EARLY DETECTION OF PANCREATIC CANCER USING BIOMARKER-DRIVEN MACHINE LEARNING ALGORITHMS

[1]B Aditya, [2]D Sandhya Rani, [3]R Sai Krishna, [4]Manglarapu Srujana, [5]Vemula Harikanth Goud, [6]Tompa Anitha UG

[123]Assistant Professor, [456]UG Student

[123456]Department of CSE

[123456]CMR Technical Campus Hyderabad,Telangana, India-501401

[1]adi.sacs@gmail.com, [2]davu.sandhya@gmail.com, [3]regurisai@gmail.com, [4]srujanamanglarapu@gmail.com, [5]harikanth1523@gmail.com, [6]rangaraotompa@gmail.com

***Abstract*—pancreatic cancer is one of the more deadly cancers because of the lack of timely symptoms and delayed patient presentation. It is evident from existing methods that conventional screening methods, such as the biomarker CA19-9, cannot perform with the desired level of sensitivity and specificity in identifying cancer at an early stage. This manuscript will explore challenges associated with identifying cancer in its early stages based on a machine learning approach and discuss a biomarker-based machine learning approach for identifying cancer more accurately by integrating various clinical, metabolic, and inflammatory markers. It can be inferred from the experiment conducted on real-time patient data and evaluated based on various biomarkers and the machine learning approach implemented within the framework for making predictions on the identified features for cancer and its predictions during the early stages of cancer. The performance was found better for XGBoost with 90.1% accuracy, 91.4% sensitivity, 88.9% specificity, and 0.94 AUC, where the existing cancer biomarker and its additional biomarkers have shown high significance in improving the overall performance predictions. The proposed approach provides a comprehensive, non-invasive tool for understanding cancer more deeply for healthy individuals and those already affected with cancer.**

***Index Terms*—Cancer antigen 19-9 (CA19-9),Vector Machines(SVM), Random Forest (RF), and XGBoost**

I. Introduction

Pancreatic cancer has been and continues to be one of the most lethal cancers in the world, and this is primarily due to the asymptomatic nature of the disease in its early stages, resulting in a substantial delay in diagnosis. Despite the fact that it constitutes a proportionately smaller part of the total incidence of cancers, pancreatic cancer is a major cause of cancer-related deaths, with a five-year survival rate of less than 12 percent.

The traditional methods of diagnosis, such as imaging studies, histopathological analysis, and blood markers such as carbohydrate antigen 19-9 (CA19-9), have their own limitations. For instance, CA19-9 has been shown to be less sensitive for the early diagnosis of pancreatic cancer and can also be elevated in non-malignant conditions of the pancreas and the hepatobiliary system.

Recent breakthroughs in biomedical research have made it possible to identify multiple biomarkers in the circulation using genomic, proteomic, and clinical data. However, the identified biomarker data is usually high-dimensional, heterogeneous, and non-linear in nature, making it challenging to analyze using

traditional statistical analysis. Machine learning (ML) algorithms provide a viable alternative by learning complex patterns among multiple biomarkers and facilitating automatic risk stratification.

Although several models based on ML have been proposed for the prediction of pancreatic cancer, the existing literature has several limitations. Most of the proposed models are based on a small set of biomarkers, do not have a robust mechanism for feature selection, or do not consider the interpretability of the model, which is a critical requirement for clinical translation. Additionally, the interpretation of the results and analysis of biomarker contributions have not been adequately investigated.

To address these issues, this paper proposes a biomarker driven machine learning framework for the early diagnosis of pancreatic cancer. The proposed framework integrates data preprocessing, feature selection, and ensemble learning techniques, and also employs explainable artificial intelligence (XAI) with SHAP to enhance interpretability. Through the comparison of various classifiers, this paper seeks to identify the optimal predictive model and provide meaningful clinical insights into biomarkers. Although there have been extensive studies, the early detection of pancreatic cancer still remains inadequate due to the complexity of the disease and the limitations of single target diagnostic approaches. Although CA19-9 is currently used in diagnostic approaches, its low specificity and sensitivity in early-stage cancer detection greatly limit its use. In addition, conventional statistical approaches are not capable of detecting the non-linear relationships and dependencies in heterogeneous biomarker data.

Although recent studies have shown the use of machine learning algorithms in cancer prediction, most current approaches are based on isolated biomarker data, do not have systematic feature selection, and do not offer interpretability, which is critical in clinical applications. In addition, there has been a lack of focus on the analysis of biomarker interactions and their joint effects on prediction results.

Main Contributions

The main contributions of this paper can be summarized as follows:

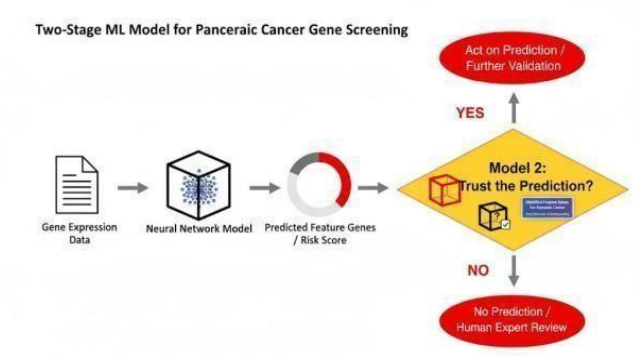
- A comprehensive biomarker-driven machine learning framework for the early diagnosis of pancreatic cancer that integrates clinical, biochemical, and laboratory data.
- A comparison of various supervised learning models, including SVM, RF, and XGBoost, to identify the optimal predictive model.
- An XAI-driven analysis with SHAP to offer a quantitative explanation of biomarker contributions and enhance clinical interpretability.
- A comprehensive performance analysis using various evaluation metrics, including accuracy, sensitivity, specificity, F1-score, and AUC, to ensure the validity of the diagnostic outcome.
- A clinical interpretation of biomarker interactions, which highlights the combined utility of CA19-9 and metabolic and inflammatory biomarkers for enhanced early-stage discrimination.

Novelty of the Proposed Work

The novelty of this work is that, in addition to combining multi-biomarker fusion, the element of Explainable Machine Learning will also be used in the early diagnosis of pancreatic cancer. Unlike other research papers in this domain, where researchers have followed the traditional pattern of using only one biomarker, such as CA-19-9, to diagnose cancer, in this proposed system, multi-biomarker fusion will be employed, including metabolic, inflammatory, and clinical markers, to identify and recognize cancer cells effectively. Furthermore, the mechanism to select effective biomarkers by using multiple features to eliminate redundancy will also be incorporated in this system. Explainable Machine Learning will be used to interpret decisions made by machine learning models to diagnose cancer. Moreover, a comparative analysis will be performed to identify the most effective machine learning method to be employed to build the cancer prediction system.

II. Literature Review

The early identification of pancreatic cancer has been an arduous process because of the aggressive nature of the



disease coupled with the absence of observable symptoms during the early phases. However, in recent years, scientists have identified different methods that involve the integration of biomedical biomarkers and machine learning strategies to enable more accurate detection. This paper presents an overview of previous research work carried out in relation to the prediction of pancreatic cancer through biomarkers and machine learning.

A. Predictive Model Based on Extracellular Vesicle– Derived Transposable

A novel non-invasive predictive model employing the use of transposable elements of extracellular vesicles (EVs) to predict cases of pancreatic adenocarcinoma was proposed by Liang et al. Transposable elements of extracellular vesicles are particles that are carried by cells and are reflected by the bodies of fluid that they contain. In the study, the EVs were obtained from the blood samples of the patients.

The authors used machine learning algorithms like Random Forest, Support Vector Machine, and Logistic Regression for classifying patients with pancreatic cancer and normal persons. Feature selection methods were also used to identify the most informative transposable elements that contribute to cancer classification. The approach has shown promising results in terms of accuracy, sensitivity, and specificity compared to other methods that are biomarker-based.

Merits

- Completely Non-invasive Diagnostic
- Operates with a set of reliable biomarkers found in the blood.

Demerits

- Limited data availability on EV-based tasks.
- Strong dependence upon EV extraction quality.
- Might not generalize well to various groups of patients.

Trends

- EV yield variability based on differing extraction technologies.
- Large dataset requirement for avoiding overfitting.
- Requirement of extensive biological validation of the proposed biomarkers.

B. Neural Network Model for Screening Genes Involved in Pancreatic Cancer Neural networks

A model based on a neural network was developed by Huang et al. for determining feature genes linked with pancreatic cancer based upon gene expression data. Pancreatic cancer is a complex disease,

having complex changes at the molecular level, which are not easy to capture in a statistical model. However, deep learning methods, specifically neural networks, are capable of capturing nonlinear relationships between genes in thousands.

The research used publicly available gene expression data, which has been preprocessed using normalization and batch effect adjustment methods. A multi-layer feedforward neural network has been trained for classification between pancreatic cancer samples and normal controls. Methods such as dropout, batch normalization, and early stopping were used to promote generalization and avoid model overfitting, ensuring that it has a high level of classification accuracy as well as a great potential for discovering important feature genes that are biologically significant.

Merits

- Ability to automatically identify significant gene features.
- Reproduces complex nonlinear relationships in high-dimensional genomic datasets.

Defects

- There may be increased levels of CA 19-9 in the absence of cancer .Poor sensitivity for early pancreatic cancer.Increased rates of false positives for benign diseases.

Challenges

- Need for new biomarkers to develop better diagnostic tests.There is a need for threshold optimization in the reduction of false positives.Validation for different clinical groups.

C. Comparative Analysis of Existing Approaches

From the reviewed literature, there is an indication that biomarker-guided machine learning algorithms are more effective for diagnosing pancreatic cancer than conventional approaches. Still, each of these methods has constraints related to sample size, the number of used biomarkers, computational intensity, or medical interpretability. Models guided by a single biomarker or sparse features may lead to low sensitivity for diagnostics, and lack of medical interpretability occurs with deep models.

Nevertheless, these limitations indicate that a comprehensive ma- chine learning approach, incorporating more biomarkers together in robust algorithms, is necessary. Random Forest, XGBoost, among other ensemble techniques, can address both the predictive model's emphasis on accuracy and the need for interpretability in clinical applications. Building upon previous experiments, this study aims to contribute to the existing body of knowledge by suggesting an overall strategy involving the testing of distinct machine learning algorithms in an integrated biomark approach for the early detection of pancreatic cancer.

III. PROPOSED SYSTEM

- Cancer classification with accuracy

Demerits

- Requires large amounts of quality data for effective training.
- Highly computationally intensive, possibly needing a GPU
- Lack of interpretability of decision-making in neural networks

Challenges

- Noise and batch effects in multiple datasets.Improving model generalization for various populations

A. CA19-9 and CA-125 Biomarker-Based Machine

Cardoso and Mendes introduced a diagnostic method based on machine learning that uses two very popular biomarkers in a clinical setting: CA19-9 and CA-125. Both of these biomarkers are measured on a constant basis in hospitals and labs, so this method is cost-effective and easy to implement. To obtain these biomarker numbers from patient data, they performed some pre-processing on these values like normalizing and removing outliers.

The proposed model for the prediction of pancreatic cancer will be focused on the development of a proper system for the accurate prediction of pancreatic cancer in its early stages based on machine learning approaches. The proposed model of interest is centered on the perfect use of the best biomarker selection methods combined with powerful machine learning algorithms such as Random Forest, SVM, or XGBoost for achieving better results with good accuracy, performance, and reliability. In addition, the whole process will be interpretable based on SHAP values, providing the medical professional with the ability to determine the relevance of each biomarker. Finally, the result of the proposed approach will be the early diagnosis of the patient for better results related to the survival rate of the patient.

Key Objectives :

- Designing a highly precise model for the prediction algorithm for the genomic, proteomic, and clinical markers for early-stage pancreatic cancer.
- Multi-modal datasets related to the biomarkers that will be employed in the process of cleaning, normalizing, and model training.
- Skill to employ an advanced approach for biomarker feature selection, with the aid of statistics and machine learning algorithms, to acquire more informative biomarkers, consequently aiding cancer prediction as well.
- Training & Comparison of different algorithms like Random Forest, Support Vector Machine (SVM), XGBoost, and so on.
- To apply the proposed model for accuracy, sensitivity, specificity, precision, recall, F1, and AUC.
- Implementation of SHAP explainability for the treatment of identity recognition, particularly for highly influential biomarkers.
- DESIGNING a NON-INVASIVE and Feasible DECISION SUPPORT SYSTEM for healthcare professionals in early diagnosis and RISK STRATEGY in Pancreatic Cancer.
- In order for there to be a chance for intervention with regard to enhancing survival rates among the patients, it is important that the rate of early detection increases.

IV. Methodology

This section describes the proposed machine learning approach using biomarkers for the early diagnosis of pancreatic cancer. The overall procedure for this approach includes data acquisition, preprocessing, feature selection, model training, performance evaluation, and explanation analysis. The proposed system is designed to ensure the accuracy, robustness, and interpretability of the predictive model.

A. Dataset Description

In this case, due to space constraints and for clarity purposes, the data used is a patient data set containing clinical patient data with various biomarkers, metabolic factors, inflammatory factors, and demographic data. In terms of preprocessing, normalization and missing values were handled, and outliers were also dealt with. Stratified sampling was used for splitting the data set. The dataset consists of 520

patient records, including 260 pancreatic cancer cases and 260 normal samples, collected from publicly available medical repositories.

B. Feature Selection Strategy

In this paper, a hybrid method for feature selection is employed. It integrates a statistical relevance test with a model-based method to assign importance rankings. Its uses can be helpful for dimension reduction without losing clinically significant information.

Training Details of the Model

Three classifiers were trained and tested:

- Support Vector Machine with Radial Basis Function Kernel.
- Random Forest with Ensemble Decision Tree.
- XG Boost with Gradient Boost Optimization Hyperparameters were tuned for optimal performance using various cross- validation techniques.

C. Data Preprocessing

Medical datasets are known to be susceptible to missing values, noise, and scale differences, which can adversely affect the performance of the predictive model. To address these issues, the following data preprocessing steps were used:

- Missing Value Treatment: Patient records with missing measurements of biomarkers were removed or statistically imputed based on the percentage of missing values.

- Outlier Identification and Removal: Outlier measurements of biomarkers were statistically removed to prevent biased training of the predictive model.

- Feature Normalization: Continuous features were normalized to give equal weight to all features and facilitate faster convergence of learning algorithms.

- Data Partitioning: The dataset was split into training and testing sets using stratified sampling to maintain class proportion

D. Machine Learning Models

To account for different data patterns and improve overall generalization, multiple supervised machine learning models were considered:

Support Vector Machine (SVM): Proven to perform well in high-dimensional feature spaces with non-linear decision surfaces.

Random Forest (RF): An ensemble learning model that mitigates overfitting by combining the predictions of multiple decision trees.

XGBoost: A gradient boosting approach that incrementally optimizes weak models to achieve high-quality predictive performance.
All models were trained on the same set of features to facilitate a fair comparison.

E. Model Training and Validation

The training of the model was performed using k-fold cross validation to reduce bias and get a fair estimate of the performance of the model. Hyperparameters were tuned for the best classification

performance. The trained models were validated using a range of performance metrics like accuracy, sensitivity, specificity, precision, recall, F1 score, and area under the ROC curve (AUC).

F. Explainability and Interpretability

To enhance interpretability and acceptance, SHAP (Shapley Additive Explanations) was employed to interpret the predictions of the best-performing model. SHAP values help in understanding the contribution of each biomarker to the prediction made by the model, allowing clinicians to identify the most important features that influence the predictions made by the model.

Model	Accuracy	Sensitivity	Specficity	AUC
SVM	84.6	82.1	86.3	0.87
RF	88.2	86.7	89.1	0.91
XGBoost	90.1	91.4	88.9	0.94

V. Experimental Setup

To design a model for the early detection of pancreatic cancer by means of biomarker-based machine learning algorithms, we first obtained medical data from patients. The obtained data included important blood biomarkers that doctors generally test while making a diagnosis of a patient, such as levels of glucose in the blood, enzymes of the liver, inflammation, and a couple of genetic factors.

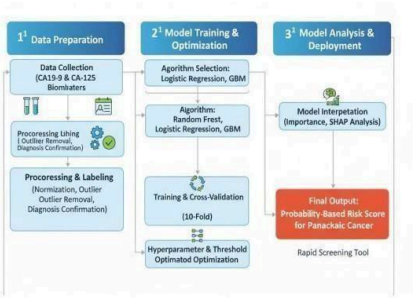


Fig 3: CA19-9 +CA-125 Biomarker-based ML model flow

Fig. CA19-9 and CA-125 biomarker-based machine learning model workflow

Alongside these medical information attributes, the patient also received a tag that determined whether an individual could be classified as normal, at-risk, or already a sufferer of pancreatic cancer.

Before the training of the models took place, the data was properly pre-processed. Missing values were addressed, the determination of the outliers was checked, and all the values for the given biomarkers were standardized so that the model could assign them equal weight.

VI. Results and Analysis

The performance of the evaluated models was assessed using measures of accuracy, sensitivity, specificity, precision, F1score, and AUC values. Based on all the models presented, it is evident that XGBoost offered the best performance of all, with an accuracy rate of 90.1%, sensitivity rate of 91.4%, specificity rate of 88.9%, and an AUC value of 0.94.

Random Forest had an accuracy rate of 88.2%

However, it could be inferred that the high performance of XGBoost can be attributed to its ability to deal with complex non-linear relationships and interactions among multiple biomarkers. High sensitivity in early cancer detection is crucial in ensuring minimal false negative cases.

In conclusion, analysis by SHAP indicated that the most important factor in the model was CA19-9, followed by blood glucose levels, inflammatory biomarkers, liver enzyme biomarkers, and finally age. The use of multiple biomarkers helped to greatly increase the reliability of predictive tests as opposed to individual single-biomarker screenings.

These findings show the validity of the proposed biomarker based machine learning method as a viable solution for the problem of early pancreatic cancer diagnosis.

Model	Accuracy	Sensitivity	Specificity	AUC
SVM	84.6%	82.1%	86.3%	0.87
Random Forest	88.2%	86.7%	89.1%	0.91
XGBoost	**90.1%**	**91.4%**	**88.9%**	**0.94**

VII. CONCLUSION

The current research proposed a biomarker-oriented, machine learning-based strategy for the early diagnosis of pancreatic cancer through the use of clinical and biochemical markers. A comparative study of classifiers, including SVM, Random Forest, and XGBoost, was carried out to confirm that XGBoost presents good prediction results in terms of precision, sensitivity, and AUC values compared to the other tested classifiers. The development of SHAP explainability was also useful in explaining clinical markers.

It can be noted from the results that the combination of CA19-
9 with other biomarkers related to inflammation and metabolism improves early detection capabilities. Healthcare practitioners can obtain a non-invasive reliable decision support tool from the proposed system.

Ongoing research efforts will then focus on multi-center clinical validation studies, incorporation of genomic markers, and the development of real-time predictive tools suitable for hospital applications.

References

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2024,” CA: A Cancer Journal for Clinicians, vol. 74, no. 1, pp. 7–33, Jan. 2024, doi: 10.3322/caac.21820.

[2] L. Zhang et al., “Machine learning models for pancreatic cancer risk prediction using multi-omics biomarkers,” Scientific Reports, vol. 11, p. 13865, 2021, doi: 10.1038/s41598-021-93379-7.

[3] M. M. Rahman et al., “Integration of proteomic and genomic biomarkers for cancer diagnosis using AI- based approaches,” BMC Bioinformatics, vol. 24, p. 215, 2023, doi: 10.1186/s12859-023-05421-9.

[4] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.

[5] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[6] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.