation Metrics

Using a variety of evaluation metrics, such as accuracy, precision, recall, F-score, AUC, and log-loss, we evaluated the results of the classification algorithms based on the number of True Positives (TP), False Positives (FP), True Negatives (TN), and True Positives (TP) in this study. Correctness: Accuracy is a measure of a model's correctness [21] and is from the confusion matrix, with respect to XGB, 24 samples were correctly classified as cystic, all the 72 samples were correctly classified as polycystic and 83 samples were correctly classified as normal.

But five normal samples were incorrectly classified as cystic and four cystic samples were incorrectly classified as normal. Classification accuracy of the LGBM classifier was 95%. Similarly 84 samples were correctly classified as normal. Also no cystic or polycystic sample was incorrectly classified.

But 4 normal samples were incorrectly classified as cystic contributing to 2.1% of the total data. So 97.9% predictions were correct and 2.1% was wrong thus giving an accuracy of 97.9% depicting a better classification.

IV. RESULTS

In this study only 7% of the 349 patient records in the dataset for this study were missing, and those missing values were filled in by imputing the mean values. After removing any entries with missing values, we were left with the data from 106 patients, of which 44 had benign tumours and 62 had ovarian cancer tumours.

The standard deviation of the mean values was estimated using the data-scaling technique. The entire dataset was split up into 20% for testing and 80% for training. The classifier performance is tested using log-loss evaluation metrics, F1-score, AUC, accuracy, precision, and recall. In order to

identify the significant risk factors for ovarian cancer, we also used the Mann-Whitney U-test.

The Mann-Whitney U-test revealed a significant difference between the two independent groups (U = [U value], p < [p-value]). This indicates that there is a statistically significant distinction in [variable of interest] between [Group 1] and [Group 2]. The output of graph explains the accuracy of the algorithms. The first graph (Fig 4.3) explains Loss function curve for DCCN. It depicts the loss function over time or iterations during the training of a neural network.

The y-axis of the curve represents the loss, while the x-axis typically denotes the training epochs or iterations. At the beginning of training, the loss is usually high as the network makes random predictions. A descending curve indicates decreasing loss, suggesting improved performance in capturing the relationship between input and output in the neural network. Monitoring this curve is crucial for optimizing the network's parameters and achieving better accuracy in DCCN-related tasks.

The second graph (Fig 4.4) illustrates loss function curve for DCCN. The loss function quantifies the difference between the predicted output and the actual target values. As the neural network undergoes training iterations, the loss function value ideally decreases, reflecting improved model accuracy.

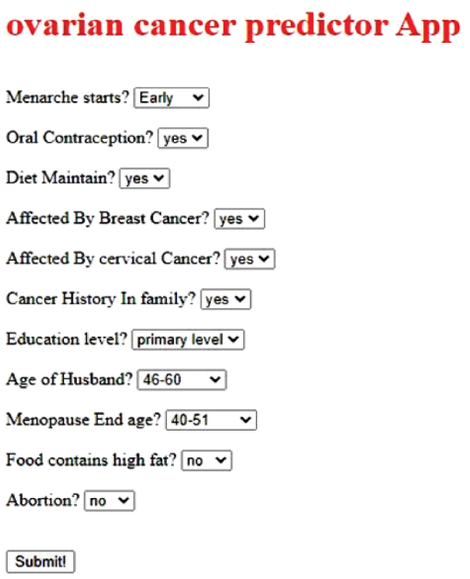

Figure 1 : Output Screen Shot 1

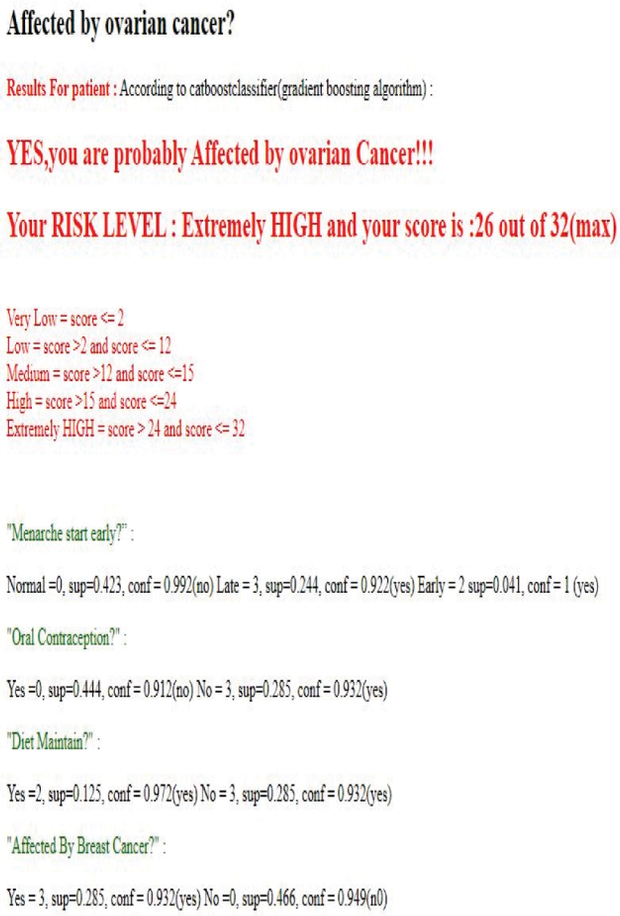

Figure 2 : Output Screen Shot 2

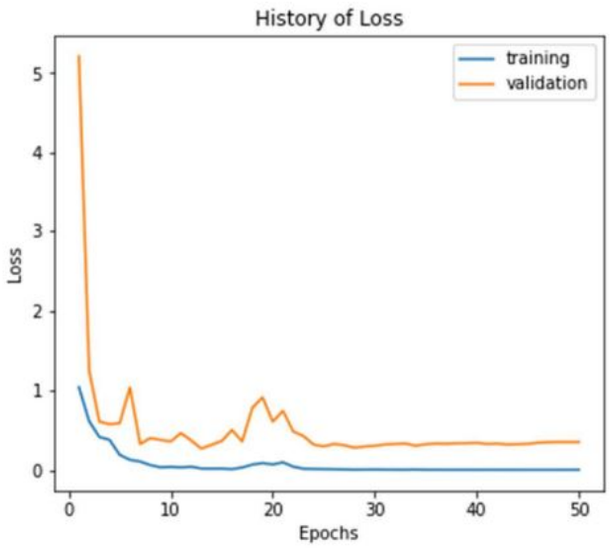

Figure 3 : Loss function curve for DCNN

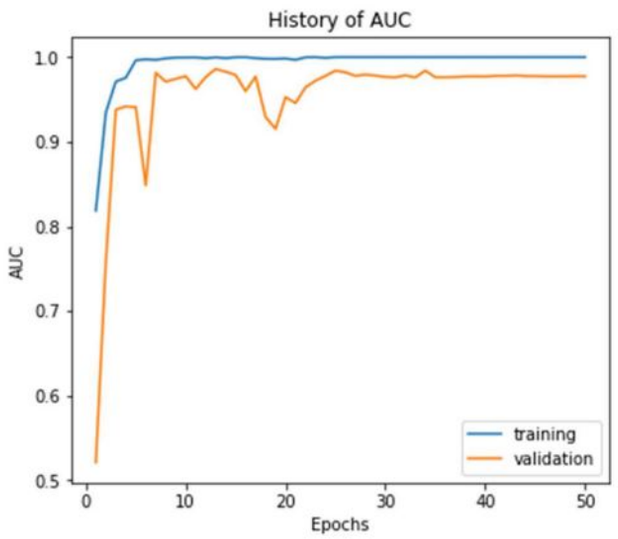

Figure 4 : AUC curve for DCNN

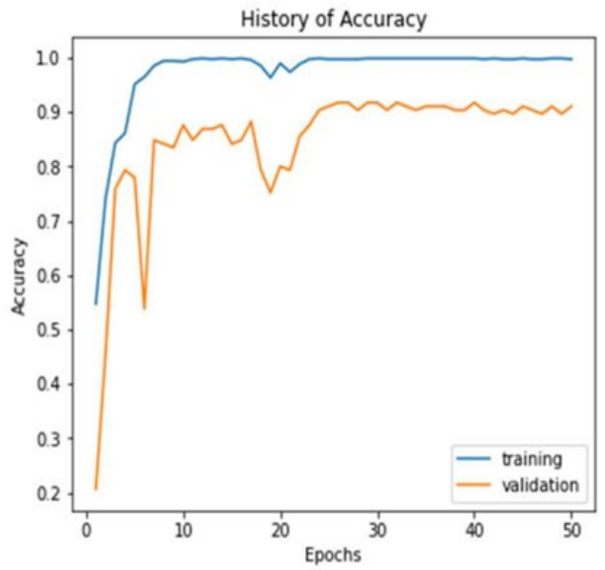

Figure 5 : Accuracy curve for DCNN

V. CONCLUSION

In order to identify key characteristics and to detect ovarian cancer earlier in patients, machine learning and statistical analysis techniques were applied to the dataset. Carcinoem-bryonic antigen, the human epididymis amino acids 4 biomarkers, are the most important ones linked to the cancer in the ovary. Furthermore, a high level of classification accuracy was discovered for LGBM classification algorithms, indicating that our work may find uses in computer-assisted medical diagnosis to aid doctors in less expensive ovarian cancer diagnosis. Our work also has the significant consequence of shortening the time it takes to identify cancer. Our research's primary constraint is the volume of data. We will investigate ovarian cancer using more data in the future, including the patient control group. Accordingly, we think our multi-classification work was crucial in determining the disease's stages and offering patients the necessary medicines to prolong their lives. Combine clinical data with multi-omics data (genomics, transcriptomics, proteomics, etc.) to gain a more comprehensive understanding of ovarian cancer. In future machine learning models will be developed that can effectively integrate and analyze diverse data types to identify robust biomarkers. Investigate advanced feature engineering techniques to extract more informative features from clinical data. Explore deep learning methods for automatic feature extraction to capture intricate patterns in the data. Address the issue of imbalanced datasets, which is common in medical datasets. Explore techniques such as over sampling, under sampling, or generating synthetic samples to improve model performance on minority classes.

REFERENCES

[1] Ashish, K., Manish, K., Yongyeon J., Hongkook, K. and Moongu, J.(2020). Despeckling of medical ultrasound images using Daubechies complex wavelet transform. Signal Processing., 90(2): 428-439.

[2] Balen, A.H., Laven, J.S.E., Tan, S.L. and Didier, D.(2019). Ultrasound assessment of the polycystic ovary: International consensus definitions. Human Reproduction Update.,9(6): 505–514.

[3] Benacerraf, B.R., Abuhamad, A.Z., Bromley, B., Goldstein, S.R., Groszmann, Y., Shipp, T.D. and Timor,T.I.E. (2019). Consider ultrasound first for imaging of the female pelvis. American Journal of Obstetrics and Gynecology., 212(4):450–455.

[4] Chen, Z.J. and Chen, C.H.Y.(2019). Efficient statistical modeling of wavelet coefficients for image denoising. International Journal of Wavelets,

Multiresolution and Information Processing., 7(5):629-641.

[5] Chitalia, R.D. and Kontos, D.(2019). Role of texture analysis in breast MRI as a cancer biomarker: A review. Journal of Magnetic Resonance Imaging.,49(4):927-938.

[6] Dewi,R.M., Adiwijaya., Wisesty,U.N. and Jondri.(2018).Classification of polycystic ovary based on ultrasound images using competitive neural network.Journal of Physics Conference Series.,971-012005:1-8.

[7] Gao, S., Peng, Y., Guo, H., Liu, W., Gao, T., Xu, Y. and Tang, X. (2018). Texture analysis and classification of ultrasound liver images. Bio-medical Materials and Engineering.,24(1):1209-1216.

[8] Goswami, B., Patel, S., Chatterjee, M., Koner, B.C. and Saxena, A.(2019). Correlation of prolactin and thyroid hormone concentration with menstrual patterns in infertile women. Journal of Reproduction and Infertility., 10(3):207–212.

[9] Himani, S. and Sunil, K.(2019). A Survey on decision tree algorithms of classification in data mining. International Journal of Science and Research., 5(4):2094-2097.

[10] Hiremath, P. S. and Tegnoor, J. R.(2020). Automatic detection of follicles in ultrasound images of ovaries by optimal thresholding method. International Journal of Computer Science and Information Technology.,3(2): 217-220.

[11] Verma, C.; Illes, Z.; Stoffova, V.; Bakonyi, V.H. Comparative Study of Technology With Student's Perceptions in Indian and Hungarian Universities for Real-Time: Preliminary Results. IEEE Access 2021, 9, 22824–22843.

[12] Lukanova, A.; Kaaks, R. Endogenous hormones and ovarian cancer: Epidemiology and current hypotheses. Cancer Epidemiol. Biomarkers Prev. 2016, 14, 98–107.

[13] Paik, E.S.; Lee, J.W.; Park, J.Y.; Kim, J.H.; Kim, M.; Kim, T.J.; Choi, C.H.; Kim, B.G.; Bae, D.S.; Seo, S.W. Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. J. Gynecol. Oncol. 2019, 30, e65.

[14] Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2019: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2019, 68, 394–424

[15] Marchetti, C.; Pisano, C.; Facchini, G.; Bruni, G.S.; Magazzino, F.P.; Losito, S.; Pignata, S. First-line treatment of advanced ovarian cancer: Current research and perspectives. Expert Rev. Anticancer Ther. 2019, 10, 47–60.