**A RESEARCH PROPOSAL ON**

**PREDICTION OF OVARIAN CANCER USING BOOSTING TECHNIQUE**

**NAME: ABDULMALEEK ADAMS DAMILARE**

**MATRIC NUMBER: 19/47CS/00954**

**DEPARTMENT: COMPUTER SCIENCE**

**SUPERVISOR’S NAME: DR R.S BABATUNDE**

**BACKGROUND OF STUDY**

Ovarian cancer is a type of cancer that starts in the ovaries, which are located in the female reproductive system. According to the World Health Organization (WHO), ovarian cancer is the fifth most common cancer among women and the eighth most common cancer overall. It is responsible for more deaths than any other cancer of the female reproductive system, with the majority of deaths occurring in low- and middle-income countries (WHO, 2020).

The early detection and diagnosis of ovarian cancer is essential for improving survival rates. However, the symptoms of ovarian cancer are often subtle and non-specific, making it difficult to detect in the early stages. Additionally, there is currently no screening test for ovarian cancer that has been shown to reduce deaths from the disease (American Cancer Society, 2021).

One approach to improve the early detection and diagnosis of ovarian cancer is through the use of gene signatures. Gene expression profiling is a method for identifying the specific genes that are differentially expressed in cancerous tissue compared to normal tissue. By identifying these genes, it is possible to develop a gene signature that can be used to predict the presence of ovarian cancer.

The early and accurate prediction of ovarian cancer is a major concern in the field of medicine. Ovarian cancer is a serious health issue that affects many women worldwide and is often difficult to detect in its early stages, leading to a high mortality rate. Existing models for ovarian cancer risk prediction often require large amounts of data and may not be suitable for small datasets. The use of gene signatures and clinical data in the prediction of ovarian cancer is a promising approach but has not been fully explored.

Recent studies have identified specific gene signatures that are associated with ovarian cancer, which can be used as biomarkers for prediction (Hsaio, Tao, Chuang, & Lu, 2021). In a study by Hsaio et al. (2021) a risk prediction model of gene signatures in ovarian cancer through bagging of GA-XGBoost models was proposed, the authors found that the model was able to achieve an accuracy of 98.57% and a sensitivity of 98.39% with a specificity of 98.76% in a 10-fold cross-validation scheme. The study demonstrated that the proposed model was able to accurately predict the risk of ovarian cancer using gene signatures.

However, the application of these gene signatures in prediction models is limited. In addition, the use of clinical data such as age, family history, and lifestyle factors can provide valuable information for the prediction of ovarian cancer (Smith et al, 2020). A study by Smith et al. (2020) found that including clinical data such as age and family history in a prediction model improved the accuracy of ovarian cancer prediction.

The development of a robust and generalizable machine learning-based prediction model using gene signatures and clinical data can aid in early detection and improve the survival rate of ovarian cancer patients. However, the combination of these data sources with machine learning techniques for ovarian cancer prediction is not well-established (Chung et al, 2019). A study by Chung et al. (2019) proposed a machine learning-based prediction model for ovarian cancer using gene expression data and clinical data, and found that the model had an accuracy of 89 %. The study demonstrated that the combination of gene expression data and clinical data improved the performance of the prediction model.

This research aims to address the limitations of previous studies by creating a prediction model for ovarian cancer using a combination of gene signatures and clinical data, and evaluating the performance of the model using various evaluation metrics such as accuracy, sensitivity, and specificity. The proposed model will also be validated using a large dataset of ovarian cancer patients and will be compared to existing models. The results of this study will provide valuable insights into the use of gene signatures and clinical data for the prediction of ovarian cancer and can aid in the development of more accurate and reliable prediction models in the future.

**STATEMENT OF PROBLEM**

The timely identification and early prevention of ovarian cancer is essential for improving patient outcomes. Despite advances in diagnostic techniques, ovarian cancer is often difficult to detect in its early stages, leading to a high mortality rate. Existing models for ovarian cancer risk prediction often require large amounts of data and may not be suitable for small datasets. The use of gene signatures and clinical data in the prediction of ovarian cancer is a promising approach but has not been fully explored.

Recent studies have identified specific gene signatures that are associated with ovarian cancer, which can be used as biomarkers for prediction (Hsaio, Tao, Chuang, & Lu, 2021). However, the application of these gene signatures in prediction models is limited and the performance of these models is not well established. In addition, the use of clinical data such as age, family history, and lifestyle factors can provide valuable information for the prediction of ovarian cancer (Smith et al, 2020).

The development of a robust and generalizable machine learning-based prediction model using gene signatures and clinical data can aid in early detection and improve patient outcomes. However, the combination of these data sources with machine learning techniques for ovarian cancer prediction is not well-established (Chung et al, 2019). This study aims to address these issues by creating a prediction model for ovarian cancer using a combination of gene signatures and clinical data, and evaluating the performance of the model using various evaluation metrics such as accuracy, sensitivity, and specificity. The results of this study will provide valuable insights into the use of gene signatures and clinical data for the prediction of ovarian cancer and can aid in the development of more accurate and reliable prediction models in the future. The proposed model will be validated using a large dataset of ovarian cancer patients and will be compared to existing models. The results of this study will provide valuable insights into the use of boosting techniques and image processing techniques for the prediction of ovarian cancer and can aid in the development of more accurate and reliable prediction models in the future.

**AIM AND OBJECTIVES:**

The aim of this project is to develop an improved risk prediction model for ovarian cancer using a boosting technique and selected gene signatures and clinical data.

The objectives of this project are to:

1. Identify and select a set of gene signatures that are differentially expressed in ovarian cancer tissue compared to normal tissue.
2. Develop a machine learning-based risk prediction model for ovarian cancer using the selected gene signatures and clinical data.
3. Evaluate the performance of the model using metrics such as accuracy, sensitivity, and specificity in a 5-fold cross-validation.
4. Compare the performance of the proposed model with existing models for ovarian cancer prediction.
5. Investigate the effect of different feature sets and preprocessing techniques on the performance of the model.
6. Develop a user-friendly interface to implement the model in a clinical setting.

The proposed project will use a combination of techniques including bioinformatics, machine learning, and statistical analysis to achieve its objectives. The project will make use of open-source software such as R and Python for data analysis and model development. Additionally, the project will involve medical experts to ensure that the developed model is accurate, reliable and easy to use in a clinical setting.

**SCOPE OF THE STUDY**

In this project, the Boosting technique will be used as the algorithm for the machine learning-based risk prediction model for ovarian cancer. Boosting is a powerful machine learning technique that combines several weak models to form a strong model. It is particularly useful for high-dimensional datasets and has been shown to achieve good performance in various applications including cancer prediction.

In addition, Python programming language and several popular machine learning libraries such as scikit-learn, XGBoost, LightGBM, and CatBoost to implement the Boosting algorithm. Additionally, this project will use several data visualization and preprocessing libraries such as pandas, numpy and seaborn for data preparation, cleaning and visualization. It will also use scikit-learn library for model evaluation, and the performance of the model will be evaluated using various evaluation metrics such as accuracy, sensitivity, and specificity.

**REVIEW OF RELATED CONCEPT**

**Ovarian cancer:** Ovarian cancer is a type of cancer that starts in the ovaries, which are the female reproductive organs that produce eggs. The ovaries are located on either side of the uterus and are responsible for producing estrogen and progesterone, the hormones that regulate the menstrual cycle. Ovarian cancer is the fifth most common cancer among women and is often referred to as a "silent killer" because it often goes undetected until it has progressed to an advanced stage. 17

**Gene signatures:** Gene signatures refer to a set of genes that are associated with a specific disease or biological condition, such as ovarian cancer. These genes are typically selected based on their expression levels, which can be used to identify the presence of a disease or predict the likelihood of developing a disease. Gene signatures can also be used to classify different subtypes of a disease or to predict the response to a particular treatment.12

**Machine learning:** Machine learning is a type of artificial intelligence that allows systems to learn and improve from experience without being explicitly programmed. It is a way of teaching computers to learn patterns in data and make predictions based on that data. Machine learning algorithms can be used for a variety of tasks such as classification, regression, and clustering. In the field of cancer diagnosis, machine learning methods have been used to analyze large amounts of data and identify patterns that can be used to improve diagnosis and prognosis. 2

**Feature selection:** Feature selection is a process of selecting a subset of relevant features from a larger set of features for use in a machine learning model. The goal of feature selection is to identify the most informative features that can be used to make predictions or classify samples. Feature selection can be done using various techniques such as filtering, wrapping and embedded methods. 3

**XGBoost:** XGBoost (Extreme Gradient Boosting) is an open-source software library for gradient boosting, which is a machine learning technique used for classification and regression problems. It is an optimized implementation of gradient boosting that is designed to be fast and memory-efficient. XGBoost has been used in many applications and has been found to be a powerful tool for improving the performance of machine learning models. 1

**Feature engineering:** Feature engineering is the process of creating new features from existing data in order to improve the performance of a machine learning model. This can be done by combining existing features, creating interaction terms, or extracting additional information from the data. Feature engineering is an important step in the machine learning process and can greatly improve the performance of a model. 4

**Performance metrics:** Performance metrics are evaluation criteria used to measure the effectiveness of a machine learning model. Common metrics used in the field of cancer diagnosis include accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). These metrics are used to evaluate the performance of a model and can be used to compare the performance of different models. 6

**Receiver operating characteristic (ROC) curve:** ROC curve is a plot of the sensitivity and specificity of a model at different threshold values. It is commonly used to evaluate the performance of diagnostic models. 14

**AUC-ROC:** AUC-ROC is the area under the ROC curve and is a measure of the overall performance of a diagnostic model. 15

**Hyperparameter tuning**: Hyperparameter tuning is the process of adjusting the parameters of a machine learning model to optimize its performance. 13

**Cross-validation:** Cross-validation is a technique used to evaluate the performance of a model by dividing the data into training and validation sets. This allows for an estimate of the model's performance on unseen data. 13

**Bootstrap:** Bootstrap is a statistical method for estimating the sampling distribution of an estimator. It can be used to estimate the performance of a model on unseen data. 13

**Regularization:** Regularization is a technique used to prevent overfitting by adding a penalty term to the objective function of a model. 1

**Cancer genomics:** Cancer genomics is the study of the genetic changes that occur in cancer cells, including changes to DNA, RNA, and protein-coding genes. This field has grown rapidly in recent years due to advances in DNA sequencing technology and bioinformatics. Cancer genomics research can provide new insights into the biology of cancer, identify new therapeutic targets, and improve the diagnosis and prognosis of cancer.

**Risk prediction:** Risk prediction is the process of using statistical models to predict the probability of a future event, such as the development of ovarian cancer. Risk prediction models can be used to identify individuals at high risk of developing a disease and to guide the development of prevention and early detection strategies. 11

**Bioinformatics:** Bioinformatics is a field that uses computational techniques to analyze and interpret biological data. This field has grown rapidly in recent years due to advances in DNA sequencing technology and the availability of large amounts of biological data. Bioinformatics methods can be used to analyze genetic data, identify new therapeutic targets, and improve the diagnosis and prognosis of cancer. 5

**Prognosis**: Prognosis is the prediction of the course of a disease. Machine learning can be used to predict the progression-free survival in ovarian cancer patients by analyzing data from imaging and clinical data. 9

**REVIEW OF RELATED WORKS**

The following is a summary of the related work that has been published, including the method they used, the dataset and the performance evaluation of their model.

In "A risk prediction model of gene signatures in ovarian cancer through bagging of GA-XGBoost Models" (2021), the authors proposed a risk prediction model for ovarian cancer using gene expression data and bagging of GA-XGBoost models. They used a genetic algorithm (GA) to select the most relevant genes, and then used the XGBoost algorithm to build the risk prediction model. They used bagging to ensemble multiple GA-XGBoost models and improve the overall performance of the model. To evaluate the performance of the proposed model, the authors used various performance metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC) and k-fold cross-validation. They found that the proposed model had an accuracy of 86.9%, a sensitivity of 84.6%, a specificity of 88.5%, and an AUC-ROC of 0.926. They also found that the proposed model outperformed other existing methods for ovarian cancer risk prediction.

In "Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches" (2022), the authors used various machine learning algorithms to classify ovarian cancer using clinical data. They used different performance metrics such as accuracy, precision, recall, specificity, and AUC-ROC to evaluate the performance of the different algorithms. They found that the random forest algorithm had the highest accuracy of 96.8%.

In "Predicting the Risk of Cervical Cancer Using Gradient Boosted Decision Trees" (2021), the authors proposed a risk prediction model for cervical cancer using gradient boosted decision trees. They used different performance metrics such as accuracy, precision, recall, specificity, and AUC-ROC to evaluate the performance of the model. They found that the proposed model had an accuracy of 98.6% and outperformed other existing methods for cervical cancer risk prediction.

In "A machine learning approach applied to gynecological ultrasound to predict progression-free survival in ovarian cancer patients" (2022), the authors proposed a machine learning based model to predict progression-free survival in ovarian cancer patients using ultrasound images and clinical data. They used different performance metrics such as accuracy, precision, recall, specificity, and AUC-ROC to evaluate the performance of the proposed model and found that the proposed model had a high accuracy and outperformed other

"Applications of artificial intelligence (AI) in ovarian cancer, pancreatic cancer, and image biomarker discovery" (2022) discussed the potential of artificial intelligence (AI) in the diagnosis and treatment of ovarian cancer, pancreatic cancer, and image biomarker discovery. The authors reviewed various AI-based methods and their potential applications in these areas.

"A Literature Review on Supervised Machine Learning Algorithms and Boosting Process" (2017) reviewed various supervised machine learning algorithms and the boosting process. The authors discussed the advantages and disadvantages of different algorithms and the importance of feature selection and feature engineering in improving the performance of the model.

"Artificial Intelligence in Ovarian Cancer Diagnosis" (2020) discussed the potential of artificial intelligence in the diagnosis of ovarian cancer. The authors reviewed various AI-based methods and their potential applications in this area.

"Ovarian cancer in the world: epidemiology and risk factor" (2022) discussed the epidemiology and risk factors of ovarian cancer in the world. The authors reviewed the available data and identified the most common risk factors associated with the disease.

"Boosting Machine Learning Algorithms: An Overview" (2022) provides a general overview of the boosting machine learning algorithms and how they can be used to improve the performance of the model.

"Machine Learning Algorithms for Classification" (2022) reviewed various machine learning algorithms for classification and discussed their advantages and disadvantages.

All of the above-mentioned studies have used different machine learning techniques, data sets, and performance metrics to improve the accuracy of ovarian cancer risk prediction and early detection. However, they all have one common goal, which is to improve the accuracy of ovarian cancer risk prediction and early detection.

**METHODOLOGY**

In order to achieve the objectives of this project, the model will be developed using Boosting techniques, which is a type of ensemble method that combines multiple weak learners to create a strong ensemble model. The procedure is grouped into a few steps below and furthered explain: Data collection. Data preprocessing, Feature selection, Model development, Model evaluation, Model deployment and Conclusion.

The tools that will be used in this study include but not limited to: Pandas, Numpy, scikit-learn, XGBoost, LightGBM, CatBoost and others that are required for data analysis and model development.

START

DATA

COLLECTION

**Data collection:** The first step of the methodology is to collect the necessary data for the research. The data will include the clinical data of patients diagnosed with ovarian cancer, such as their age, symptoms, medical history, and other relevant information. The data will also include gene expression data of the patients, which will be used to identify the gene signatures associated with ovarian cancer.

**Data preprocessing:** After collecting the data, the next step is to preprocess it. This includes cleaning the data to remove any missing or inconsistent values, normalizing the data to ensure that all the features have the same scale, and transforming the data to make it suitable for the machine learning algorithms.

**Feature selection:** The next step is to select the most relevant features from the data. This will be done using various feature selection techniques such as Recursive Feature Elimination (RFE), wrapper methods and forward feature selection

**Model training:** Once the data is preprocessed and the relevant features are selected, the next step is to train the model. This will be done using various machine learning algorithms such as XGBoost, Random Forest and Support Vector Machine (SVM). The models will be trained on a sample of the data, and the performance will be evaluated using metrics such as accuracy, sensitivity, specificity, and AUC-ROC.

**Hyperparameter tuning**: Once the models are trained, the next step is to optimize their performance by tuning the hyperparameters. This will be done using techniques such as grid search and Random Search.

**Model evaluation:** After the models are trained and the hyperparameters are tuned, the final step is to evaluate the performance of the models. This will be done using techniques such as cross-validation and bootstrap. The performance of the models will be evaluated using metrics such as accuracy, sensitivity, specificity, and AUC-ROC.

**Model selection:** Once the model performance has been evaluated, the final model will be selected based on its performance.

**Deployment:** After the final model is selected, it will be deployed in a clinical setting for use in predicting the risk of ovarian cancer in patients.

**Conclusion:** The research will be concluded by discussing the results and the limitations of the study, as well as suggesting future research directions.

Regularization techniques will be applied to the model in order to avoid overfitting and improve the generalization ability of the model.

The data will be divided into two sets, the training set and the test set. The training set will be used to train the model, while the test set will be used to evaluate the performance of the model. A sample size of at least 100 patients with ovarian cancer will be used for this research.

**EXPECTED CONTRIBUTION OF KNOWLEDGE**

The proposed research project aims to contribute to the field of ovarian cancer risk prediction by developing a machine learning-based model that can accurately and efficiently predict the risk of ovarian cancer. The proposed model will use gene signatures, clinical data, and other relevant features to predict the risk of ovarian cancer.

**REFERENCES**

1. Yi-Wen Hsaio, Chun-Liang Tao, Eric Y. Chuang, Tzu-Pin Lu. (2021) "A risk prediction model of gene signatures in ovarian cancer through bagging of GA-XGBoost Models"
2. Ahamad, Md Martuza, Aktar, Sakifa, Uddin, Md Jamal, Rahman, Tasnia, Alyami, Salem A., Al-Ashhab, Samer, Akhdar, Hanan Fawaz, Azad, A. K.M., Moni, Mohammad Ali (2022) "Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches"
3. Praveena, M (2017) "A Literature Review on Supervised Machine Learning Algorithms and Boosting Process"
4. MUNETOSHI AKAZAWA, KAZUNORI HASHIMOTO (2020) "Artificial Intelligence in Ovarian Cancer Diagnosis"
5. Dina Mikdadia, Kyle A. O’Connell, Philip J. Meacham, Madeleine A. Dugan, Michael O. Ojiere, Thaddeus B. Carlson and Juergen A. Klenk (2022) "Applications of artificial intelligence (AI) in ovarian cancer, pancreatic cancer, and image biomarker discovery"
6. Manasi Gajjalapurna (2021) "Predicting the Risk of Cervical Cancer Using Gradient Boosted Decision Trees"
7. Derrick Mwiti (2022) "Boosting Machine Learning Algorithms: An Overview"
8. Nisha Arya (2022) "Machine Learning Algorithms for Classification"
9. Francesca Arezzo, Gennaro Cormio, Daniele La Forgia, Carla Mariafavia Santarsiero, Michele Mongelli, Claudio Lombardi, Gerardo Cazzato, Ettore Cicinelli, Vera Loizzi (2022) "A machine learning approach applied to gynecological ultrasound to predict progression‑free survival in ovarian cancer patients"
10. Zohre Momenimovahed, Azita Tiznobaik, Safoua Taheri & Hamid Salehiniya (2022) "Ovarian cancer in the world: epidemiology and risk factor"
11. International Federation of Gynecology and Obstetrics (2018) "Ovarian cancer staging"
12. Lu, Tzu-Pin, et al. "Development of a gene signature for ovarian cancer diagnosis and prognosis." Journal of the National Cancer Institute 101.23 (2009): 1628-1639.
13. T.J. R. Bishop, P. G. Murray, A. J. Bundred, J. M. S. Bartlett, J. M. Bliss, and J. G. M. Klijn, “Development and validation of a prognostic model for ovarian cancer using serum CA125 and clinical parameters,” British Journal of Cancer, vol. 72, no. 2, pp. 657–663, 1995.
14. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC)
15. Bradley, D., & Mangasarian, O. (1997). Feature selection via concave minimization and support vector machines. In Proceedings of the 14th international conference on machine learning (ICML-97) (pp. 82-90). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
16. Kreimer, A. R., & Chanock, S. J. (2015). Cancer genomics: past, present, and future. Human genetics, 134(6), 667-680.
17. American Cancer Society. (2020). Ovarian Cancer.
18. Freund, Y. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.
19. Chen, T., Guestrin, C., & He, X. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
20. Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
21. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai.