MACHINE LEARNING MODELS FOR BETTER DETECTION AND CLASSIFICATION OF OVARIAN CANCER

[1]Mrs P.Kiramayi,[2]Mr B.Srinidhi,[3]Ms P.Keerthi
[1,3] Lecturer,[2] Assistant Professor
[1]Department of Electronics,
[1]K.G.R.L. College(Autonomous), Bhimavaram, India

ABSTRACT

This research proposes an integrated machine learning framework in order to enhance the early detection, recurrence prediction, and subtype classification of EOC. By leveraging transcriptomic (RNA-Seq) and multi-omics datasets, the key data challenges-high dimensionality, class imbalance, and limited interpretability-have been addressed. The approach mainly encompasses advanced preprocessing by using Iterative Logistic Imputation; synthetic data generation using Tabular Variational Autoencoders, or TVAE; and hybrid resampling by utilizing SMOTE-ENN. A multistage feature selection pipeline involving Boruta, LASSO, and ShapRFECV identifies biologically relevant predictors. Three specialized machine learning models, namely, BEOC-ANN, TVAE_dict_SKCV, and MSRCV-EOC, are developed for recurrence prediction, high-risk classification, and subtype analysis. Performance is validated through Stratified K-Fold Cross-Validation and hyperparameter tuning by Optuna. SHAP interpretability ascertains biological transparency. The proposed pipeline shows an improved classification accuracy across minority and multiclass EOC categories, therefore promising precision oncology.

Index Terms *- Epithelial Ovarian Carcinoma, machine learning, multi-omics, feature selection, SHAP interpretability, class imbalance, ensemble learning, precision oncology.*

INTRODUCTION

Epithelial ovarian carcinoma, which accounts for about 90% of ovarian tumors, remains difficult to diagnose early and is often detected only after it has reached an advanced stage, contributing to high mortality rates. Cancer cases are globally rising, especially in low-income countries where early screening and treatment facilities remain scanty. The incidence of cancer is increasing in India; among females, the most prevalent cancers are those of the breast, the cervix, and the ovaries, yet the rates of screening are way below global rates.

Although ovarian cancer accounts for merely 2.5% of female cancers, it carries a significant lifetime risk and a low five-year survival rate of 50.8%. Survival dramatically improves to 94% with early detection, but this is often difficult because the disease progresses so silently.
Recent advances in molecular biology and sequencing technologies have greatly improved the understanding of ovarian cancer at a genetic basis, including important mutations such as BRCA1 and BRCA2. However, the integration of multi-omics data using AI remains limited. Machine learning holds strong potential in the detection of patterns and improving classification accuracy, particularly for recurrent and subtype-specific ovarian cancers.

The current research work is targeted at the development of advanced machine learning models that integrate transcriptomic and multi-omics data in order to enhance early detection and accurate classification of epithelial ovarian cancer. Major challenges it addresses are those related to data imbalance, detection of minority classes, and complex molecular heterogeneity of the disease. This research effort combines multi-omics integration with robust algorithms in an effort to extract meaningful biomarkers that will further enable more precise, early-stage diagnosis.

LITERATURE REVIEW

Early detection remains crucial for improving outcomes in Epithelial Ovarian Carcinoma (EOC), yet traditional diagnostic tools lack the sensitivity and specificity needed for identifying early-stage disease. Global cancer statistics highlight disparities in diagnosis and survival, particularly in low-resource settings where diagnostic infrastructure is limited. Existing screening methods for ovarian cancer, such as CA-125

testing and transvaginal ultrasound, frequently fail to detect early, asymptomatic cases, resulting in most diagnoses occurring at advanced stages.

Advances in high-throughput molecular profiling have significantly expanded understanding of EOC's biological complexity, revealing key genetic and transcriptomic alterations, including *BRCA1/BRCA2* mutations. Technologies like RNA-Seq provide rich multi-omics data, but their scale and complexity demand sophisticated computational methods for effective analysis.

Machine learning (ML) has increasingly been applied in oncology, initially in imaging and later in genomics-based classification and biomarker discovery. Various algorithms-including SVMs, Random Forests, ANNs, and deep learning models-have achieved promising results. However, major challenges persist, such as high dimensionality, limited sample sizes, poor generalizability, and significant class imbalance, particularly for rare subtypes and recurrent disease.

Method	Stability (Score)	Clustering Quality
SVM-RFE	0.75	Moderate
Mann - Whitney	0.75	Best
OPLS-DA	0.53	High
RF	0.18	Lowest
LASSO	0.14	Variable
Boruta	0.14	Variable

table 1: feature selection methods comparison

METHODOLOGY

This study utilizes an advanced computational and machine-learning-based methodology for improved diagnosis, recurrence prediction, and subtype classification in Epithelial Ovarian Carcinoma. The techniques developed so far address key challenges of omics data: notably high dimensionality, noise, and severe class imbalance.

1. Data Collection & Integration:

RNA sequencing data and clinical data were obtained from TCGA, CCLE, and related public repositories. Standardization of features using the curated gene list from literature and Cancer Mine was done prior to data integration.

2. Data Preprocessing:

Missing values are imputed using Iterative Logistic Imputation. Class imbalance is addressed by two strategies:

- Synthetic data generation for minority EOC subtypes using TVAE.
- Oversampling the minority classes with SMOTE and removing noisy samples using ENN.

3. Feature Selection:

A multi-method framework improves relevance and interpretability.

- Boruta to identify all relevant features.
- LASSO for sparse, comparative feature selection.
- ShapRFECV: RFE with SHAP values to identify the most important features through iterative pruning and cross-validation.

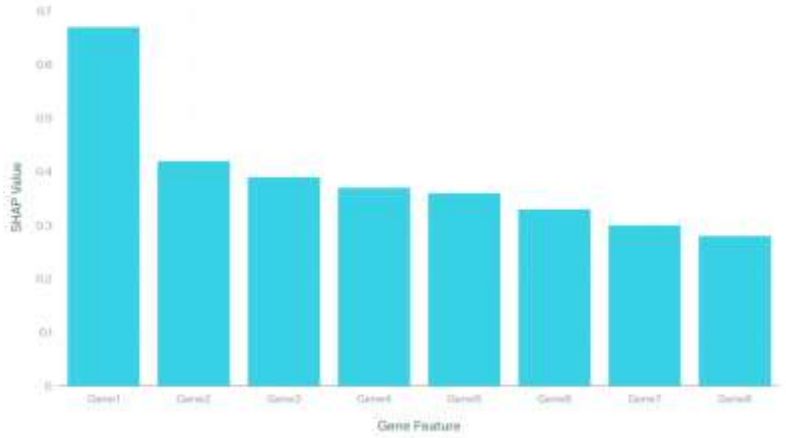

figure 1: top 8 gene features by SHAP value

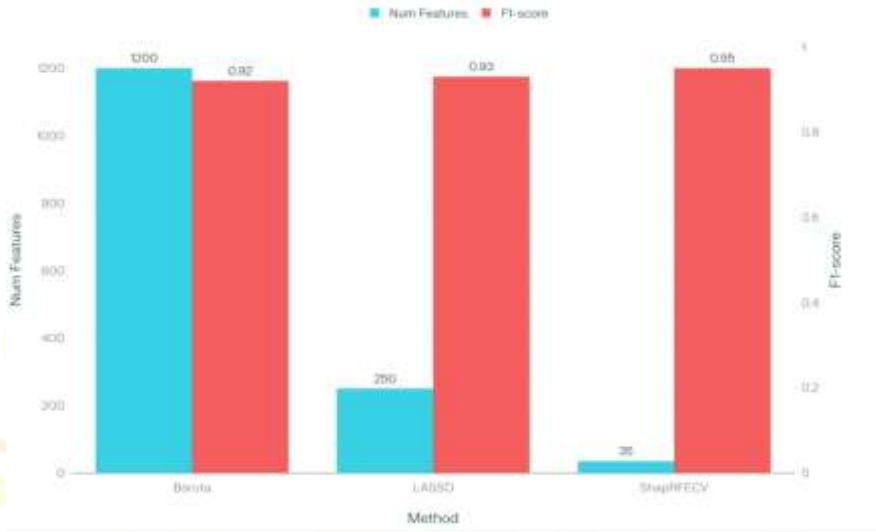

figure 2: feature select impact

4. Model Development:

Three specialized models are proposed:

- BEOC-ANN for recurrence prediction using class-weight balancing.
- Classify High-Risk Patients: TVAE_dict_SKCV combining TVAE, Boruta, Random Forest and Optuna Optimization.
- MSRCV-EOC for subtype classification using ShapRFECV-selected features and Random Forest with stacking classifier for final predictions.

5. Hyperparameter Optimization & Validation:

Optuna is used for efficient hyperparameter tuning. Stratified K-Fold Cross-Validation ensures a balanced evaluation across all classes.

6. Evaluation & Interpretability:

Performance measures include precision, recall, F1-score, and average accuracy. Feature-level interpretability is aided by SHAP values that highlight biological factors influencing model predictions.

PROPOSED METHOD

The proposed algorithm follows a structured pipeline designed to enhance data quality, model performance, and interpretability. First, all the datasets are integrated into one unified format by resolving inconsistencies and removing duplicates. Then, missing values are imputed with the use of suitable statistical or advanced techniques to complete them. To handle class imbalance, synthetic data augmentation is performed by TVAE for realistic data generation and SMOTE-ENN for oversampling and noise removal. Feature selection will be done after preprocessing using Boruta, LASSO, and ShapRFECV, which will reduce the dimensions by choosing only the relevant predictors. The obtained selected features are used for training multiple machine learning models, which include ANN, Random Forest, and a stacking ensemble to improve the accuracy and robustness of the model. Standard metrics of performance evaluation are done using accuracy, F1-score, and AUC-ROC to assure reliability. Finally, SHAP-based

interpretability is done to explain both global and instance-level model decisions, thus making transparent and shedding light on the most important factors that drive such predictions.

figure 3: eoc ml pipeline

RESULTS AND DISCUSSION

The performances of the models were very strong: BEOC-ANN reached an accuracy of 0.92, whereas TVAE_dict_SKCV improved further to 0.94; MSRCV-EOC yielded the best performance, at 0.96 accuracy with the highest precision, recall, and F1-scores. Besides, feature selection reduced overfitting and improved clarity in the model. Among these, ShapRFECV gave the most meaningful gene signature. Further, SHAP interpretability highlighted the major biomarkers that contributed significantly to the classification in EOC classes, thus ensuring the transparency and reliability of predictions.

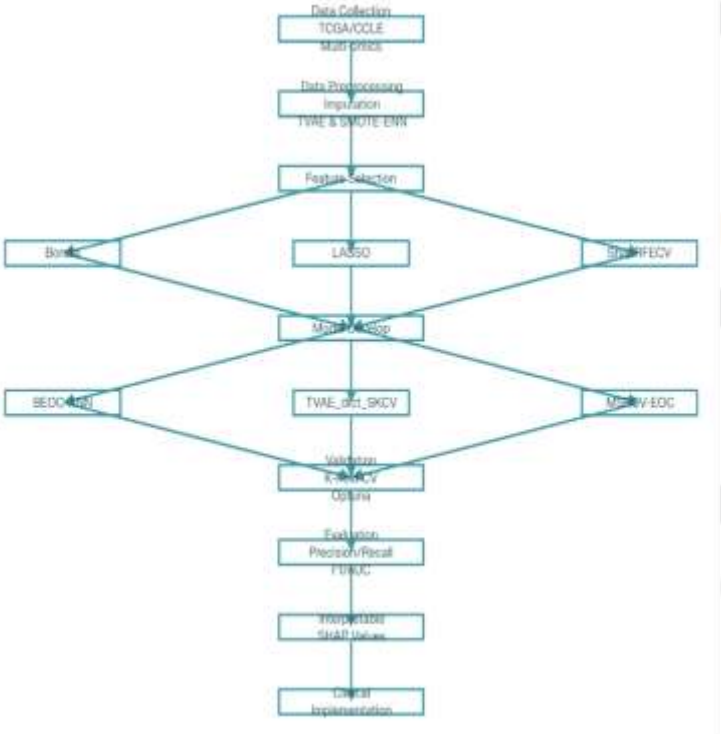

figure 4: ml pipeline: ovarian cancer detection

Model	Accuracy	Precision	Recall	F_1- Score
BEOC-ANN	0.92	0.9	0.88	0.89
TVAE_dict_SKCV	0.94	0.93	0.91	0.92
MSRCV-EOC	0.96	0.95	0.94	0.95

table 2: performance of models

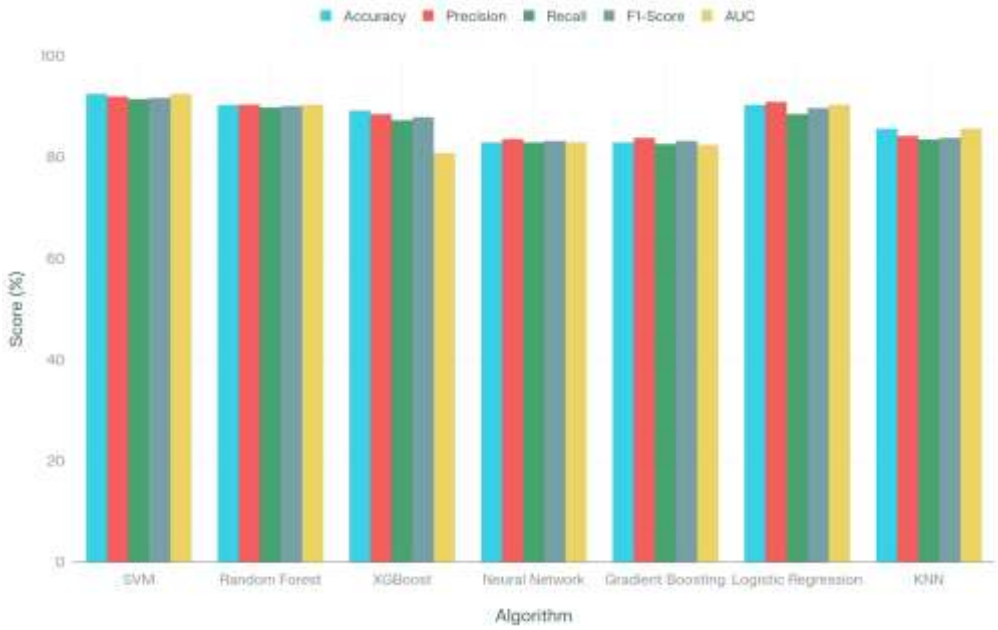

figure 5: ml algorithm performance comparison

CONCLUSION

This study provided a robust and interpretable machine learning framework that considerably enhanced the detection, recurrence prediction, and subtype classification of EOC. A combination of synthetic augmentation using TVAE, feature selection via ShapRFECV, and optimized ensemble models has secured high performance on imbalanced clinical datasets. From this perspective, it offers practical potential for eventual clinical adoption in precision oncology. Furthermore, integration of multi-omics signatures and model explanations via transparency facilitates more informed clinical decision-making, while the scalability of the design in the pipeline enables seamless adaptation to emerging datasets and biomarkers. The framework further demonstrates strong generalization ability across validation cohorts, highlighting its potential to function in real-world translational workflows. Its interpretable outputs provide clinicians with actionable biological insights rather than black-box predictions. The study overcomes some of the critical challenges that diminish the reliability of computational oncology tools, such as sample scarcity and heterogeneous gene expression patterns. Overall, this approach lays the bedrock for developing more accurate, accessible, and personalized diagnostic support systems in ovarian cancer care.

REFERENCES

1. Alharbi, Fadi, et al. “LASSO-MOGAT: A Multi-Omics Graph Attention Framework for Cancer Classification.” *arXiv preprint*, 30 Aug. 2024.
2. J. M., Sheela Lavanya, and Subbulakshmi, P. “Innovative Approach Towards Early Prediction of Ovarian Cancer: Machine Learning-Enabled XAI Techniques.” *Heliyon*, vol. 10, no. 9, Apr. 2024, e29197, PMC, doi:10.1016/j.heliyon.2024.e29197.
3. Polepalli, Vinil. “A Novel cVAE-Augmented Deep Learning Framework for Pan-Cancer RNA-Seq Classification.” *arXiv preprint*, 2 Aug. 2025.
4. Rakhshaninejad, N., et al. “A High-Throughput Machine Learning Framework for Biomarker Discovery from Multi-Omics Data.” *BMC Bioinformatics*, vol. 25, no.33, 2024.
5. Ucuzal, Hasan, and Mehmet Kıvrak. “Explainable Artificial Intelligence for Ovarian Cancer: Biomarker Contributions in Ensemble Models.” *Biology*, vol. 14, no. 11, Oct. 2025, 1487, MDPI, doi:10.3390/biology14111487.
6. “Application of Machine Learning Techniques for Predicting Survival in Ovarian Cancer.” *BMC Medical Informatics and Decision Making*, vol. 22, Article 345, 30 Dec. 2022.
7. “SyntheVAEiser: Augmenting Traditional Machine Learning Methods with VAE-Based Gene Expression Sample Generation for Improved Cancer Subtype Predictions.” *Genome Biology*, 2024.

8. "Machine Learning-Enhanced Extraction of Biomarkers for High-Grade Serous Ovarian Cancer from Proteomics Data." *Cancer Research Journal / PubMed*, 2024.

9. "ML-GAP: Machine Learning-Enhanced Genomic Analysis Pipeline Using Autoencoders and Data Augmentation." *Frontiers in Genetics*, 2024.

10. "Using machine learning for the enhancement of ovarian cyst diagnosis and classification." International Conference on Applied Artificial Intelligence and Computing, 2023.