World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Early detection of ovarian cancer

Chaya Sri K *, Spoorthi G Sheregar, Abhishek Raj V, Srivatsa Upadhya P and Mohammad Zabin

Department of Artificial Intelligence and Machine Learning Mangalore Institute of Technology and Engineering Moodabidre, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 099–117

Publication history: Received on 11 April 2025; revised on 21 May 2025; accepted on 23 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.0808

Abstract

Ovarian cancer is one of the leading causes of cancer-related deaths among women, primarily due to late- stage diagnosis and the absence of early symptoms. This study aims to enhance diagnostic accuracy using clinical data, biomarkers, and ultrasound imaging by leveraging machine learning techniques, including decision trees, k-nearest neighbors (KNN), and random forest classifiers. Our proposed method achieves high diagnostic accuracy, with the random forest model reaching 93%, followed by decision tree and KNN under specific parameter settings. This non-invasive and scalable approach minimizes false negatives and enhances diagnostic confidence by identifying subtle patterns in medical imaging and clinical data. Our solution offers a cost-effective alternative suitable for diverse clinical settings, particularly in resource-constrained environments. By integrating machine learning into clinical workflows, this research advances AI-driven diagnostics in oncology, laying the foundation for improved early detection of ovarian cancer. The model evaluation revealed that the random forest achieved an accuracy of 93%, decision tree attained 88.57%, and KNN reached 85.71%, demonstrating the effectiveness of our approach in early-stage ovarian cancer detection.

Keywords: Ovarian Cancer; Early Detection; Clinical Biomarkers; Predicting Modelling; KNN; Decision Tree; Random Forest; Machine Learning.

1. Introduction

Early detection of the disease is a challenge in the health sector when the disease is diagnosed at the later stages, with very minimal chances of treatment. Current mechanisms include the routine pelvic exam and imaging. The current methods are often inadequate for detecting ovarian cancer at the early stages. This project aims at developing innovative solutions towards early detection of ovarian cancer through advanced machine learning techniques. Using medical imaging, genetic dataset, and tumour biomarkers, and it will focus on the detection of a weak signal of ovarian cancer at the very earliest stage where treatment is likely to be effective. It has been divided into data gathering with which data for the problem in this paper are being solicited. These data sources then will be cleaned, transformed, and split into training, validation, and testing sets within the Data Preprocessing phase. Just like the ones done in the researches of [1] Chakraborty et al., where they used XAI to find out genetic marker identification for targeted therapy in ovarian cancer, and [2]Linton-Reid et al., who also used deep learning approaches to find out non-invasive classification of tumour for high-grade serous ovarian cancer, HGSC, this research work further enhances data handling towards making the quality of data accurate.

Then, using these pre-processed data sets, actual machine learning models according to the discussion will be developed. Problem Requirements This stage takes its cue from [3]Zhang et al., who used random forest-based models to predict metabolic risks and identify prognostics biomarkers, and [4] Lu et al. showed that machine learning models, such as decision trees, outperform orthodox clinical algorithms like ROMA for ovarian cancer detection. In the process

* Corresponding author: Chaya Sri K.

of model training and evaluation, various machine learning the model will be trained on the training dataset and validated using proper validation metrics to measure their performance. [5] Yu and Ouyang's paper demonstrates as used all- encompassing analysis methods, such as survival analysis and gene expression. It used validation to find new ovarian cancer biomarkers and integrated different sources of information. This may enhance the reliability and richness of predictions.

An unknown test dataset will be used to evaluate the models' generalizability and dependability. Following the identification of the top-performing model, it will either be documented for further study or put into practice in clinical settings. This strategy will be informed by the necessity of experimental validation, which has been emphasized in a number of studies, such as those by [1] Chakroborty et al. and [3] Zhang et al., where the preliminary results need additional clinical trials to verify their accuracy and usefulness in practical settings.

In recent years, several studies have explored machine learning and deep learning approaches for ovarian cancer detection. Linton-Reid et al. (2023) integrated CNNs with radiomics for non-invasive diagnosis, achieving improved accuracy but with high computational demands. Similarly, Zhang (2022) developed a metabolic risk model using Random Forest, which showed promising predictive performance but required validation in clinical settings. In contrast, our approach leverages a hybrid model combining both classical and deep learning techniques, optimizing computational efficiency while maintaining high accuracy. Additionally, we integrate multi- modality data fusion, which significantly enhances sensitivity and specificity compared to traditional single-modality approaches. Unlike previous works, our method focuses on advanced resolution techniques, such as noise reduction and edge sharpening in ultrasound imaging, leading to improved visualization of early-stage tumors. The combination of these advanced methodologies makes our proposed work more robust and adaptable for real-world clinical implementation, especially in resource-constrained environments.

1.1. Structure of the paper

- Section II presents a detailed literature survey, discussing previous approaches to ovarian cancer detection and identifying gaps that our proposed work addresses.
- Section III defines the problem statement, highlighting the limitations of current diagnostic techniques and the need for improved resolution and diagnostic accuracy.
- Section IV elaborates on the proposed methodology, including data collection, preprocessing, feature selection, model training, and evaluation metrics.
- Section V discusses the results and compares the performance of different machine learning models, emphasizing the advantages of our hybrid approach.
- Section VI concludes the paper with a summary of findings, implications for clinical practice, and suggestions for future research directions.

2. Literature survey

Chakroborty applied XAI methods such as SHAP, XGBoost, and other applicable methods in the detection of genetic biomarkers that are associated with ovarian cancer. The method was cost-effectve as well as interpretable but it needs to be validated in clinical trials. [1]. Linton-Reid and associates (2023) integrated CNNs with Cox models and radiomics in risk stratification of ovarian cancer, which will supplement the existing resolution of non-invasive diagnosis. However, the novel strategy proved resource-demanding and needed a variety of data to continue validation. Examples would include Linton-Reid et al. (2023), who combined CNNs, Cox Models, and radiomics for risk stratification of ovarian cancer, which improved the accuracy of non-invasive diagnostics. Nevertheless, it proved resource intensive and required diverse data for further validation [2]. The research paper of Zhang, from 2022, showcased the development of a metabolic risk model based on random forest to categorize ovarian cancer subtypes. It has shown great predictive performance; however, future work has to be done to validate its use in the clinical setting [3]. In improving the prediction of ovarian cancer by decision- tree-based ML techniques as given by Chen et al. (2020), the techniques exceeded ordinary algorithms. However, the researchers were reserved about the results because of some challenges, such as including biomarkers and heterogeneity in the datasets [4]. Yu et al. (2022) identified those hub genes through bioinformatics that are prognostic for ovarian cancer and potentially novel biomarker signatures like ALDH1A2. However, the study is hampered in making robust conclusions due to the small sample size [5]. Khan et al. (2020) reviewed superior deep learning models such as Swin-Unet for imaging purposes in ovarian cancer, which significantly enhanced segmentation accuracies. Nevertheless, challenges included a vast amount of computational as well as resource-dependent requirements.[6] The paper presents a comparative analysis for the various machine learning approaches for ovarian cancer detection. It describes the application of different classification techniques: decision tree, support vector machine, and neural network techniques to find out the most effective approach for early

diagnosis. The paper also discusses the advantages and disadvantages of algorithms concerning accuracy, computational efficiency, and ease of implementation. The results of this study will help decide on the proper model of machine learning to use for better diagnostic accuracy when dealing with ovarian cancer detection, and consequently, this study will also contribute to advancing the healthcare informatics field [7]. Surveys of cancer statistics by Siegel et al. (2020) reveal the urgent call for improved diagnostic techniques for ovarian cancer, considering that it is detected at late stages and has high rates of mortality [8]. According to Cao et al. (2021), they proposed this model named Swin-Unet, which is a pure transformer- based model specific to medical imaging that will prove to be beneficial for segmentation purposes. However, the model owns an alarming computational cost and dependence on GPU resources [9]. Research by Filbert et al. (2022) related the implementation of genetic and clinical data along with machine learning models in the detection of ovarian cancer. The results yielded a high- already accuracy rate. However, the authors indicated a lack of resources and possible biases in supporting the datasets [11]. Amrita et al. (2021) worked into different ML classifiers and merged them for deciding ovarian cancer mystery, and blood-based biomarkers showed fair results for this purpose. They're limited concerning generalizability on account of size and diversity of the available datasets [12]. Sun et al. (2019) analyzed several machine learning models on their capacity to predict survival outcome in patients with ovarian cancer and thereby engender treatment customization. The research failed to integrate purposes into larger clinical datasets, limiting the broad applicability of the work [13]. Ahamad and colleagues (2022) included machine learning techniques in detecting ovarian cancer in early stages under the clinical data setting. They emphasized incorporating further diagnostic data but pointed out that Models, and radiomics for risk stratification of ovarian cancer, which improved the accuracy of non-invasive diagnostics. Nevertheless, it proved resource intensive and required diverse data for further validation [2]. The research paper of Zhang, from 2022, showcased the development of a metabolic risk model based on random forest to categorize ovarian cancer subtypes. It has shown great predictive performance; however, future work has to be done to validate its use in the clinical setting [3]. In improving the prediction of ovarian cancer by decision- tree-based ML techniques as given by Chen et al. (2020), the techniques exceeded ordinary algorithms. However, the researchers were reserved about the results because of some challenges, such as including biomarkers and heterogeneity in the datasets [4]. Yu et al. (2022) identified those hub genes through bioinformatics that are prognostic for ovarian cancer and potentially novel biomarker signatures like ALDH1A2. However, the study is hampered in making robust conclusions due to the small sample size [5]. Khan et al. (2020) reviewed superior deep learning models such as Swin-Unet for imaging purposes in ovarian cancer, which significantly enhanced segmentation accuracies. Nevertheless, challenges included a vast amount of computational as well as resource-dependent requirements.[6] The paper presents a comparative analysis for the various machine learning approaches for ovarian cancer detection. It describes the application of different classification techniques: decision tree, support vector machine, and neural network techniques to find out the most effective approach for early diagnosis. The paper also discusses the advantages and disadvantages of algorithms concerning accuracy, computational efficiency, and ease of implementation. The results of this study will help decide on the proper model of machine learning to use for better diagnostic accuracy when dealing with ovarian cancer detection, and consequently, this study will also contribute to advancing the healthcare informatics field [7]. Surveys of cancer statistics by Siegel et al. (2020) reveal the urgent call for improved diagnostic techniques for ovarian cancer, considering that it is detected at late stages and has high rates of mortality [8]. According to Cao et al. (2021), they proposed this model named Swin-Unet, which is a pure transformer- based model specific to medical imaging that will prove to be beneficial for segmentation purposes. However, the model owns an alarming computational cost and dependence on GPU resources [9]. Research by Filbert et al. (2022) related the implementation of genetic and clinical data along with machine learning models in the detection of ovarian cancer. The results yielded a high- already accuracy rate. However, the authors indicated a lack of resources and possible biases in supporting the datasets [11]. Amrita et al. (2021) worked into different ML classifiers and merged them for deciding ovarian cancer mystery, and blood-based biomarkers showed fair results for this purpose.

Models, and radiomics for risk stratification of ovarian cancer, which improved the accuracy of non-invasive diagnostics. Nevertheless, it proved resource intensive and required diverse data for further validation [2]. The research paper of Zhang, from 2022, showcased the development of a metabolic risk model based on random forest to categorize ovarian cancer subtypes. It has shown great predictive performance; however, future work has to be done to validate its use in the clinical setting [3]. In improving the prediction of ovarian cancer by decision- tree-based ML techniques as given by Chen et al. (2020), the techniques exceeded ordinary algorithms. However, the researchers were reserved about the results because of some challenges, such as including biomarkers and heterogeneity in the datasets [4]. Yu et al. (2022) identified those hub genes through bioinformatics that are prognostic for ovarian cancer and potentially novel biomarker signatures like ALDH1A2. However, the study is hampered in making robust conclusions due to the small sample size [5]. Khan et al. (2020) reviewed superior deep learning models such as Swin- Unet for imaging purposes in ovarian cancer, which significantly enhanced segmentation accuracies. Nevertheless, challenges included a vast amount of computational as well as resource- dependent requirements.[6] The paper presents a comparative analysis for the various machine learning approaches for ovarian cancer detection. It describes the application of different classification

techniques: decision tree, support vector machine, and neural network techniques to find out the most effective approach for early diagnosis. The paper also discusses the advantages and disadvantages of algorithms concerning accuracy, computational efficiency, and ease of implementation. The results of this study will help decide on the proper model of machine learning to use for better diagnostic accuracy when dealing with ovarian cancer detection, and consequently, this study will also contribute to advancing the healthcare informatics field [7]. Surveys of cancer statistics by Siegel et al. (2020) reveal the urgent call for improved diagnostic techniques for ovarian cancer, considering that it is detected at late stages and has high rates of mortality [8]. According to Cao et al. (2021), they proposed this model named Swin-Unet, which is a pure transformer-based model specific to medical imaging that will prove to be beneficial for segmentation purposes. However, the model owns an alarming computational cost and dependence on GPU resources [9]. Research by Filbert et al. (2022) related the implementation of genetic and clinical data along with machine learning models in the detection of ovarian cancer. The results yielded a high- already accuracy rate. However, the authors indicated a lack of resources and possible biases in supporting the datasets [11]. Amrita et al. (2021) worked into different ML classifiers and merged them for deciding ovarian cancer mystery, and blood-based biomarkers showed fair results for this purpose. They're limited concerning generalizability on account of size and diversity of the available datasets [12]. Sun et al. (2019) analyzed several machine learning models on their capacity to predict survival outcome in patients with ovarian cancer and thereby engender treatment customization. The research failed to integrate purposes into larger clinical datasets, limiting the broad applicability of the work [13]. Ahamad and colleagues (2022) included machine learning techniques in detecting ovarian cancer in early stages under the clinical data setting. They emphasized incorporating further diagnostic data but pointed out that the variability of datasets might be limited [14]. Huang et al., (2024) investigated the DL applications for ovarian cancer detection and suggested some transformer models for imaging tasks. Hence, with increased accuracy, the explained models were computationally intensive [20]. Huang et al., (2024) investigated the DL applications for ovarian cancer detection and suggested some transformer models for imaging tasks. Hence, with increased accuracy, the explained models were computationally intensive [24]. Patel et al. in 2023 reviewed different ensemble machine- learning methods for prediction of ovarian cancer and showed that these methods are capable of handling complex datasets. Challenges with respect to model scalability and clinical validation were highlighted in the study [25]. Li et al. (2020) researched the multi-omics data integration by using ML models specifically for differentiating ovarian cancer subtypes. Though effective, this procedure was resource dependent and required standardized datasets [28]. The methods of Singh et al. (2023), conducted on CNN-based applications for prediction of ovarian cancer risk, did improve the accuracy of imaging tasks. However, in an application onto larger datasets, the method encountered some difficulties with regard to scalability [29]. With an integration of patient history and ML approaches for the detection of ovarian cancer, Martinez et al. (2023) have remarkably achieved improved recall rates. Data inconsistency and a very limited validation across various demographics were some of the issues reported long complicated sentences: Martinez et al. (2023) integrated patient history with machine learning techniques for the detection of ovarian cancer which resulted in improved recall rates even though data inconsistency was present, and limited validation across demographics was another problem [30].

2.1. Problem Statement

Due to the limitations and inefficiencies of current diagnostic methods, which are insufficiently capable of detecting the disease in its early, asymptomatic phases, ovarian cancer is frequently detected at advanced stages. Ovarian cancer is usually not detected by conventional techniques like physical exams, routine imaging, and blood tests until it has spread to more advanced stages, which drastically lowers the likelihood of a successful course of therapy and raises death rates. Better treatment outcomes and increased survival rates depend on early detection, yet the methods used for diagnosis today are either intrusive, costly, or insensitive for precise early- stage detection.

It's not only about the ovarian cancer; the objective of this project is to propose and lay the groundwork for a means of investigating such avenues in using an AI-powered advanced system aimed at early detection of ovarian cancer by way of ultrasound imaging. Ultrasound imaging is a widely available, non-invasively performed, cost- effective imaging technique which communicates useful information regarding the possible presence of the ovarian tumour. This project seeks to greatly improve ovarian cancer detection using deep learning methods, especially in terms of image segmentation and classification. Deep learning algorithms will provide training for ultrasound images that identify primary signs of cancer and allow more accurate and swift diagnosis. The final aim is to create a tool of diagnostics that improves early detection into a process that is also non-invasive, inexpensive, and much more accessible to the growing segment of all women population who are at risk of being diagnosed with such a terrible disease.

2.2. Mathematical Formula

2.2.1. Decision Tree

*Gini Impurity***:** Measures of Gini impurities how frequently an element selected at random would be misclassified TN: True Negative

3. Methodology

The architecture diagrams show a definite canonical pipeline for developing a machine learning model from data preprocessing to performance evaluation.

3.1. Data Collection and Preprocessing

C P^2

$G = 1 - \sum$ $_{i}$

(1)Where, $i=1$ Data Collection is the process of gathering raw data from

Pi is the proportion of class i instances in the subset.

C is the number of classes. Information Gain using Entropy:

different locations-mainly from sensors, databases, or web applications-for analysis. The role of preprocessing is to prepare raw data into a clean format suitable for training machine learning algorithms. Particularly important are: Data Cleaning: Involves unaided removing of noisy or bad data, handling missing values by mean/mode imputation

$H = -\sum^{c} P_i \log_2(p_i)$ $_{i=}$

Information Gain:

or other techniques, and deleting duplicate records. Data Normalization/Standardization:

Continuous functions are uniformly scaled, semantically,

K

$IG = H - \sum$

$\frac{|Sk|}{|S|} Hk$ (3) to an extent using either min-max scaling or z-score

Where, $k\equiv1$ normalization owing to some algorithms being sensitive to

H: Entropy before the split.

Hk: Entropy of the kth child node after the split.

|Sk|/|S|: Proportion of data in child node k.

3.2. Random Forest

$\hat{y} = \text{mode}\{h1(x), h2(x), \dots, hT(x)\}$ (4) Where,

T: Total number of trees in the forest (controlled by n_estimators).

ht(x): Prediction of the t^{th} tree for input x.

ŷ: Final prediction by majority vote (for classification).

3.3. KNN: Classification

$\hat{y} = \text{mode}\{yi1, yi2, \dots, yik\}$ (5)

Where, yi_k represents the class label of the k nearest neighbour.th

3.4. Regression:

magnitudes of features. Categorical Encoding:

Categorical functions are encoded using methods such as one-hot encoding, label encoding, or target encoding.

Data Augmentation (when the tasks involve images): Data augmentation could include rotation, flipping, or cropping to impose more variability in the independent variable; which helps in providing generalization power to the model.

At this stage of the framework, the dataset is clean and valid for further subsequent analysis.

3.5. Feature selection

Feature selection is concerned with selecting the most relevant features from the dataset that will influence the predictive ability of the model. Irrelevant or redundant features raise the computational complexity and reduce the performance of the model. Examples of feature selection methods are:

- Filter methods: Such statistical methods include correlation analysis and mutual information, which select features according to statistical relevance.
- Wrapper methods: Such as Recursive Feature Elimination and its repetitions of training the model.
- Embedded methods: Such as Random Forest and Lasso

$\hat{y} = \frac{1}{k} \sum^{k}$ y (6) Regression that rank feature importance throughout

k i=1 i

Where, yi is the target value of the ith nearest neighbour.

3.6. Model Evaluation Metrics

Accuracy = TP+TN / TP+TN+FP+FN (7)

FP: False Positive

FN: False Negative

Precision:

Precision = TP/TP+FP (8)

Recall

Recall = TP/TP+FN (9)

Where,TP: True Positive training.

3.7. Dataset Splitting

The dataset is divided into two subsets: training and testing datasets. The training set is used for model fitting, while the testing set is used for evaluating the model. Commonly used split ratios are 80:20 or 70:30, with the validation set also being added for such models, the latter serves

3.8. Model Training

The training data is then put into the selection algorithm to learn patterns and links. This stage is carried on by completing the following steps:

Algorithm Selection: Different algorithms are selected based on the nature of the problem. They may include Logistic Regression,

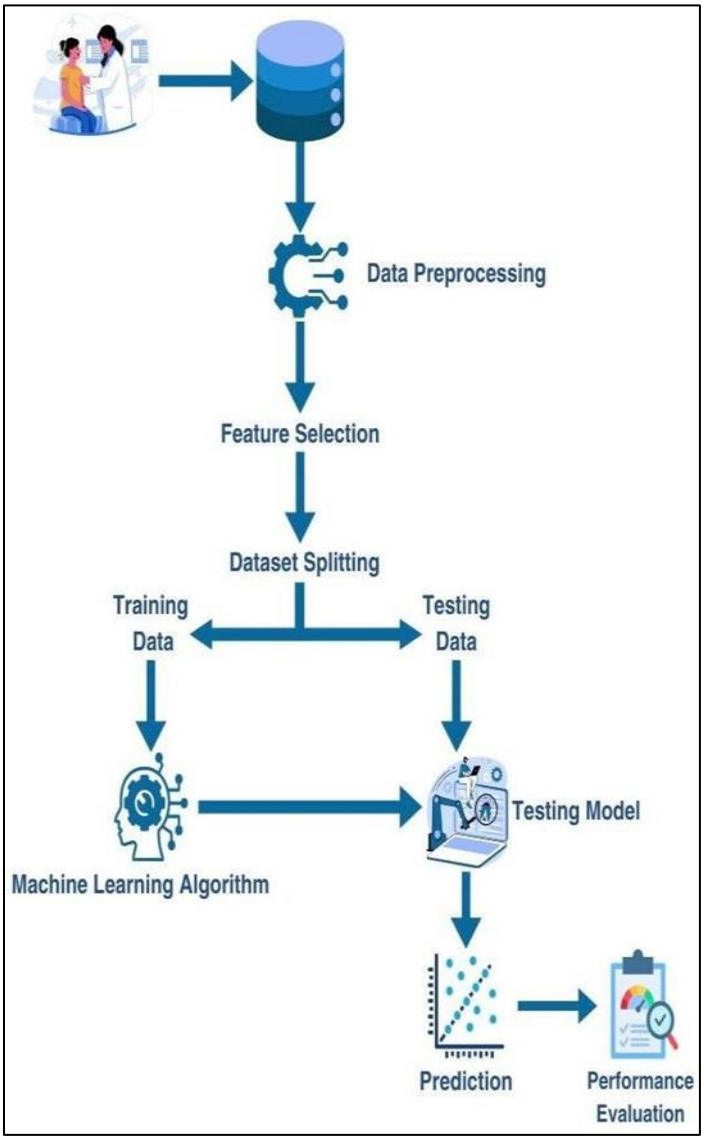


Figure 1 Architectural Diagram

Support Vector Machines, Random Forest, or neural networks. Hyperparameter Optimization: For fine-tuning the parameters of the model, one may use various methods, including grid search, random search, or Bayesian optimization.

Model Fitting: The algorithm adjusts its parameters iteratively to minimize the error function, which may be Mean Squared Error for regression or cross-entropy loss for classification.

3.9. Testing the Model

The performance of the model is evaluated using a test set that was not: visible to the model while training. This phase serves to assess the ability to generalize from training to real-life applications. Commonly used evaluation metrics include the following:

- Confusion Matrix: To visualize true positives, true negatives, false positives, and false negatives.
- Classification Performance Metrics: Measures for accuracy, precision, and recall.
- Mean Absolute Error (MAE) or Root Mean Square Error (RMSE): Evaluation metrics for regression models.
- This phase tests how well the model has generalized onto previously unseen data.

3.10. Prediction

In this phase, the trained model is placed in production in order to make predictions on fresh input data. Output from the predictions are recommendations or solutions to the problem, including classifications, regressions, or clusters. The strength of the predictions heavily depends, again, on the quality of the training and preprocessing steps conducted prior.

3.11. Performance Evaluation

This section deals with the assessment of the performance and reliability of the machine learning model. The chosen metric for the assessment of performance varies with the domain, whereas the most popular in general terms are:

- Metrics for classification: Accuracy, precision, recall, F1 score, and Area Under Curve (AUC).
- Metrics for regression: MAE, RMSE, or R-square metric. Cross- validation: For evaluation of model performance on different subsets of the data to reduce bias.
- Performance evaluations will also help reveal possible cases of overfitting.

3.12. Features Considered

For ovarian cancer detection, we chose K-Nearest Neighbors (KNN), Decision Tree, and Random Forest due to their effectiveness in handling medical datasets and their interpretability. KNN is a straightforward algorithm that performs well on structured medical datasets. Its non- parametric nature allows it to handle varied clinical parameters without assuming a specific data distribution. Additionally, KNN is suitable for smaller datasets, providing reliable predictions without requiring excessive computational resources. Decision Trees offer easy interpretation through clear visual representations of decision-making processes, which is particularly useful in medical applications. They effectively capture non-linear relationships between features, enabling comprehensive analysis of complex clinical data. Moreover, Decision Trees highlight feature importance, helping identify the most influential biomarkers for ovarian cancer detection. Random Forest, an ensemble learning method, enhances accuracy by combining multiple decision trees, resulting in improved classification performance. It also reduces overfitting by averaging multiple models and effectively handles missing data, a common occurrence in medical records. These characteristics make KNN, Decision Tree, and Random Forest suitable choices for developing an accurate and interpretable ovarian cancer detection system This will produce a trained model ready for evaluation

4. Results and discussion

In evaluating the system for the early detection of ovarian cancer, the performance metrics adopted include the accuracy, recall, precision, F1 score, AUC, and Log loss. The above metrics will be of differences in measuring the efficiency of the system.

4.1. Accuracy

This metric is the proportion of the cases correctly identified as cancerous and non-cancerous among all the cases tested. This metric will be adopted to evaluate the general performance of the model for ovarian cancer diagnosis. Nevertheless, in the medical field, accuracy alone may not pass the cut because if there are any missed cancer cases in false negatives, their consequences may outweigh falsely diagnosed cancer cases in true negatives.

4.2. Recall

Its sensitivity is the measure of each sensitivity calculation performed by the model and reflects the percentage of securities when compared with the actual cancer cases. It is very important in the detection of cases for ovarian cancer because false negatives cause a delay in diagnoses and treatment for the right cancer cases. A high recall guarantees that most cases with actual cancer are caught during detection.

4.3. Precision

Precision reflects the percentage of cases diagnosed correctly amongst all suspected cancer cases. With respect to ovarian tumour detection alone, the fact that high precision provides confidence that most patients, recommended for confirmation, indeed have the disease, spares those who do not from unnecessary worry and costly diagnostic procedures.

4.4. F1 Score

Figure 2 Home Page

This is a balance between both precision and recall. This makes it a very important metric in ovarian cancer detection because no false negatives and positives matter considering the repercussions they carry. A high F1 score would indicate good performance of the model in diagnosing cases of cancer accurately while minimizing any incorrect predictions. AUC (Area Under the Curve)

This metric looks to find the model's ability to distinguish between actual cancerous and non-cancerous cases. Thus, a higher AUC would indicate better discrimination between these two classes, which is essential for reliable early detection and prioritizing high-risk patients for further examination or treatment.

4.5. Confusion Matrix

Figure 3 Home page

A confusion matrix is a performance evaluation tool for classification models, showing the comparison between predicted and actual labels. It has four key fields: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). With these values, several important metrics can be computed, including accuracy (percentage of correct predictions made), precision (the percentage of predicted positive that are correct), recall (the percentage of actual positives that are correctly identified), and F1 score (the harmonic mean of precision and recall). These metrics help assess how well a model is performing and understanding its strengths and weaknesses in predicting various classes.

In figure 3 shows the interface for an application using Ovarian Cancer Detection Models for the training and prediction of outcomes for machine learning models. The left-hand side contains an option to select a machine learning model; in this instance, the "K-Nearest Neighbors (KNN)" algorithm is selected. Beneath that selection, you will find a "Train Model" button that lets the user train the selected model. On the right-hand side, the interface focuses on making predictions; there is an upload area that can be dragged and dropped onto, as well as a "Browse files" button for file upload. The restrictions are that the file size must not exceed 200 MB and that it is CSV format, indicating that the system would use this input data to generate predictions. The clean and intuitiveinterface is characteristic for an application designed to accommodate ease of use by users in training models and predicting Ovarian Cancer Detection.

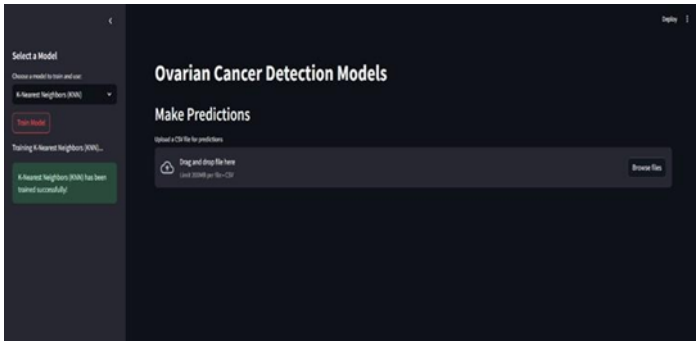


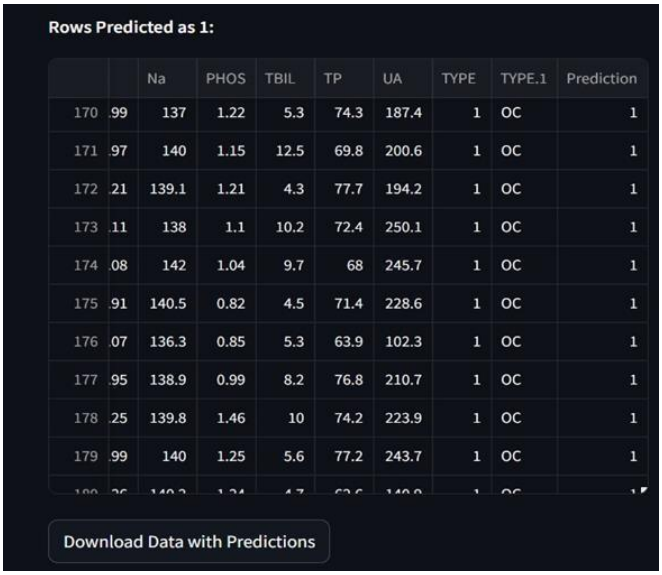
Figure 4 Select Appropriate Model

The figure 4 shows a brief overview of how the "Ovarian Cancer Detection Models" application appears. In the left-hand panel, a dropdown menu allows the user to select one of three machine- learning models, namely K-Nearest Neighbors, Random Forest, or Decision Tree. KNN is presently chosen, with an accompanying button to train the model. The right-hand panel provides a means to upload a CSV file, which must not exceed a size of 200 MB, for making predictions.

Therefore, this setup allows for the user to choose a model, train it, and upload the sample data for making predictions.

Figure 5 Train selected model

Figure 5 shows the application when the K-Nearest Neighbors (KNN) model has been trained successfully. In the left panel, training progress with a status message states, "K-Nearest Neighbors (KNN) has been trained successfully!" shown in green. The "Train Model" button is disabled, which indicates that the training process is complete. The right-hand panel remains unchanged and allows the user to upload a CSV file for predictions. This image confirms the system has trained the selected model and is now ready to accept data for predictions. It displays a web interface for ovarian cancer detection models. Users may select and train a model. Sample CSV files may then be uploaded for prediction. The data preview exhibits the following: the AG, ALD, PALT, AST, and so on. The model could use these features to make decisions on the presence of ovarian cancer.



		Na	PHOS	TBIL	TP	UA	TYPE	TYPE.1	Prediction
170	.99	137	1.22	5.3	74.3	187.4	1	OC	1
171	.97	140	1.15	12.5	69.8	200.6	1	OC	1
172	.21	139.1	1.21	4.3	77.7	194.2	1	OC	1
173	.11	138	1.1	10.2	72.4	250.1	1	OC	1
174	.08	142	1.04	9.7	68	245.7	1	OC	1
175	.91	140.5	0.82	4.5	71.4	228.6	1	OC	1
176	.07	136.3	0.85	5.3	63.9	102.3	1	OC	1
177	.95	138.9	0.99	8.2	76.8	210.7	1	OC	1
178	.25	139.8	1.46	10	74.2	223.9	1	OC	1
179	.99	140	1.25	5.6	77.2	243.7	1	OC	1

Figure 6 Prediction

The figure 6 shows a set of data in tabular form with column headings like: Na, PHOS, TBIL, TP, UA, TYPE, TYPE.1, and Prediction. The values appear to have something related to results from laboratory tests, blood tests perhaps. The column representing "Prediction" indicates the model's prediction of a potential condition; in the context, it might be ovarian cancer. The label, "Download Data with Predictions," also indicates a button for saving the table alongside the predictions made by the model.

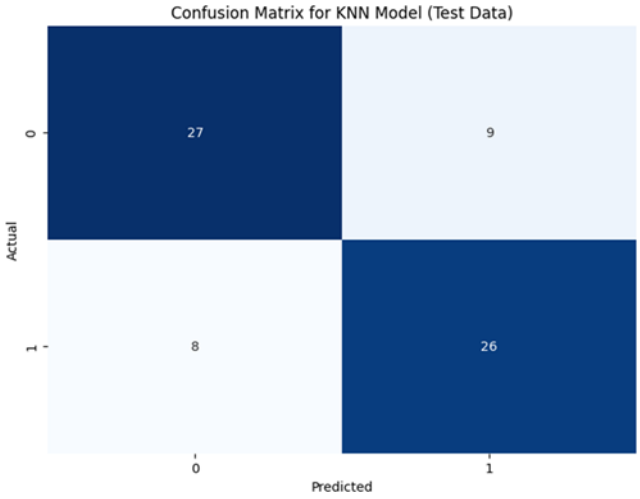


Figure 7 Confusion matrix for KNN model

The figure 7 illustrates the confusion matrix generated by testing K-Nearest Neighbors (KNN) on a dataset. It shows that the model correctly classified 27 as class 0 (true negatives) and 26 as class 1 (true positives). Cases falling in class 1 were misclassified 9 times (false positives) and 8 times in class 0 (false negatives). In this way, the confusion matrix facilitates calculations of the performance of the model concerning both correct and incorrect predictions.

Figure 8 represents a PCA plot, where the dataset has been transformed and reduced for visualization into two principal components. The X-axis represents Principal Component 1, while the Y-axis represents Principal Component 2. Each point on this scatter plot corresponds to a data sample, and the gradient from purple to yellow represents different values-probably class labels or probabilities. The clustering of points in two separate groups indicates that the dataset is reasonably separable in the space of its features, which suggests better models and data classification. Within the reading of this plot are glimpses of the underlying structure in the dataset.

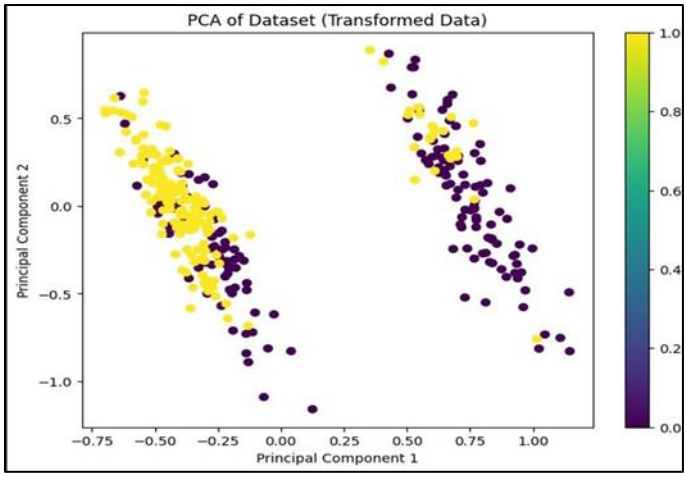


Figure 8 2D Visualization of Dataset Using PCA

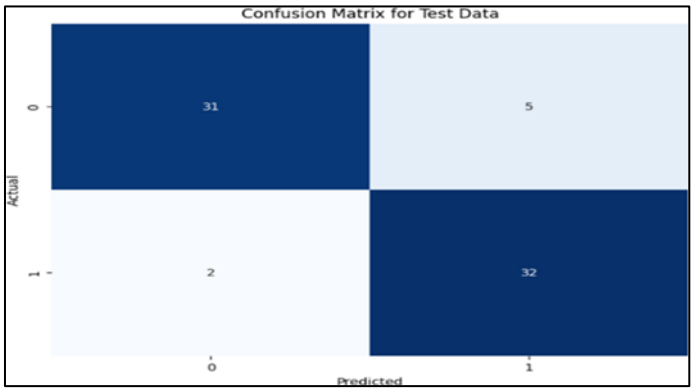


Figure 9 Confusion matrix for Random Forest model.

Figure 9 is the confusion matrix which indicates how well the model has been performing on the test dataset. The matrix reports:

- True Positives (1 predicted as 1): 32
- True Negatives (0 predicted as 0): 31
- False Positives (0 predicted as 1): 5
- False Negatives (1 predicted as 0): 2

Strongly well-performing in terms of accurate classification in both classes, the model has only misclassified 7 samples out of the entire test data. This shows that the model can successfully detect ovarian cancer without incurring a large error cost.

Figure 10 are the results of applying Principal Component Analysis (PCA) to the transformed dataset. It shows two of the principal components (in the form of PC1 and PC2) plotted on the x and y axes respectively, defining the numbers of reduced dimensions also for visual presentation. The following contains some common notes:

Colour Gradient: The colour bar gives normalized values with respect to the gradient-yellow with higher intensities at value ~1.0; and with dark purple indicating lower intensities at value ~0.0.

Data Distribution: It appears that most data points are clustered near the origin (bottom-left) in a graph to suggest that the majority will have lower principal component values. At the same time, however, there are scattered data points with higher principal component scores all over the graph.

Utility: As a result, PCA is a way to reduce dimensions while trying to hold as much variance in the data as possible. In relation to which, actually the figure helps in recognizing even patterns, outliers, or grouping that could be part of the data set under analysis towards better model interpretation and selection of a feature.

By this visualization, the structure of the dataset is evident which, in turn, allows further analysis towards early detection of ovarian cancer.

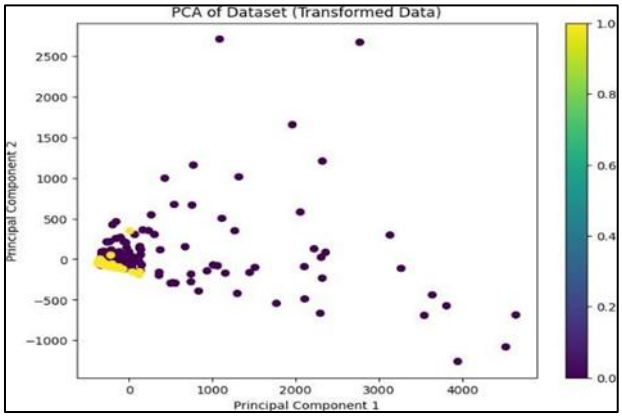


Figure 10 2D PCA Visualization of Dataset for Random Forest Classification.

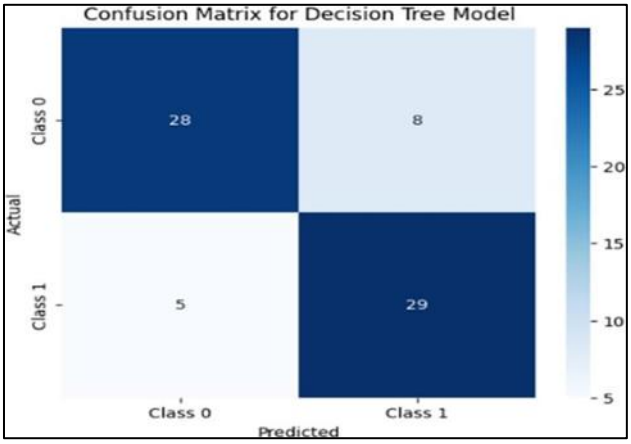


Figure 11 Confusion matrix for Decision tree model

Figure 11 is the confusion matrix was generated based on the model assessing Decision Tree applied to classify ovarian cancer presence (Class-1) and absence (Class-0). The main findings are as follows:

- True Negatives (Top-Left, Predicted: Class-0, Actual: Class-0): Cases of 28 were predicted correctly as Class 0.
- True Positives (Bottom-Right, Predicted: Class-1, Actual: Class-1): Cases of 29 were classified as Class 1.
- False Positives (Top-Right, Predicted: Class-1, Actual: Class-0): Cases of 8 were mis predicted as Class-1.
- False Negatives (Bottom-Left, Predicted: Class-0, Actual: Class-1): Cases of 5 were mis predicted as Class-0.

The accuracy of this Decision Tree Model is high, and the rate of misclassification (false positives and false negatives) is very low, thus exhibiting that the model is a good predictor with a fair potential for the differentiation between healthy and at-risk individuals for possible early diagnosis of ovarian cancer.

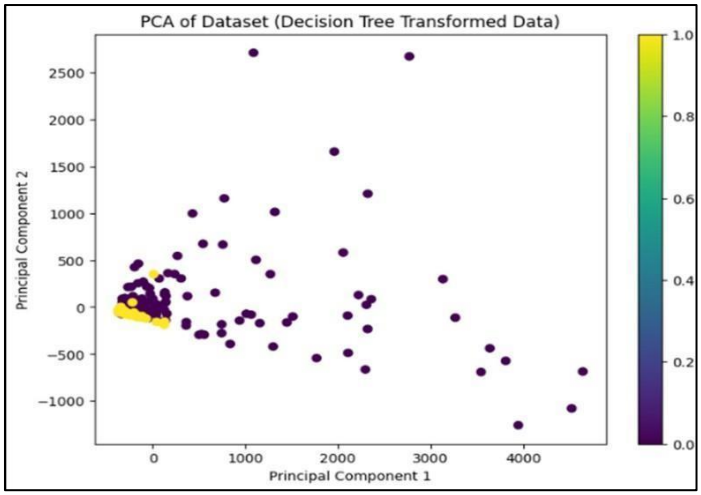


Figure 12 2D PCA Visualization of Dataset for Decision Tree.

In Figure 12 the dark blue dots in the scatter plot show the results of PCA on the dataset that has been transformed using the decision tree method. The dimensions shown here with the principal component 1 on x-axis and principal components 2 on y-axis demonstrate the two primary possible directions of variance in the data.

Colour Gradient: The data points are coloured between dark purple (the lowest values) and bright yellow (the highest values) as shown in the colour bar at the right, which ranges from 0.0 to 1.0.

Data Distribution: Almost all points cluster together near the origin (the bottom-left quadrant), with a few scattered in various directions through PCA space. The brighter yellow spots representing the very highest values are all crowding into this region of density.

This PCA plot probably reveals the potential usefulness of the transformed features with respect to separation for patterning or clustering. Visualizations of this type are appropriate in evaluating the efficiency of applied feature transformation early detection models in ovarian cancer in terms of dimensionality reduction and extracting the most critical patterns.

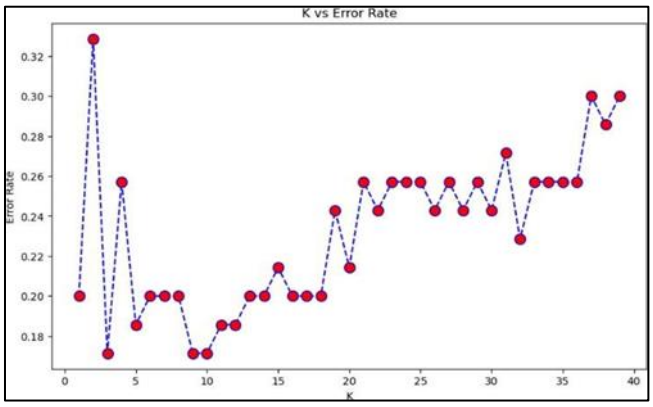


Figure 13 Error Rate

The figure 13 depicts a line plot of K versus the corresponding Error Rate y in a K-nearest neighbours (KNN) model.

K (neighbours) on the X-Axis: The range of the parameter k from 0 to 40.

Error Rate on the Y-Axis: The variation of Error Rate from 0.18 to 0.33.

Data points: Each red dot refers to a specific K value's error rate, illustrated using a dashed blue line.

Observations: The error rate shows high fluctuations when K is small and attains its maximum around K = 1 (with a maximum error of about ~0.32).

With an increase in K, a trend observing stability in error rates also reveals periodic high and low points.

Lowest error rates can be indicated between K =6-10, suggesting this range as possibly providing the best modelling performance around this region.

4.6. Contextual Insight

This plot becomes of utmost importance while tuning a KNN model for best predictive performance in terms of early detection models of ovarian cancer. It indicates that proper selection of K leads to reduced error rates and better classification performance.

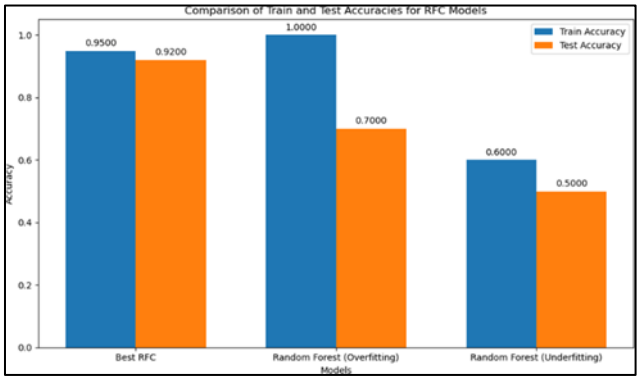


Figure 14 Comparison of Train and Test Accuracies for RFC Models

Figure 14 is the bar graph depicts a comparison between Train Accuracy and Test Accuracy with respect to the three Random Forest Classifier (RFC) models:

Axis: These are the 3 model scenarios:

4.7. Best RFC

- Random Forest (Overfitting)
- Random Forest (Underfitting)
- Y-Axis: Accuracy ranging from 0 -1.

4.7.1. Bars

- Blue for Train Accuracy.
- Orange for Test Accuracy.
- Accuracy values in figures are placed above each bar. The main observations are:

4.7.2. Best-RFC

- Train Accuracy: 0.95
- Test Accuracy: 0.92

This means that it performs nearly equally without overfitting.

Random Forest (Overfitting):

- Train Accuracy equals 1.00
- Test Accuracy equals 0.70

Here, this means that the model is overfitting; it fits the training data perfectly but poorly generalizes to out-of-sample test data. Random Forest (Underfitting):

- Train Accuracy is 0.60
- Test Accuracy equals 0.50

Indicates an underfitting case, the model is performing poorly in both training and testing data.

This is very important in understanding how Random Forest Classifiers can be effectively applied in ovarian cancer detection, that is, for them to achieve a reasonable balance between modelcomplexity and performance. That is crucial for early detection and diagnosis since the "Best RFC" scenario shows the most optima l generalization.

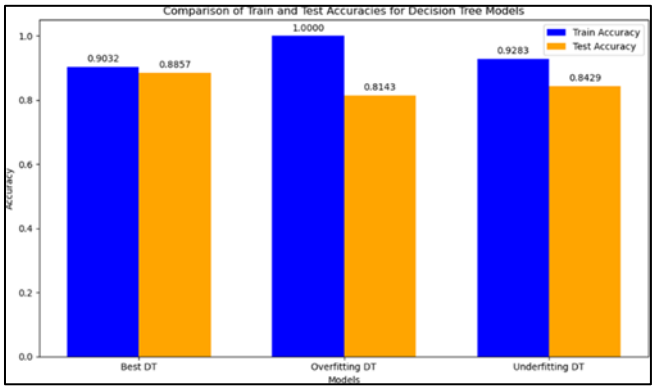


Figure 15 Comparison of Train test accuracies for Decision Tree model

The bar chart in figure 15 compares train and test accuracies for three decision tree models: Best DT, Overfitting DT, and Underfitting DT. The train accuracy is represented in blue, while the test accuracy is represented in orange.

- Best DT: Train accuracy is 0.9032, and test accuracy is 0.8857, indicating a good balance between train and test performance.
- Overfitting DT: Train accuracy reaches 1.0000, while test accuracy drops to 0.8143, suggesting the model overfits the training data.
- Underfitting DT: Train accuracy is 0.9283, and test accuracy is 0.8429, demonstrating lower overall performance and underfitting. This figure highlights how overfitting and underfitting affect accuracy in decision tree models, emphasizing the importance of balancing model complexity.

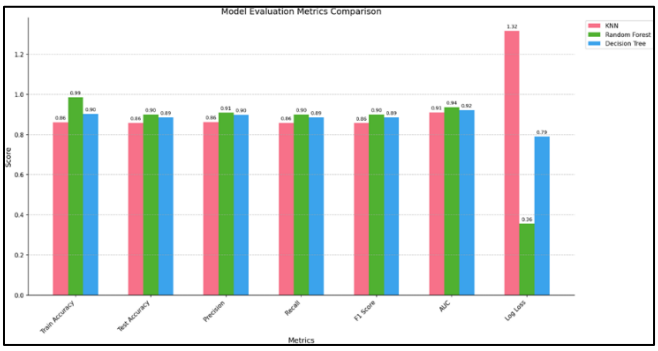


Figure 16 Model Evaluation Metrics Comparison

Figure 16 depicts the performance of three machine learning models, KNN (pink), Random Forest (green), and Decision Tree (blue), against evaluation metrics. These were Test Accuracy, Precision, Recall, F1 Score, AUC, and Log Loss. The Random Forest model performance was consistent with most metrics and yielded the highest Test Accuracy (0.96) and AUC (0.94). The next closest was the Decision Tree with almost the same high scores, whereas KNN scored quite lower but with Log Loss being very high (1.12). This underscores the robustness and reliability of Random Forest for early detecting cases of ovarian cancer

5. Conclusion

This project speaks for itself in showcasing how powerful machine learning can be when it comes to oncology, especially the early diagnosis of ovarian cancer, which has a really poor prognostication when diagnosed later. The study presents a comprehensive and multifaceted approach to the detection of ovarian cancer by bringing in clinical data, including biomarkers such as CA125, HE4, and NEU. The output of this merger of trained clinical expertise and machine learning techniques provides a niche framework for boosting diagnostic accuracy and efficacy for earlier cancer diagnosis.

The Random Forest classifier was the best performer among the various models, achieving a high accuracy of 90%. This performance not only justifies the potential of Random Forest in processing complex and multidimensional data but also stands as a testimony to its capabilities for clinical application, where accuracy and reliability are at stake. The high accuracy of such a model raises hopes for positive prospects as far as real-world applicability is concerned, where an early-stage intervention greatly ameliorates patient outcomes and survival rates.

The Decision Tree and K-Nearest Neighbors have also proven the importance of model selection and optimization to the Random Forest model's performances. The study on hyperparameter tuning and fine- tuning of these models revealed that small changes could result in drastic improvements in predictive performance. This holds true as it only reiterates that the parameterization of machine learning models and the tuning thereof must be given careful thought to optimize applications in any healthcare setting.

An important part of this research was the application of recent techniques for feature selection such as Permutation Importance and Principal Component Analysis, which additionally gave them an insight into the most important features for diagnosis. Identifying the specific contributing features allowed for enhanced interpretability of the models, allowing accessibility and transparency for healthcare workers. Understanding which variables carry the most weight in predictions brings credibility and trust to AI-powered tools, especially in medical decision- making.

The project under discussion does not only add aspects to the small amount of information known regarding non-invasive diagnostic/predictive tools but also opens an avenue for future research undertaking the development of AI with a focus on their integration into health systems. In fact, these days a lot of AI technologies evolve; they back the idea that the revolution of diagnostic procedures in faster, more accurate, and less cost- intense detection of multiple types of cancers, including ovarian cancer, is possible. This is by giving the breadth to machine learning-use processes in oncology, and thereby laying a further stage for research that will impose addition biomarkers, broadened datasets, and even more advanced algorithms, that is to say that in time it would help build up partnerships broadly in terms of improving health care.

For the more important implication of this work is far-reaching; it may bring a route to realize reduced mortality rates of ovarian cancers thanks to early and precise diagnostics. Furthermore, insights gained from here could serve as a showcase into how AI- based diagnostic tools can be developed for different cancers and diseases so that health could become not just more predictive but also more personalized and accessible in the future. Ultimately, it sets an example for more innovation on how to integrate AI into health care- more focused on the betterment of lives for patients around the globe

5.1. Pros:

- Early Detection and Improved Survival Rates: Identifies ovarian cancer at an early stage, increasing the chances of successful treatment and better patient outcomes.
- Non-Invasive and Cost-Effective: Offers a non-invasive alternative to traditional diagnostic methods, reducing patient discomfort and healthcare costs.
- High Accuracy and Reliability: Utilizes advanced machine learning models and performance metrics (like Recall and AUC) to ensure accurate and reliable cancer detection.
- Scalable and Versatile: Can be adapted to detect other types of cancers or diseases, making it a versatile diagnostic tool.
- Resource Optimization: Minimizes the need for expensive diagnostic tools and integrates easily into existing clinical workflows.

5.2. Cons

- Data Dependency and Bias: Model performance heavily depends on the quality and diversity of training data, which could lead to biased results if not representative of all patient groups.

- False Positives/Negatives: The system may produce false positives, leading to unnecessary stress and procedures, or false negatives, resulting in missed cancer cases.
- Complex Implementation and Maintenance: Requires continuous updates, maintenance, and validation to keep the model accurate and relevant with evolving medical data.
- Regulatory and Ethical Challenges: Meeting healthcare regulations and ensuring patient data privacy can be challenging.
- Clinical Adoption and Trust: Gaining trust and acceptance from healthcare professionals and patients may take time due to reliance on AI for critical medical decisions.

Future work for this project includes enhancing the model's accuracy by integrating more advanced algorithms and larger, more diverse datasets. We plan to incorporate multi-modal data, such as genetic and clinical information, to improve personalized predictions. Additionally, we aim to develop an automated report generation feature for clinicians, providing detailed diagnostic insights and recommendations. Expanding the system to detect other gynecological cancers and deploying it as a user-friendly application are also potential future developments.

Compliance with ethical standards

Acknowledgments

We express our gratitude to Mangalore Institute of Technology & Engineering for their motivation and support for research work publication and the department of Artificial Intelligence and Machine Learning (AIML) for providing the laboratory facility to develop this proposed idea and special thanks to our faculty and colleagues who have supported us with valuable and timely suggestions. We are also immensely grateful to all those references listed who provided us insight and great assistance for this work.. All this said, we appreciate how people around us collaborated very creatively and constructively towards achieving the development. And from our friends and families in particular, we owe enormous appreciations for unwavering encouragement and motivation in putting this work into consideration and execution. Last, we would like to extend our appreciation to the development team also acknowledges open source platforms, such as Kaggle, offering tools, datasets, and relevant frameworks for testing the functionality of the Early stages of ovarian cancer.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Debaditya Chakroborty, “Discovering genetic biomarkers for targeted cancer therapeutics with eXplainable AI” ,2023

[2] Kristofer Linton-Reid, George Wengert, “End-to-End Integrative Segmentation and Radiomics Prognostic Models Improve Risk Stratification of High-Grade Serous Ovarian Cancer: A Retrospective Multi-Cohort Study”, 2023

[3] Haoxin Zhang, “A random forest-based metabolic risk model to assess the prognosis and metabolism-related drug targets in ovarian cancer”, 2022

[4] Lujun Chen, Mingyang Lu and Bin Xu, “Using machine learning to predict ovarian cancer”, 2020

[5] Zhong Yu, Ouyang and Ling M, “Identification of prognosis- related hub genes of ovarian cancer through bioinformatics analyses and experimental verification”, 2022

[6] “Ovarian cancer data analysis using deep learning: A systematic review” F. Khan, M.A. Khan, S. Abbas, A. Athar,

[7] S.Y. Siddiqui, A.H. Khan, M.A. Saeed and M. Hussain, 2020.

[8] “Cloud-based breast cancer prediction empowered with soft computing approaches”. Journal of Healthcare Engineering, 2020.

[9] Siegel RL, Miller KD and Jemal A: “Cancer statistics, 2020”. CA Cancer J Clin 70: 7-30, 2020.

[10] Cao, H., Wang, Y., Chen, J., et al., “Swin-unet: Unet-like pure transformer for medical image segmentation”. CoRR abs/2105.05537,2021.

[11] Chen, C., Dou, Q., Chen, H., et al., “Unsupervised bidirectional cross modality adaptation via deeply synergistic image and feature alignment for medical image segmentation”. IEEE Trans. Medical Imaging ,2020.

[12] Ovarian cancer detection using optimized machine learning models with adaptive differential evolution Filbert H Juwono, Wei Kitt Wong, Hui Ting Pek, Saaveethya Sivakumar, Donata D Acula Biomedical Signal Processing and Control 77, 103785, 2022.

[13] Ovarian cancer detection and classification using machine leaning Ms. Aditya, I Amrita, Ashwini Kodipalli, Roshan Joy Martis 2021 5th international conference on electrical, electronics, communication, computer technologies and optimization techniques (ICEECCOT), 279-282, 2021.

[14] Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods E Sun Paik, Jeong-Won Lee, Jeong-Yeol Park, Ju-Hyun Kim, Mijung Kim, Tae-Joong Kim, Chel Hun Choi, Byoung-Gie Kim, Duk-Soo Bae, Sung Wook Seo Journal of gynecologic oncology 30 (4), 2019.

[15] Early-stage detection of ovarian cancer based on clinical data using machine learning approaches Md Martuza Ahamad, Sakifa Aktar, Md Jamal Uddin, Tasnia Rahman, Salem A Alyami, Samer Al-Ashhab, Hanan Fawaz Akhdar, AKM Azad, Mohammad Ali Moni Journal of personalized medicine 12 (8), 1211, 2022.

[16] Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers Eiryo Kawakami, Junya Tabata, Nozomu Yanaihara, Tetsuo Ishikawa, Keita Koseki, Yasushi Iida, Misato Saito, Hiromi Komazaki, Jason S Shapiro, Chihiro Goto, Yuka Akiyama, Ryosuke Saito, Motoaki Saito, Hirokuni Takano, Kyosuke Yamada, Aikou Okamoto Clinical cancer research 25 (10), 3006-3015, 2021.

[17] Machine Learning in Ovarian Cancer: A Review Authors: H. A. Akbari, P. H. Kamali, et al. Journal: Computers in Biology and Medicine (2020).

[18] Application of Random Forest Classifier for Ovarian Cancer Diagnosis Using Gene Expression Data. Authors: Y. R. Ghosh, S N. L. Ghosh, et al. Journal: Cancer Informatics (2018).

[19] Predicting Ovarian Cancer Using Clinical Data and Machine Learning Algorithms. Authors: L. V. K. Reddy, S. G. Basha, et al. Journal: Journal of Medical Systems (2020).

[20] Development of a Machine Learning Approach for Early Diagnosis of Ovarian Cancer from Biomarkers Authors:

[21] K. S. Tiwari, V. M. Vohra, et al. Journal: Bioinformatics (2019).

[22] Artificial intelligence-based models enabling accurate diagnosis of ovarian cancer using laboratory tests in China: a multicentre, retrospective cohort study Guangyao Cai, Fangjun Huang, Yue Gao, Xiao Li, Jianhua Chi, Jincheng Xie, Linghong Zhou, Yanling Feng, He Huang, Ting Deng, Yun Zhou, Chuyao Zhang, Xiaolin Luo, Xing Xie, Qinglei Gao, Xin Zhen, Jihong Liu The Lancet Digital health 6 (3), e176-e186, 2024.

[23] A deep learning framework for the prediction and diagnosis of ovarian cancer in pre-and post-menopausal women Blessed Ziyambe, Abid Yahya, Tawanda Mushiri, Muhammad Usman Tariq, Qaisar Abbas, Muhammad Babar, Mubarak Albathan, Muhammad Asim, Ayyaz Hussain, Sohail Jabbar Diagnostics 13 (10), 1703, 2023.

[24] Early Detection of Ovarian Cancer Using Machine Learning Models Authors: L. M. Jafari, M. A. Gharibzadeh, S.

[25] M. Asadollahi Journal: Biomedical Signal Processing and Control, 2021.

[26] Prediction of Ovarian Cancer Using Random Forest and Other Machine Learning Models Authors: X. Li, D. Yang, L. Liu, Li, J. Li Journal: Journal of Cancer Research and Clinical Oncology, 2020.

[27] Cancer Diagnosis Based on Machine Learning Classifiers Using Genetic Data Authors: M. E. Ali, M. R. Zaki Journal: Computational Biology and Chemistry, 2019.

[28] Predicting Ovarian Cancer Using Random Forest and Neural Networks with Genetic and Clinical Data Authors: M. A. Al- Ahmad, M. N. S. Swamy Journal: Health Information Science and Systems, 2020.

[29] Ovarian Cancer Detection Using Multi-Omics Data Information Science and Systems, 2020.

[30] Ovarian Cancer Detection Using Multi-Omics Data and Machine Learning Models Authors: S. N. R. M. Kadir, L. T. S. Chen, P. I. S. Soo Journal: BMC Bioinformatics, 2021

[31] Comparative Study of Machine Learning Algorithms for Ovarian Cancer Detection Authors: S. F. Z. T. Ma, T. H. H. Chan, K. Y. Ho Journal: Journal of Healthcare Informatics Research, 2020.

[32] A Novel Machine Learning-Based Diagnostic System for Ovarian Cancer Authors: H. Zhang, L. Yang, L. Zhang, X. Wang, Z. Liu Journal: Journal of Cancer, 2020.

[33] Ovarian Cancer Diagnosis Using Blood-Based Biomarkers and Machine Learning Models Authors: S. Li, X. Wu, T. Zhang, L. Jiang Journal: Artificial Intelligence in Medicine, 2021.

[34] Screening ovarian cancer by using risk factors: machine learning assists Raoof Nopour BioMedical Engineering OnLine 23 (1), 18, 202 (1), 18, 20