I.J. Information Technology and Computer Science, **2025, 6, 127-142**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.06.07

Pancreatic Cancer Prediction Using Machine Learning: An Investigation of Different Algorithms

Radha Singh Jadaun
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India
E-mail: radhasingh16012002@gmail.com
ORCID iD: https://orcid.org/0009-0002-9287-7370

Nij Mehar Grover
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India
E-mail: nijmehar16@gmail.com
ORCID iD: https://orcid.org/0009-0002-7103-0808

K. Srinivas
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India
E-mail: ksri12@gmail.com
ORCID iD: https://orcid.org/0009-0002-3884-6282

A. Charan Kumari*
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India
E-mail: charankumari@dei.ac.in
ORCID iD: https://orcid.org/0000-0002-3160-1912
*Corresponding author

Received: 14 July 2025; Revised: 06 September 2025; Accepted: 02 October 2025; Published: 08 December 2025

Abstract: Pancreatic cancer, characterized by its high mortality rate and scarce treatment options, poses a formidable challenge in the field of oncology. Now, we live in a reality that requires immediate progress in diagnostic and prognostic methodologies to find pancreatic cancer early and understand its stage. This study deals with the pressing requirement for better diagnostic tools by evaluating and deciding the suitable machine learning (ML) algorithms for detecting pancreatic cancer at an early stage. This work uses a publicly available dataset with 590 urine samples which included control, benign hepatobiliary disease as well as Pancreatic Ductal Adenocarcinoma (PDAC) samples. The primary objectives of the research included developing a predictive model based on clinical data, examining various machine learning (ML) algorithms for their diagnostic precision, and improving the early detection rates for pancreatic cancer. The study assessed the efficacy of a broad array of ML algorithms in forecasting outcomes associated with pancreatic cancer. This analysis systematically explored Random Forest, Support Vector Machine, Decision Trees, K-Nearest Neighbours, XGBoost, ADABoost, CatBoost, and GradientBoost. The assessment focused on standard performance metrics such as accuracy, precision (also known as positive predicted value or PPV), recall (sometimes called sensitivity or true positive rate), F1-score, and support. Notably, CatBoost achieved the highest accuracy of 75%, outperforming other models such as Random Forest (74%) and XGBoost (74%), demonstrating its superior classification performance in distinguishing between pancreatic cancer, benign conditions, and non-cancerous cases. In addition to performance evaluation, this study integrates SHAP (Shapley Additive Explanations) analysis to enhance model interpretability, ensuring transparency in feature contributions. SHAP analysis revealed that Plasma CA19-9, LYVE1, and TFF1 were the most influential biomarkers across all classifications, reinforcing their diagnostic significance. This research emphasizes the critical importance of early detection, model interpretability, and clinical applicability, demonstrating that ML algorithms, particularly CatBoost, not only enhance diagnostic precision but also provide explainable predictions that support real-world medical decision-making.

Index Terms: Pancreatic Cancer, Machine Learning Algorithms, Biomarkers, Oncology, Benign Hepatobiliary.

1. Introduction

The pancreas, a crucial accessory organ of the human body, functions dually as both an endocrine and exocrine gland. In a healthy adult, it typically weighs around 100g and ranges from 14 to 25 cm in length [1]. Its structure is lobular and elongated, consisting of five distinct anatomical segments: the head, uncinate process, neck, body, and tail. Acting as an exocrine gland, its main function is to secrete digestive enzymes that assist in the breakdown and absorption of nutrients, particularly fats. Concurrently, as an endocrine gland, it plays a vital role in regulating the body's blood sugar levels and ensuring efficient nutrient absorption by cells. Disorders like diabetes and pancreatitis arise from malfunctions in the human pancreas, which is responsible for maintaining glucose balance and aiding in digestion. Recent researches have highlighted the novel roles of a local renin-angiotensin system (RAS) in the pancreas and its clinical relevance; its inappropriate activation leads to pancreatic endocrine and exocrine disease, notably type 2 diabetes [2].

Pancreatic cancer is among the deadliest diseases related to the pancreas, where cells grow uncontrollably and quickly in this organ. Because it is very aggressive and usually detected at later stages, pancreatic cancer often has a small rate of survival. This underlines how important it is to find and predict pancreatic cancer early on to help patients more effectively and give them timely medical treatments. In the present time, Machine Learning (ML) is showing great promise in medical research and diagnostics. It could play a major role in improving predictive models for pancreatic cancer to better detect it at an early stage.

The term "pancreatic cancer" is usually talking about exocrine pancreatic cancer. Pancreatic cancer is more common in Black Americans than any other racial or ethnic group. They have the highest rate of getting this type of cancer in the United States, according to National Cancer Institute SEER data (SEER stands for Surveillance, Epidemiology, and End Results program). People who are at a higher risk should pay careful attention to their bodies. They need to be aware if they experience any symptoms that are out of ordinary and express it openly with their healthcare team. Some examples include back or stomach pain, unexplained weight loss, yellow skin colour (jaundice), as well as problems related to digestion such as having trouble digesting food properly - all these signs could be indicative of pancreatic cancer. The percentage of Black Americans taking part in clinical trials for treating cancer is very low even though they make up 13% of the total population in America; currently only 4% participate accordingly [3].

Pancreatic cancer arises when there are abnormal changes or mutations in the DNA of pancreatic cells. These changes make the cells grow and divide without control, which results in a tumour forming. This tumour may spread or metastasize to other parts of the body such as liver, abdominal wall, lymph nodes, lungs, or bones. Certain risk elements can boost one's chance to get pancreatic cancer. Factors that contribute to developing pancreatic cancer are smoking, obesity, diabetes I that lasts quite long, strong family background with the disease, consuming processed food, and red meat in diet excessively high as well as having a condition termed chronic pancreatitis. Pancreatic adenocarcinoma is a serious disease caused by abnormal cells in the pancreas that divide and grow out of control, forming a tumour [4]. Pancreatic cancer holds the position of being fourth most common reason for people dying from cancer in the western part of Earth, making it an important health issue.

Cancer survival rates can be improved by detecting it early on. Certain types of cancer, such as pancreatic cancer, are difficult to diagnose or catch at an early stage and their stages progress quickly. This paper presents the most modern methods for predicting survival in cases of cancer, making a proposal on how these techniques could be applied to forecast the overall life expectancy among patients having pancreatic ductal adenocarcinoma (PDAC) cancer. Due to the confusion and large data sets, recent studies have emphasized on machine learning (ML) algorithms such as support vector machines and convolutional neural networks. The predictions of studies are that the survival rate for pancreatic ductal adenocarcinoma cancer is within limits of 41.7% at one year, 8.7% at three years, and 1.9% after five years with no significant correlation found between disease stages and overall survival rate [5]. Techniques of prediction like ML, which encompass statistical multivariate regression and deep learning (DL), could be employed for forecasting the survival rate of pancreatic ductal adenocarcinoma (PDAC). Nevertheless, it is necessary to fine-tune and optimize these techniques for achieving optimal prediction abilities.

The pancreas is located within the abdominal cavity, positioned horizontally across the upper abdomen behind the stomach. It is in proximity to significant organs such as the small intestine, liver, and gallbladder. Within the digestive system, the pancreas performs a crucial role by producing enzymes for food breakdown in the small intestine and releasing hormones, including insulin, which regulates blood sugar levels.

The primary contributions of this research include:

- The development of a predictive model for pancreatic cancer utilizing extensive clinical data, significantly improving early detection and diagnosis.
- Utilization of various machine learning algorithms like Support Vector Machine (SVM), Random Forest, K-Nearest Neighbours, Decision Trees, XGBoost, ADABoost, GradientBoost, and CatBoost for the diagnosis of pancreatic cancer. This analysis provides insights into the effectiveness of these algorithms.
- Emphasis on the significance of early detection in enhancing pancreatic cancer prediction rates and demonstrating the transformative impact of machine learning in diagnostics.

- Demonstration of the efficacy of machine learning techniques through standard performance metrics, establishing a benchmark for future research.
- Exploration of the correlation between support metric and model performance, particularly in dealing with imbalanced datasets, underscoring its importance in creating equitable healthcare models.

The rest of the paper is organized as follows. Section 2 presents the research work reported in the literature for pancreatic cancer prediction. The methodology adopted for this research is described in section 3. Section 4 presents the results, section 5 presents explainability analysis of CatBoost classifier using SHAP values and section 6 concludes.

2. Literature Review

In recent years, there has been a growing utilization of Machine Learning (ML) in predicting cancer, including pancreatic cancer. The interest in employing ML technology stems from its capacity to analyze various data types such as genomic, proteomic, and clinical data, enabling the identification of patterns and relationships associated with pancreatic cancer. An examination of previous studies demonstrates the evolving role of ML in cancer prediction, with a particular emphasis on early detection, especially in pancreatic cancer cases. ML algorithms have been deployed across a range of datasets, including patient records, medical imaging data, and genetic information, providing valuable insights that could potentially reveal patterns indicative of early-stage pancreatic cancer when processed through ML algorithms. The enhancement of specific algorithms like support vector machines, neural networks, and ensemble methods has exhibited promising results in the early detection of pancreatic cancer [6].

A comprehensive analysis of research literature on the prediction of pancreatic cancer highlights a collective emphasis on early detection utilizing various contemporary technologies and algorithms. These technologies encompass Random Forest, SVMs, K-Nearest Neighbours (KNN), Decision Trees, among others. The field of pancreatic cancer prediction and diagnosis has made substantial progress in recent years. Noteworthy research conducted by Bakasa and Viriri explored diverse machine learning techniques, including advanced frameworks such as TensorFlow and PyTorch. This study employed the TCIA Public Access dataset, focusing on 3D CT scans to predict the overall survival of pancreatic cancer patients. Additionally, the U-Net model, renowned for its accuracy, played a pivotal role in image processing tasks within this study.

Davide Placido *et al*. [7,8] investigated into disease trajectories utilizing various deep learning algorithms, introducing a practical prediction-surveillance process through machine learning. By leveraging DNPR disease trajectories and CPR36 demographic data, this investigation identified high-risk patients and forecasted the progression of pancreatic cancer. Another study by Bahrudeen and Krishnan [9] explored the integration of artificial neural networks and diverse imaging modalities to enhance cancer diagnosis. Despite the absence of specific data utilization, the study provided valuable insights into harnessing advanced AI techniques for precise diagnostic purposes.

Furthermore, H.S. Saraswathi and Mohamed Rafi [10] introduced a novel hybrid KNN-IM approach for managing missing values, which exhibited superior performance compared to existing methods, demonstrating its potential to enhance predictive accuracy in pancreatic cancer diagnosis. Additional research by Haidy Nasief *et al.* [11] concentrated on evaluating treatment response using weekly DRFs data. This study effectively employed a linear regression model and delta-radiomic process to anticipate treatment outcomes in patients with pancreatic cancer. Additionally, Seiya Yokoyama [12] conducted hierarchical clustering analysis on mRNA expression levels to identify clusters with poor prognoses, offering valuable insights into patient outcomes based on molecular characteristics.

Several investigations have made significant progress in the prediction and prognosis of pancreatic cancer. Faur *et al.* [13] collected a total of 97 studies on pancreatic neoplasms. Similarly, Osman [14] analyzed data from 14,631 patients, considering variables such as age, tumor size, surgery, and tumor extension. Through the application of the Random Forest algorithm, various performance metrics of the machine learning model, including AUC, Precision, Accuracy, and Sensitivity, were assessed over different survival periods. The study emphasized the potential of machine learning in predicting survival, assisting in treatment decisions and care planning.

Pancreatic cancer, despite its challenging nature in detection and treatment, remains a significant concern in the field of oncology due to its aggressive characteristics and limited therapy options. Typically diagnosed in advanced stages, this disease poses a challenging task for clinicians and researchers alike. Nonetheless, recent advancements in machine learning methods have brought optimism towards the prediction of pancreatic cancer. These methods facilitate early detection, potentially enhancing patient outcomes and survival rates. In the oncology domain, pancreatic cancer represents a daunting obstacle, often identified at late stages, and linked to restricted treatment choices. Nevertheless, the recent progress in employing Machine Learning (ML) techniques has demonstrated potential in predicting pancreatic cancer. These approaches support early detection, potentially improving patient outcomes and survival rates. Through the utilization of ML capabilities, researchers and clinicians may be able to combat this aggressive disease more effectively, offering prospects for enhanced detection, treatment, and patient results.

3. Methodology

The methodology outlined here presents a systematic and structured approach to predicting pancreatic cancer.

Through careful planning and rigorous execution, this methodology aims to utilize various tools and techniques to achieve precise and dependable predictions regarding pancreatic cancer prognosis. By employing an interdisciplinary framework that integrates data analysis, machine learning algorithms, and clinical knowledge, this methodology seeks to enhance comprehension of pancreatic cancer and facilitate early detection and intervention strategies. This paper sets the foundation for a detailed examination of the methodology employed in predicting pancreatic cancer outcomes.

The employed methodology comprises several key steps:

- Firstly, the process begins with Data Collection and Preprocessing, which involves gathering pertinent datasets containing various patient details related to pancreatic cancer, such as age, sex, medical history, and genetic markers. The data is then cleansed, missing values are addressed, and features are normalized or standardized to ensure accuracy and uniformity.
- Next, Model Selection is carried out by selecting a diverse range of machine learning algorithms suitable for predictive modeling, including traditional ones like decision trees, SVM, and random forest, as well as more advanced techniques such as boosting methods.
- Subsequently, Evaluation Metrics are defined to gauge the performance of each algorithm. Common metrics, like accuracy, precision, F1-score, support, and recall, are used for classification tasks in cancer prediction.
- Lastly, Performance Comparison is conducted to assess the accuracy and reliability of algorithms using various evaluation metrics, aiding in the identification of the most precise and dependable models.

By adhering to this structured framework, individuals can effectively assess and pinpoint the machine learning algorithm or combination of algorithms that yield the most precise outcomes in predicting pancreatic cancer. This process contributes to the advancement of the field by shedding light on the most promising methods for clinical applications, ultimately enhancing diagnostic accuracy and patient care.

3.1. Data Set

In this research, a comprehensive and representative dataset has been compiled, incorporating patient histories and genetic profiles. This dataset is pivotal for predicting pancreatic cancer stages in patients. It was acquired from multiple institutions, including University College London, Spanish National Cancer Research Center, Barts Pancreas Tissue Bank, Cambridge University Hospital, University of Liverpool, and the University of Belgrade. The dataset [15] used in this study consists of 590 urine samples collected from different patient groups, including individuals diagnosed with Pancreatic Ductal Adenocarcinoma (PDAC), patients with benign hepatobiliary diseases, and those with no pancreatic disease. It includes 14 attributes covering demographic details, biochemical markers, and clinical information relevant to pancreatic cancer diagnosis. The demographic features considered are age and sex, while the biochemical markers include Plasma CA19-9, a commonly used biomarker for pancreatic cancer detection, along with Creatinine (a kidney function marker for biomarker normalization), LYVE1 (a lymphatic marker linked to tumor progression), REG1B (associated with pancreatic disease), TFF1 (linked to gastrointestinal cancers), and REG1A (another relevant biomarker for pancreatic cancer). Additionally, the dataset includes clinical information, such as the stage of the disease, benign sample diagnosis, and the diagnosis label, which classifies samples into "Pancreatic Cancer," "Benign Hepatobiliary Disease," or "No Pancreatic Disease." This dataset provides a comprehensive representation of clinical and biochemical markers, supporting the development of machine learning models for early pancreatic cancer detection.

The dataset provides a diverse representation of diagnostic categories, ensuring a comprehensive evaluation of the machine learning models. It includes many samples from individuals with no pancreatic disease, a moderate representation of cases with benign hepatobiliary disease, and a significant subset of samples from patients diagnosed with Pancreatic Ductal Adenocarcinoma (PDAC). This distribution allows the study to assess model performance across different patient groups, highlighting its potential for early pancreatic cancer detection while ensuring robust classification across varied clinical conditions. This study maintains the real-world distribution of diagnostic classes, ensuring that models learn from naturally occurring case proportions. While pancreatic cancer cases are underrepresented, this approach enhances the model's applicability to real clinical settings.

3.2. Data Preprocessing

Data preprocessing played a crucial role in maintaining the integrity and quality of the dataset, ultimately enhancing the accuracy and reliability of the predictive models. A key aspect of this process involved addressing missing data to minimize bias and ensure the robustness of the findings. Various imputation techniques were employed based on the characteristics and distribution of missing values. Mean substitution was applied to features with a limited number of missing values to maintain consistency without introducing significant distortions. For features exhibiting strong correlations with other variables, regression imputation was utilized to estimate missing values based on existing data relationships. Additionally, multiple imputation methods were implemented for features with complex missing patterns, particularly for biomarkers such as REG1A and plasma CA19-9, ensuring that critical information was preserved while mitigating potential data loss.

In addition to handling missing values, feature scaling and normalization were applied to standardize the numerical range of different attributes and prevent any individual feature from disproportionately influencing the model's predictions. The choice of scaling method was determined based on the nature of the feature. Min-Max scaling was employed for

continuous biomarker variables, including LYVE1, REG1B, TFF1, and REG1A, to normalize their values within a fixed range of 0 to 1, ensuring uniformity across all samples. Conversely, Z-score standardization was applied to variables such as age, transforming them to have a mean of 0 and a standard deviation of 1, facilitating better comparability across different scales. By implementing these preprocessing techniques, the dataset was effectively optimized for machine learning algorithms, ensuring equitable feature contributions, and reducing the risk of bias introduced by missing data or variations in numerical scales.

3.3. Model Development

During the model development phase, the study delved into a variety of Machine Learning (ML) algorithms to build resilient predictive models for pancreatic cancer. Algorithms such as Random Forest, SVM, Decision Trees, KNN, XGBoost, ADABoost, CatBoost, and GradientBoost were implemented, each with distinct strengths and weaknesses that varied in effectiveness depending on the context. In this study, all the models were trained using their default hyperparameter settings, ensuring a baseline evaluation of their performance without manual tuning. This approach provides an unbiased assessment of how each algorithm performs under standard conditions. By conducting thorough performance evaluations, the research identified the most suitable algorithm tailored to the specific prediction task.

3.4. Model Evolution

To evaluate the performance of the machine learning models comprehensively, the following key metrics were considered: accuracy, precision, recall, F1-score, support, and confusion matrix. These metrics provide a detailed assessment of model effectiveness, particularly in medical diagnostics, where both false positives and false negatives have significant consequences.

A. Accuracy

Accuracy is defined as the proportion of accurately identified instances to all instances as shown in equation 1.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{1}$$

B. Precision

The precision metric, as described in equation 2, quantifies the percentage of true positive predictions among all the positive predictions generated by the model. It shows that the model can prevent false positives.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2}$$

C. Recall (Sensitivity)

The percentage of true positive predictions among all actual positive data instances is known as recall. It illustrates how the model may prevent false negatives by capturing all positive cases as depicted in equation 3.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

D. F1-Score

The F1-score, as described in equation 4, is the harmonic mean between recall and precision. It offers an equitable assessment of a model's effectiveness, particularly in cases where the dataset exhibits class imbalances.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

E. Support

Represents the number of actual occurrences of each class in the dataset. It helps in understanding whether performance metrics are calculated on a balanced number of samples or if certain classes dominate the dataset.

F. Confusion Matrix

A tabular overview of the model's predictions compared to the actual class labels is given by a confusion matrix. It allows for a more in-depth analysis of the model's performance, showing the counts of true negatives, true positives, false negatives, and false positives.

3.5. The Algorithms

This subsection describes the Machine Learning algorithms implemented in this work.

A. Random Forest

The Random Forest approach aims to enhance accuracy by combining the results of multiple decision trees. Each tree is trained on a subset of the original data independently, with their predictions aggregated through a voting mechanism. This technique effectively reduces overfitting, thereby improving performance on new challenges and enhancing the model's ability to generalize for regression and classification tasks in diverse domains. Essentially, Random Forest comprises a collection of decision trees that collectively determine the final output via majority voting. This algorithm stands out for its superior performance compared to individual tree classifiers and its robustness against noise. Moreover, the efficacy of the Random Forest Algorithm scales with the dataset size, showing better performance with larger datasets than with smaller ones.

B. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a potent supervised learning model known for its effectiveness in both classification and regression analyses. It operates by identifying a hyperplane that can accurately segregate classes based on data points while maximizing the margin between points of different classes. Particularly adept at handling high-dimensional data spaces, SVM leverages a kernel trick to explore nonlinear relationships among variables. Central to SVM is the concept of structural risk minimization (SRM), where the algorithm's objective is to discern distinct classes within training samples and determine the optimal separating hyperplane that efficiently distinguishes them—referred to as the "optimal separable hyperplane." Beyond classification and regression, SVM has demonstrated utility in diverse domains such as clustering and time series analysis.

C. Decision Trees

Decision Trees are machine learning models extensively utilized for classification and regression tasks, functioning by recursively partitioning data based on different features to form a tree-like structure. The internal nodes of the tree correspond to decisions made on features, while the leaf nodes provide predicted outcomes. Named for their hierarchical tree-like layout, decision trees feature internal nodes representing dataset features, branches showing outcomes of various tests, and leaf nodes offering final predictions. Through recursive partitioning, the dataset is segmented into subsets primarily composed of data points belonging to specific classes. At each internal node, the optimal split decision is made based on impurity measures. The decision tree classifier involves two key stages: tree construction, where the dataset is partitioned to ensure distinct representation of features, and tree pruning, which aims to optimize the tree structure to mitigate overfitting. Decision trees are highly favored for their efficiency and versatility across a wide range of applications.

D. K-Nearest Neighbours (KNN)

KNN or K-Nearest Neighbours is an uncomplicated and understandable algorithm that can be employed for classification as well as regression. In the case of K-nearest neighbours, they determine the class by counting which one is in majority within their range. When it comes to regression, KNN gives mean value of the k-neighbours. KNN's most important characteristic is its non-parametric nature. It does not rely on assumptions about data distribution like other algorithms, but instead uses closeness of data points in feature space as the foundation for making decisions. This helps KNN to identify patterns in the given data. KNN is an illustration of supervised learning. It applies similarities between fresh observation and known labels to carry out classification tasks. Through the process of examining training examples inside a certain region, it dedicates the new observation into a suitable category so that accurate classification becomes feasible for KNN.

E. XGBoost

XGBoost, which is the short form for Extreme Gradient Boosting, has gained popularity in machine learning because of its fast and high-quality functioning. Working on a method named gradient boosting, it uses a structure where weak learners - often decision trees - are trained one by one to fix mistakes from earlier models. XGBoost makes use of optimization methods to lessen certain loss functions; for example: log loss in classification or squared error when working with regression tasks. For stopping overfitting and promoting generalization, it applies different regularization methods such as L1 and L2 regularization. Furthermore, XGBoost gives useful knowledge about feature importance. This lets the users understand how much each feature contributes comparatively to the model's predictions.

F. CatBoost

CatBoost, which is a boosting algorithm that has strong power in dealing with gradient, shows superiority in its handling of categorical features. It does not require manual preprocessing steps like one-hot encoding as seen with other algorithms. This algorithm also has good protection against overfitting and it works quite well. This makes CatBoost appropriate for different uses. Furthermore, CatBoost handles feature scaling automatically. This means users do not have to do any data preprocessing for model training. The good performance of CatBoost on small and big datasets confirms its reliability.

G. AdaBoost (Adaptive Boosting)

AdaBoost, which is also called Adaptive Boosting, applies a method of sequential learning. In this method, there are weak learners that are trained one after another to correct mistakes made by the preceding models. AdaBoost assigns different weights to data points using instance weighting. The weights are chosen based on how accurately an instance has been classified; instances that were misclassified receive more weight in the following iterations. The final prediction is decided by combining all weak learner's predictions together and each learner contributes according to its performance level. AdaBoost shows a flexible character because it can cooperate with various base classifiers, and decision trees are often selected as these are simple yet effective.

H. Gradient Boosting

Gradient Boosting is known for its step-by-step process to improve model ability, which is done by reducing a set loss function such as mean squared error or cross-entropy loss. By using residual learning, each next model (also called weak learner) gets trained to fix the mistakes made by previous model group, thus getting better at predicting accuracy. This method makes it possible for handling complicated data relationships and adjusting well with both linear as well as non-linear patterns in Gradient Boosting. In the end, the last forecast is produced by adding together predictions from all weak learners. This is commonly done using simple average or weighted average methods.

The selection of machine learning algorithms for this study was based on their established effectiveness in medical diagnostics and classification tasks, particularly in handling structured clinical data. We considered a diverse set of models, including tree-based, distance-based, and ensemble learning approaches, to ensure a comprehensive evaluation of different learning paradigms.

- Tree-based models (Decision Trees, Random Forest, XGBoost, ADABoost, CatBoost, Gradient Boosting): These models are widely used in medical diagnosis due to their ability to handle complex feature interactions, interpretability, and robustness against missing data. Ensemble models such as XGBoost and CatBoost were included due to their enhanced predictive performance in structured datasets.
- Support Vector Machine (SVM): SVM is known for its strong classification capabilities, particularly in high-dimensional spaces, making it suitable for medical applications where features may have nonlinear relationships.
- K-Nearest Neighbors (KNN): KNN was included as a baseline model due to its simplicity and ability to capture local patterns in data, providing a comparative measure against more complex algorithms.

By including a mix of traditional classifiers, ensemble learners, and boosting techniques, we aimed to assess their relative strengths in pancreatic cancer prediction and determine the most suitable approach for early diagnosis.

4. Results and Analysis

This section reports the results obtained by various algorithms and compares their efficacy in the prediction process. The accuracies obtained by the implemented algorithms are listed in Table 1.

Table 1. Comparison of accuracies of the implemented algorithms

Algorithm	Accuracy
Random Forest	0.74
SVM	0.59
Decision Trees	0.62
KNN	0.67
XG Boost	0.74
Cat Boost	0.75
ADA Boost	0.61
Gradient Boosting	0.72

While CatBoost demonstrated the highest accuracy (75%), its performance was closely followed by Random Forest and XGBoost, both achieving 74%. These results indicate that ensemble learning methods, particularly boosting-based models, excel in handling complex feature interactions and improving classification accuracy. However, other algorithms such as Gradient Boosting (72%) and KNN (67%) also showed competitive performance, highlighting their potential utility in specific scenarios. Simpler models like Decision Trees (62%) and AdaBoost (61%) had comparatively lower accuracies, primarily due to their inherent limitations in feature representation and sensitivity to data distribution.

The performances of the implemented algorithms are presented below.

4.1. Random Forest

Based on the report of Random Forest Classifier as presented in Table 2 and its corresponding confusion matrix presented in Fig. 1, it is evident that the model performs reasonably well across the different classes. Class 3 exhibits the highest precision, recall, and F1-score values, indicating robust classification performance for this category. However, there are slight discrepancies in the precision and recall metrics for Classes 1 and 2, suggesting room for improvement in the classifier's ability to accurately predict instances belonging to these classes. Overall, while the model demonstrates effectiveness, further refinement may enhance its performance, particularly for Classes 1 and 2, ensuring more accurate classification across all categories.

Table 2. Random forest classifier report

Classes	Precision	Recall	F1-Score	Support
1	0.71	0.73	0.72	62
2	0.67	0.60	0.63	63
3	0.84	0.92	0.88	52

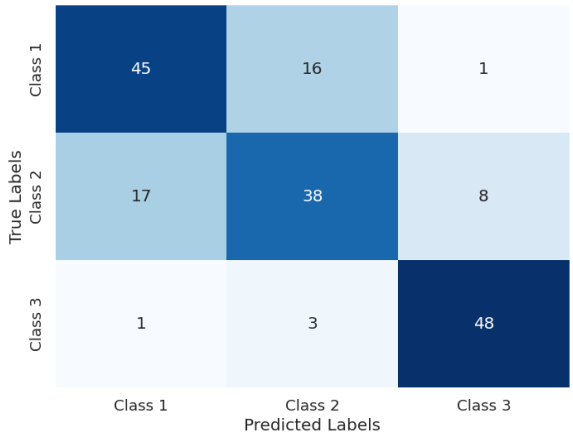

Fig.1. Confusion matrix of random forest

Table 3. SVM classifier report

Classes	Precision	Recall	F1-Score	Support
1	0.55	0.84	0.67	62
2	0.49	0.32	0.38	63
3	0.79	0.63	0.70	52

Fig.2. Confusion matrix of SVM

4.2. Support Vector Machine

The assessment of the SVM Classifier in Table 3 and confusion matrix in Fig. 2 reveals diverse performance metrics across different classes. While Class 3 showcases the highest precision and F1-score, Classes 1 and 2 exhibit relatively lower precision and recall values. Specifically, Class 1 shows higher recall but lower precision, indicating that the classifier correctly identifies more Class 1 instances but also misclassifies some instances from other classes as Class 1. On the other hand, Class 2 demonstrates lower recall, suggesting that the classifier misses a significant number of Class 2 instances. Overall, the SVM Classifier's performance highlights the significance of balancing precision and recall across all classes to achieve optimal classification accuracy.

4.3. Decision Trees

The analysis in Table 4 and confusion matrix in Fig. 3 depicts the model's performance across different classes. Notably, Class 3 demonstrates the highest precision, recall, and F1-score, signaling strong performance in accurately identifying instances of this class. However, Classes 1 and 2 show comparatively lower precision and recall values, indicating challenges in correctly classifying instances for these classes. Particularly, Class 2 exhibits the lowest precision and recall among the three classes, pointing out areas for improvement in distinguishing Class 2 instances. While the Decision Tree Classifier shows strengths in specific classes, enhancing its performance across all classes is crucial for achieving comprehensive classification accuracy.

Table 4. Decision trees classifier report

Classes	Precision	Recall	F1-Score	Support
1	0.62	0.73	0.67	62
2	0.51	0.44	0.47	63
3	0.76	0.73	0.75	52

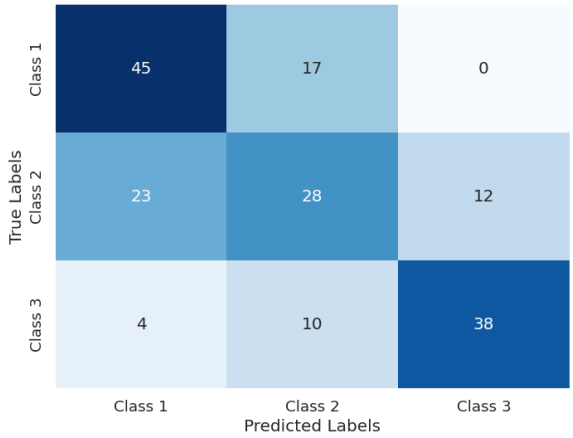

Fig.3. Confusion matrix of decision tree

4.4. K-Nearest Neighbours

The evaluation of the K-Nearest Neighbors (KNN) Classifier is provided in table 5 and its corresponding confusion matrix in Fig. 4. A detailed analysis reveals that Class 3 demonstrates the highest precision, recall, and F1-score, showcasing the model's ability to precisely identify instances belonging to this category. While Classes 1 and 2 show moderate precision, recall, and F1-score values, they exhibit some variability in performance compared to Class 3. Specifically, Class 1 displays slightly better precision and recall than Class 2, indicating superior classification performance for Class 1 instances. Overall, the KNN Classifier's performance emphasizes its effectiveness in distinguishing between different classes, with a strong capability observed in correctly classifying Class 3 instances.

Table 5. KNN classifier report

Classes	Precision	Recall	F1-Scoe	Support
1	0.63	0.74	0.68	62
2	0.62	0.62	0.62	63
3	0.85	0.67	0.75	52

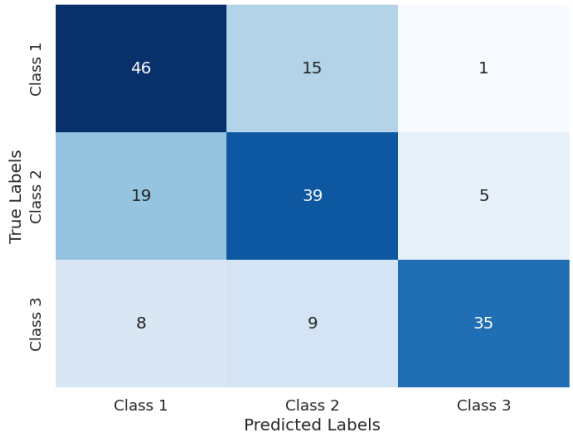

Fig.4. Confusion matrix of KNN

4.5. XGBoost

The assessment in Table 6 and confusion matrix in Fig. 5 for the XGBoost Classifier offers insights into its performance metrics across various classes. Notably, Class 3 demonstrates the highest precision, recall, and F1-score, showcasing the model's proficiency in accurately identifying instances of this class. Classes 1 and 2 also exhibit respectable precision, recall, and F1-score values, although slightly lower than Class 3. Specifically, Class 1 and Class 3 show similar precision and recall values, while Class 2 displays slightly lower values. Overall, the XGBoost Classifier displays robust performance across all classes, with notable strengths in correctly classifying Class 3 instances.

Table 6. XGBoost classifier report

Classes	Precision	Recall	F1-Score	Support
1	0.76	0.76	0.76	62
2	0.67	0.62	0.64	63
3	0.81	0.88	0.84	52

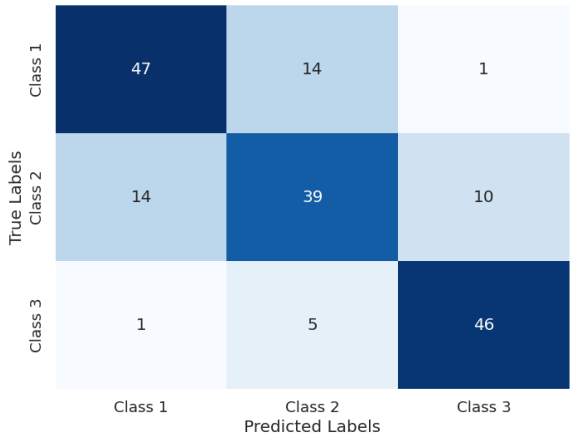

Fig.5. Confusion matrix of XGBoost

4.6. CatBoost

The classification report for the CatBoost Classifier in Table 7 and its confusion matrix in Fig. 6, highlights its performance metrics across different classes. Notably, Class 3 exhibits the highest precision, recall, and F1-score values, indicating the model's proficiency in accurately identifying instances belonging to this class. Classes 1 and 2 also demonstrate respectable precision, recall, and F1-score values, although slightly lower than Class 3. Specifically, Classes 1 and 2 showcase similar precision and recall values, while Class 3 displays higher values, indicating its strength in classification. Overall, the CatBoost Classifier demonstrates robust performance across all classes, with notable strengths observed in correctly classifying instances of Class 3.

Table 7. CatBoost classifier report

Classes	Precision	Recall	F1-Score	Support
1	0.76	0.73	0.74	62
2	0.67	0.63	0.65	63
3	0.83	0.92	0.87	52

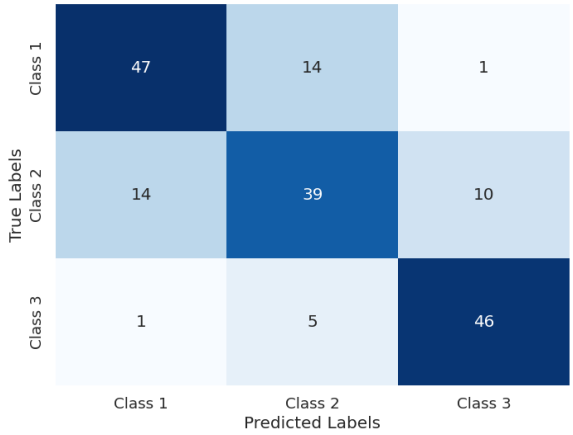

Fig.6. Confusion matrix of CatBoost

4.7. AdaBoost

The classification report for the AdaBoost Classifier depicted in Table 8 and confusion matrix in Fig. 7, outlines its performance across different classes. Class 3 displays the highest precision, recall, and F1-score values, indicating the model's effectiveness in accurately identifying instances from this class. Classes 1 and 2 exhibit relatively lower precision, recall, and F1-score values compared to Class 3, indicating some challenges in correctly classifying instances from these classes. However, the AdaBoost Classifier demonstrates acceptable performance across all classes, with notable strengths observed in correctly classifying instances from Class 3.

Table 8. AdaBoost classifier report

Classes	Precision	Recall	F1-Score	Support
1	0.62	0.58	0.60	62
2	0.48	0.49	0.49	63
3	0.75	0.79	0.77	52

4.8. Gradient Boost

The classification report is presented in Table 9 and confusion matrix in Fig. 8 for the Gradient Boosting Classifier. The results showcase its performance metrics across different classes. Class 3 demonstrates the highest precision, recall, and F1-score values, indicating the model's proficiency in accurately identifying instances from this class. While Classes 1 and 2 exhibit slightly lower precision, recall, and F1-score values compared to Class 3, the Gradient Boosting Classifier displays commendable performance overall, with notable strengths observed in correctly classifying instances from Class 3.

The comparative analysis of machine learning algorithms in this study reveals key insights into their performance variations. CatBoost emerged as the best-performing model with an accuracy of 75%, followed closely by Random Forest and XGBoost at 74%. The superior performance of CatBoost can be attributed to its handling of categorical features, efficient feature selection, and ability to reduce overfitting through ordered boosting techniques. Additionally, CatBoost's use of oblivious trees ensures robust classification performance, particularly in datasets with imbalanced class distributions like the one used in this study.

Tree-based ensemble models such as Random Forest and XGBoost also demonstrated strong performance due to their capability to handle nonlinear relationships in the data. Random Forest, with its multiple decision trees, helps in reducing variance and improving model stability, making it well-suited for complex classification tasks. Similarly, XGBoost, a gradient boosting algorithm, performed well due to its ability to iteratively correct classification errors, efficiently optimizing predictive accuracy. Gradient Boosting (72%) followed a similar trend, reinforcing the effectiveness of ensemble methods in structured medical datasets.

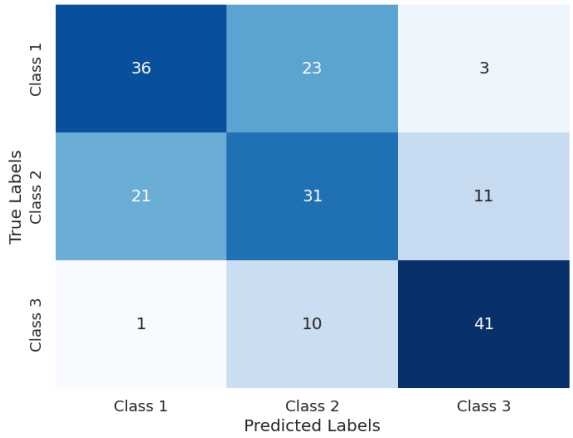

Fig.7. Confusion matrix of AdaBoost

Table 9. Gradient boost classifier report

Classes	Precision	Recall	F1-Score	Support
1	0.71	0.73	0.72	62
2	0.63	0.59	0.61	63
3	0.84	0.88	0.86	52

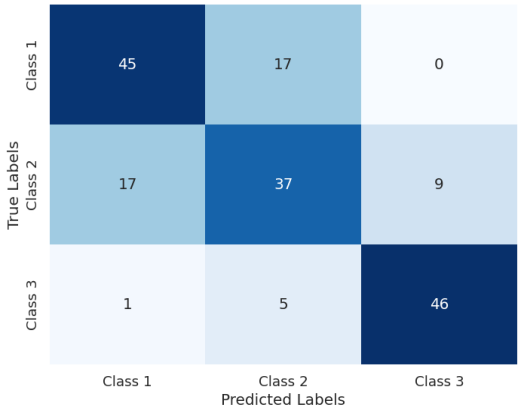

Fig.8. Confusion matrix of gradient boost

In contrast, SVM (59%) and AdaBoost (61%) exhibited lower accuracies, suggesting limitations in their ability to capture intricate feature relationships. SVM struggled due to the dataset's potential nonlinear feature dependencies, as it is more effective in high-dimensional spaces where feature separation is clearer. Moreover, AdaBoost, which relies on sequentially weighting misclassified instances, may have been more sensitive to noise in the dataset, affecting its performance.The performance of K-Nearest Neighbors (67%) highlights its reliance on local distance-based classification. While KNN can work well with small datasets, its effectiveness declines when dealing with complex, high-dimensional data, where feature relationships are more intricate.

Decision Trees (62%) performed lower than ensemble-based models because individual trees are prone to overfitting, limiting their generalization capability. However, their interpretability remains an advantage, making them useful for understanding key decision factors in pancreatic cancer prediction. Overall, the results indicate that ensemble learning models, particularly CatBoost, Random Forest, and XGBoost, offer the most promising approach for pancreatic cancer prediction due to their ability to handle complex data patterns, reduce overfitting, and optimize decision boundaries. Future improvements could focus on refining hyperparameters, integrating feature selection methods, and exploring deep learning techniques to further enhance model accuracy and generalizability.

The proposed CatBoost-based pancreatic cancer prediction model can be seamlessly integrated into clinical workflows by serving as a Clinical Decision Support System (CDSS) within Electronic Health Records (EHR), providing automated risk assessments based on biomarker profiles. It can assist physicians in diagnostic decision-making, differentiating pancreatic cancer from benign conditions, and prioritizing high-risk patients for further testing. The model's explainability through SHAP analysis enhances clinical trust, making it a valuable second-opinion tool. Additionally, it can be deployed in screening programs, integrated into pathology labs for automated report generation, and utilized in telemedicine for remote assessments.

5. SHAP-Based Explainability Analysis of the CatBoost Classifier

Based on the results presented in the previous section, CatBoost classifier is emerged as the best performing model and this section discusses the interpretability of the predictions obtained by CatBoost using SHAP values.

Explainability in machine learning is crucial for understanding model decisions, particularly in high-stakes applications like medical diagnosis. SHAP (Shapley Additive Explanations) is a widely used technique for interpreting model predictions by quantifying the contribution of each feature. SHAP assigns a Shapley value to each feature, indicating its impact on the model's output. The SHAP summary plot visually represents the distribution of these values, illustrating how feature importance varies across instances. This plot not only ranks features based on their overall influence but also shows how different feature values (high or low) affect predictions. By leveraging SHAP, this study enhances the interpretability of the CatBoost classifier, offering insights into the most critical predictors of pancreatic cancer.

The x-axis of the SHAP summary plots represents the SHAP values, indicating the impact of each feature on model predictions. A positive SHAP value implies that the feature contributes toward classifying a sample into a given category, whereas a negative SHAP value suggests that the feature pushes the prediction away from that class. Features with SHAP values around zero have minimal influence on the prediction for that instance. The spread of SHAP values highlights how the impact of features varies across different samples, providing deeper insights into the model's decision-making process. In SHAP summary plots, the colors indicate feature values, with red representing high values and blue representing low values. The x-axis shows SHAP values, which determine whether a feature increases or decreases the probability of a given classification. If a red-colored (high-value) point has a positive SHAP value, it means higher values of that feature contribute to classifying the sample into that class. Conversely, if a blue-colored (low-value) point has a negative SHAP value, it suggests that lower feature values reduce the likelihood of that class. This color representation helps interpret the directional impact of features on the model's predictions.

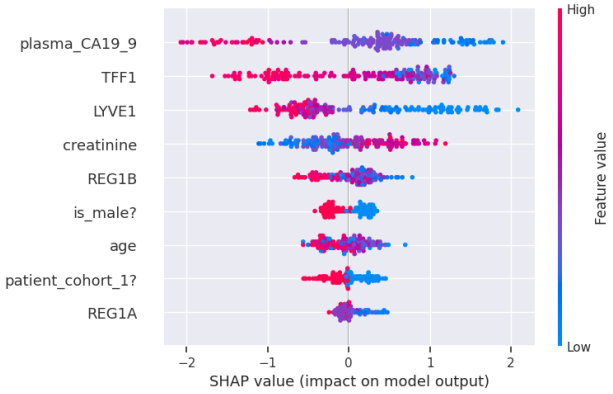

Fig.9. SHAP summary plot for No pancreatic disease class

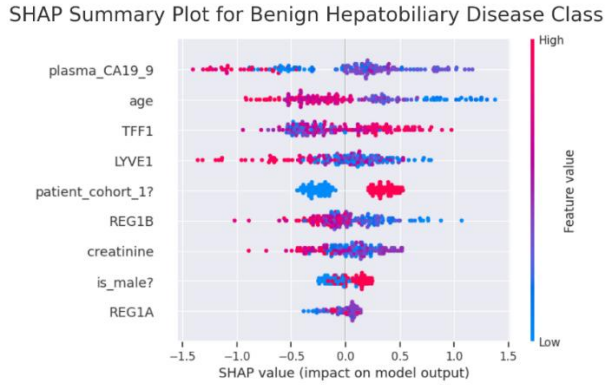

Fig.10. SHAP summary plot for benign hepatobiliary disease class

SHAP analysis as depicted in Fig. 9 highlights Plasma CA19-9, TFF1, and LYVE1 as the most influential features in predicting non-cancerous cases, with lower values increasing the likelihood of a "No Pancreatic Disease" classification. Creatinine exhibited a positive association, while REG1B and REG1A negatively impacted non-cancerous predictions, linking higher levels to pancreatic disease. Age and gender had moderate influences, suggesting demographic relevance. These findings reinforce the established role of Plasma CA19-9 and LYVE1 as key cancer biomarkers, confirming the model's ability to effectively differentiate non-cancerous cases based on biochemical markers.

SHAP analysis as shown in Fig. 10 reveals Plasma CA19-9, Age, and LYVE1 as key predictors in distinguishing benign hepatobiliary disease from pancreatic cancer and non-cancerous conditions. Lower Plasma CA19-9 values were strongly associated with benign diagnoses, while older age and higher LYVE1 and TFF1 values increased the likelihood of benign classification. REG1B and REG1A negatively correlated with benign cases, reinforcing their role in identifying pancreatic cancer. Additionally, lower creatinine levels were linked to benign conditions, suggesting a metabolic association. These findings highlight distinct biochemical profiles, emphasizing the diagnostic significance of Plasma CA19-9, TFF1, and LYVE1 in benign disease classification.

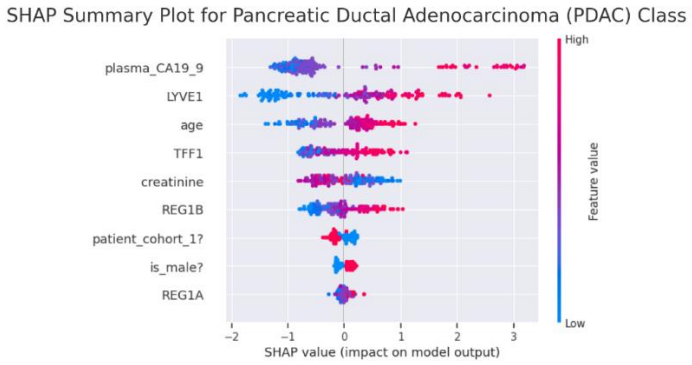

Fig.11. SHAP summary plot for pancreatic ductal adenocarcinoma class

SHAP analysis in Fig. 11 identifies Plasma CA19-9, LYVE1, and TFF1 as the most influential predictors of pancreatic ductal adenocarcinoma (PDAC), aligning with established clinical biomarkers. Higher Plasma CA19-9 values strongly increased PDAC likelihood, while LYVE1 and TFF1 reinforced differentiation from benign conditions. Older age contributed positively, suggesting an age-related risk factor, whereas lower creatinine levels were linked to PDAC, indicating metabolic alterations. REG1B further supported PDAC classification, confirming its relevance. These findings underscore the critical role of Plasma CA19-9, LYVE1, and TFF1 in pancreatic cancer diagnosis, aiding early detection and clinical decision-making.

SHAP analysis highlights Plasma CA19-9, LYVE1, and TFF1 as the most influential features across all diagnostic classifications, reinforcing their significance in pancreatic cancer detection. Plasma CA19-9 exhibited the highest SHAP values, confirming its strong diagnostic relevance, while LYVE1 and TFF1 effectively differentiated benign and malignant conditions. REG1B and REG1A played meaningful roles, particularly in pancreatic cancer classification, and creatinine showed an inverse association with disease presence. Age and gender had moderate effects, reflecting demographic trends, while cohort-based variations were also observed. These findings validate the model's robustness in leveraging clinically relevant biomarkers for accurate classification, enhancing interpretability in pancreatic cancer diagnosis.

6. Conclusions and Future Work

This study emphasizes the transformative role of machine learning in advancing early pancreatic cancer detection, a critical challenge in oncology due to the disease's high mortality and limited treatment options. By systematically evaluating a diverse set of ML algorithms, we demonstrated that machine learning can significantly enhance diagnostic accuracy, potentially aiding clinicians in identifying pancreatic cancer at an early stage. Among the tested models, CatBoost emerged as the most effective, showcasing superior accuracy and robustness in classification tasks. The results highlight the potential of ML-based diagnostic frameworks in complementing traditional medical approaches, offering data-driven insights that could improve clinical decision-making and patient outcomes. Additionally, the study contributes to the growing body of knowledge on AI-driven healthcare solutions, reinforcing the viability of ML techniques in oncology diagnostics. The findings not only validate machine learning's role in medical applications but also open new pathways for integrating AI-based tools into routine clinical workflows.

While this study effectively demonstrates the potential of machine learning models for pancreatic cancer detection, it is currently limited to a single publicly available dataset. Although the dataset is well-structured and diverse in terms of sample representation, the absence of external validation on independent datasets remains a limitation. Validation on additional datasets, preferably from different clinical sources, would further establish the generalizability and robustness of the proposed models.

Future research should focus on evaluating the models on external datasets to confirm their applicability across varied patient demographics and medical conditions. Additionally, real-world validation in collaboration with healthcare institutions would provide deeper insights into the clinical feasibility of integrating ML-based diagnostic tools. Despite this limitation, the study provides a strong foundation for further research in AI-driven oncology diagnostics and underscores the potential of machine learning in improving early pancreatic cancer detection.

References

[1] M. A. Atkinson, M. Campbell-Thompson, I. Kusmartseva, and K. H. Kaestner, "Organisation of the human pancreas in health and in diabetes," Diabetologia, vol. 63, no. 10, pp. 1966-1973, sept 2020, doi: 10.1007/s00125-020-05203-7.

[2] Leung, P.S, "The Renin-Angiotensin System:Current Research Progress in The Pancreas", *Springer*,vol 690, pp. 53, 2010.

[3] Allison Rosenzweig,"Black Americans at Increased Risk for Pancreatic Cancer", Pancreatic Cancer Action Network. https://pancan.org/news/black-americans-at-increased-risk-for-pancreatic-cancer/

[4] A. Moore and T. Donahue, "Pancreatic Cancer," JAMA, vol. 322, no. 14, pp. 1426-1426, oct 2019, doi: 10.1001/jama.2019.14699.

[5] Wilson Bakasa and Serestina Viriri, "Pancreatic Cancer Survival Prediction: A Survey of the State-ofthe- Art", Computational and Mathematical Methods in Medicine, pp. 2021:1188414, sept, 2021, doi: 10.1155/2021/11884.

[6] Wang, S., Zheng, Y., & Zhang, S., "Clinical application of artificial intelligence in pancreatic cancer" Cancer Letters, vol. 471, pp. 13-23, 2020.

[7] Davide Placido1, Bo Yuan2 Jessica X. Hjaltelin, Amalie D. Haue1, Piotr J Chmura1, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, Alexander Chowdhury, Alexandra Franz, Lauren Brais, Elizabeth Andrews, Debora S. Marks, Aviv Regev, Peter Kraft, Brian M. Wolpin, Michael Rosenthal, Søren Brunak1, Chris Sander, "Pancreatic cancer risk predicted from disease trajectories using deep learning",Nature Medicine,vol. 29, pp. 1113–1122, may 2023.[Online]. Available: https://www.nature.com/articles/s41591-023-02332-5

[8] Davide Placido, Bo Yuan, Jessica X. Hjaltelin, Chunlei Zheng, Amalie D. HauePiotr J. Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, Alexander Chowdhury, Alexandra Franz, Lauren Brais, Elizabeth Andrews, Debora S. Marks, Aviv Regev, Siamack Ayandeh, Mary T. Brophy, Nhan V, Peter Kraft , Brian M. Wolpin, Michael H. Rosenthal, Nathanael R. Fillmore, Soren Brunak & Chris, "A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories.", Nature Medicine, vol. 29, no. 5, pp. 1113-1122, May 2023, doi: 10.1038/s41591-023-02332-5.

[9] Bahrudeen shahul Hameed and Uma Maheswari Krishnan, "Artificial Intelligence-Driven Diagnosis of Pancreatic Cancer", Cancers, vol.14, no. 21, pp. 5382, oct 2022, doi: 10.3390/cancers14215382.

[10] H.S.Saraswathi, Mohamed Rafi, " Promising Urinary Biomarkers to predict Pancreatic Ductal Adenocarcinoma using Machine Learning Techniques", European Chemical Bulletin, vol.12, no. 8, pp. 1858-1869,2023.[Online].Available: https://www.eurchembull.com/uploads/paper/341a314135b011716cd2cb97b5170140

[11] Haidy Nasief, Cheng Zheng, Diane Schott, William Hall, Susan Tsai,Beth Erikson and X.Allen Li, "A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer", npj Precision Oncology, no. 25, oct 2019. [Online]. Available: https://www.nature.com/articles/s41698-019-0096-z

[12] Seiya Yokoyama1, Taiji Hamada1, Michiyo Higashi1, Kei Matsuo1, Kosei Maemura2,3, Hiroshi Kurahara3, Michiko Horinouchi1, Tsubasa Hiraki1, "Predicted Prognosis of Patients with Pancreatic Cancer by Machine Learning", Clinical Cancer Research, vol. 26, pp. 2411–2421, 2020, doi: 10.1158/1078-0432.CCR-19-1247.

[13] Alexandra Corina Faur, Daniela Cornelia Lazar, and Laura Andreea Ghenciu, "Artificial intelligence as a noninvasive tool for pancreatic cancer prediction and diagnosis", World Journal of Gastroenterology, vol.29, no. 12, pp. 1811-1823, march 2023, doi: 10.3748/wjg.v29.i12.1811.

[14] M.H. Osman, "Predicting survival of pancreatic cancer using supervised machine learning", GASTROINTESTINAL TUMOURS, NON-COLORECTAL, vol. 29, pp. VIII255.[8], oct 2018, doi:10.1093/annonc/mdy282, 1905.

[15] Alvaro martin. (2023).," Detect pancreatic cancer early using machine learning", neural designer, Aug 2023. [Online]. Available: https://www.neuraldesigner.com/learning/examples/pancreatic-cancer/

Authors' Profiles

Radha Singh Jadaun is a Graduate Engineer Trainee at Navitasys India Private Limited, Bawal. She graduated with a B.Tech in Electrical Engineering with a specialization in Computer Science from Dayal Bagh Educational Institute (2020-2024). During college, Radha participated in 'SMART INDIA HACKATHON 2022'. Her technical skills include Java, C/C++, SQL, HTML/CSS, and familiarity with Python, MATLAB, Git, and various data analysis tools. She also served as a Training and Placement Coordinator in college.

Nij Mehar Grover is a B.Tech graduate from Dayalbagh Educational Institute, Agra, India. With a strong foundation in mobile app development and a growing expertise in web technologies, he has developed real-world applications that have reached thousands of users. During a 1.5-month internship at Bhanzu, he honed skills in Python, DBMS, AWS, and API development, which led to a full-time role at the company. Currently, is working on enhancing the Bhanzu's products, transitioning hackathon projects into market-ready products, and exploring generative AI.

K. Srinivas received a B.E. in Computer Science & Technology, an M.Tech. in Engineering Systems and a Ph.D. in Electrical Engineering. He is currently working as an Associate Professor in the Electrical Engineering Department, Faculty of Engineering, Dayalbagh Educational Institute (Deemed to be University) Agra, India. His research interests include AI Machine Learning, Deep Learning, Soft Computing, Optimization using Metaheuristic Search Techniques, Search-Based Software Engineering, Mobile Telecommunication Networks, and Systems Engineering. He is an active researcher, guiding research, publishing papers in journals of national and international repute, and being involved in R&D projects. He is a member of IEEE and a life member of the Systems Society of India (SSI).

A. Charan Kumari received her Ph.D from Dayalbagh Educational Institute, in collaboration with Indian Institute of Technology, Delhi, India, under their MOU. She is currently working as an Assistant Professor in the Electrical Engineering Department, Dayalbagh Educational Institute (Deemed to be University) Agra, India. Her current research interests include Machine Learning, Search Based Software Engineering, Evolutionary computation and soft computing techniques. She has published papers in journals and conferences of national and international repute. She has delivered an invited talk at 43rd CREST open workshop on Hyper-heuristics at University College London, London. She is a member of IEEE, Computer Society of India (CSI) and Systems Society of India (SSI).

How to cite this paper: Radha Singh Jadaun, Nij Mehar Grover, K. Srinivas, A. Charan Kumari, "Pancreatic Cancer Prediction Using Machine Learning: An Investigation of Different Algorithms", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.6, pp.127-142, 2025. DOI:10.5815/ijitcs.2025.06.07