**Early Stage Detection of Ovarian Cancer**

***Literature Survey***

*Submitted to - Dr. Sanchali Das*

*Submitted By :*

*Aditi Singh - E23CSEU1484*

*Nischay Rawat - E23CSEU1485*

*Deep Bansal - E23CSEU1480*

**INTRODUCTION**

The seventh most prevalent illness in women is ovarian cancer (OC), which carries a high lifetime risk. Low survival rates are caused by late-stage diagnosis and a lack of early indications, which highlights the critical need for efficient diagnostic instruments. Our goal is to provide comprehensive research of malignant epithelial ovarian tumours by integrating clinical data and biomarkers.

Crucial Biomarkers

-> CA125

-> HE4

-> NEU

These biomarkers represent the cutting edge of our research, each providing a distinct window into the complex molecular and clinical features of ovarian cancers. In addition to the well-known ovarian cancer marker CA125, HE4 and NEU complete a trio of biomarkers that together have the potential to improve the sensitivity and specificity of our diagnostic models.

The following primary goals are the focus of this investigation:

Early Detection: Using clinical data and biomarkers together, developing new approaches for the early detection of malignant epithelial ovarian cancers.

Biomarker importance: Identifying patterns suggestive of ovarian cancers by revealing the importance and interaction of CA125, HE4, and NEU with a wide range of clinical characteristics.

Comprehensive Analysis: To conduct rigorous and statistically sound analytical studies, advanced machine learning models are applied to a broad dataset that includes blood samples, general chemistry medical tests, and ovarian cancer markers.

Keywords: malignant epithelial ovarian tumours, biomarkers, clinical data, machine learning, gene sequencing, early detection, prognostication.

**LITERATURE SURVEY**

**PAPER 1**

**Debaditya Chakroborty et.al** [[1]](https://doi.org/10.1101/2023.07.24.550346)

By utilizing eXplainable AI (XAI), researchers sought to find genetic markers for targeted cancer therapy with a particular emphasis on high grade serous carcinoma (HGSC). Their approach combined SHapely Additive Explanations (SHAP) and XGBoost to identify pertinent biomarkers, resulting in clear predictions and economical hypothesis testing. Although the study has been successful in predicting the probability of a 5-year survival, it recognizes that meaningful confirmation requires human trials, particularly considering the reliance on mouse tumor models. Although the method provides biological insights, it is still difficult to prove cause-and-effect linkages because some biomarkers show context dependency. This study emphasizes the promise of XAI and probabilistic methods in cancer research, but it also emphasizes the need for additional experimental validation.

**PAPER 2**

**Kristofer Linton-Reid et.al** [[2]](https://doi.org/10.1101/2023.04.26.23289155)

An end-to-end integrated method is used by researchers under the direction of Kristofer Linton-Reid to tackle the problem of risk classification in high-grade serous ovarian cancer (HGSC). This includes using machine learning (ML) and deep learning (DL) models, such as CNNs and Cox proportional hazard models, as well as radiomics analysis for non-invasive tumor information extraction and CT scan segmentation to locate ovarian masses. The intricacy of DL modelling and the non-invasiveness of radiomics analysis are among the benefits. DL and ML models are resource-intensive, and big medical picture datasets are required for radiomics feature validation. These are some of the drawbacks. Consistently exhibiting higher prediction accuracy, the Permutation-Variable Importance Random Forest - Random Survival Forest (PVIRF-RSF) model offers a more accurate data-driven method and shows promise in risk assessment.

**PAPER 3**

**Haoxin Zhang et.al** [[3]](https://doi.org/10.1016/j.compbiomed.2022.106432)

A novel random forest-based metabolic risk model for ovarian cancer prognostic assessment and identification of metabolism-related therapeutic targets is presented by Haoxin Zhang and colleagues. Patients are categorized into low and high metabolic subtypes based on the model's identification of 17 metabolic pathways associated with prognosis. Several datasets are used in the study to increase sample size, and different biological studies are used to identify subtype differences. WGCNA is utilized to identify genes linked to subtypes, and an XGBoost classifier is employed to predict subtypes. Although the method offers insightful suggestions for novel pharmacological targets and customizable strategies, its drawbacks include a lack of clinical validation, possible bias in the TCGA data, and the requirement to evaluate generalizability across separate datasets. When predicting metabolic conditions, the XGBoost Classifier shows excellent accuracy and a high area under the ROC curve, enhancing the understanding of cellular metabolisms role in ovarian cancer.

**PAPER 4**

**Mingyang Lu et.al** [[4]](https://doi.org/10.1016/j.ijmedinf.2020.104195)

Using ML approaches and the MRMR feature selection method on a dataset consisting of 49 variables, Mingyang Lu and colleagues address the prediction of ovarian cancer through machine learning. A decision tree model performs better than techniques like logistic regression and risk of ovarian malignancy algorithm(ROMA) when it is trained on one dataset and tested on another that includes both actual and control instances. The method's benefits include better performance over ROMA, a large dataset providing a comprehensive view, and reliance on two biomarkers (HE4 and CEA) for clinical simplicity. The dataset's exclusive Chinese patient population, possible bias resulting from missing data imputing, and the requirement for additional clinical validation and benchmarking are its drawbacks. Even though the study's early ovarian cancer detection method beat ROMA, it still highlights the need for additional validation, benchmarking, and generalizability across various populations.

**PAPER 5**

**Zhong Yu et.al** [[5]](https://journals.lww.com/md-journal/Fulltext/2022/09090/Identification_of_prognosis_related_hub_genes_of.59.aspx?context=LatestArticles)

Through a multi-step process that includes building protein-protein interaction networks, weighted gene co-expression analysis (WGCNA), survival analysis, qRT-PCR gene expression validation, mutation analysis, functional pathway analysis, drug sensitivity analysis, and assessment of tumor mutation burden (TMB), Zhong Yu and Ouyang Ling set out to identify prognosis-related hub genes in ovarian cancer. Comprehensive datasets include clinical data and the discovery of novel biomarkers (ALDH1A2, CLDN4, GPR37) are two benefits of the study. Analysis of functional pathways illuminates the mechanisms underlying ovarian cancer. Nevertheless, qRT-PCR's limited sample size, inconsistent findings with earlier research, and the critical requirement for additional experimental validation are some of its drawbacks. Although the study identifies three putative prognostic hub genes that provide enhanced understanding of ovarian cancer, it emphasizes the need for larger datasets and experimental validation to fully understand the implications and roles of these genes.

**Our Contributions**

**Integration of Ensemble Learning :**

* We implemented ensemble learning techniques by combining two well-established models, Random Forest and Decision Tree, to improve the accuracy and robustness of early-stage ovarian cancer detection.
* By leveraging the strengths of both models, we achieved a more resilient prediction system, with reduced risk of overfitting and enhanced generalization capabilities.

**Hyperparameter Tuning :**

• Optimized Model Parameters: Used GridSearchCV to systematically search for the best combination of hyperparameters across models (Random Forest, Decision Tree, etc.). This significantly boosted model performance by identifying the best-fit parameters.

* Cross-Validation: Employed k-fold cross-validation to assess model stability and prevent overfitting, leading to more reliable generalization on unseen data.

**Feature Selection and Engineering—Used model-based methods:**

* (e.g., feature importance from Random Forest or Decision Tree) to identify the most influential features, allowing for refined input and improving predictive performance.

**Visualization and Interpretability :**

* Feature Importance Plotting: Visualized the top three or more most important features for each model, allowing domain experts to understand which factors most influence predictions.
* Metric Comparison Plotting: Compared model performance across various metrics in a single visualization, providing a clear and comprehensive comparison for decision-makers.

**REFERENCES**

**Dataset Links:**

[Ovarian Cancer Clinical Data](https://github.com/martuzaiu/Ovarian_Cancer_Project)

**Literature Survey:**

**PAPER 1**

**Debaditya Chakroborty et.al**

<http://biorxiv.org/lookup/doi/10.1101/2023.07.24.550346>

**PAPER 2**

**Kristofer Linton-Reid et.al**

<https://www.medrxiv.org/content/10.1101/2023.04.26.23289155v1>

**PAPER 3**

**Haoxin Zhang et.al**

<https://www.sciencedirect.com/science/article/abs/pii/S0010482522011404?via=ihub>

**PAPER 4**

**Mingyang Lu et.al**

<https://linkinghub.elsevier.com/retrieve/pii/S1386505620302781>

**PAPER 5**

**Zhong Yu et.al**

<https://journals.lww.com/md-journal/Fulltext/2022/09090/Identification_of_prognosis_related_hub_genes_of.59.aspx?context=LatestArticles>

**Other:**

<https://pmc.ncbi.nlm.nih.gov/articles/PMC6376972/>