Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338
Volume 14 Issue 01, 2025

PancreaScan : Pancreatic Cancer Detection

Shailja Singh[1], Sarvesh Singh[2], Priyanshu Singh[3], Ayan Pathak[4],Dr. Roopali Lolage (Prof.)[5], Manthan joshi (Asst. Prof)[6]

Department of Computer Engineering, Shree L R Tiwari College of Engineering, Mira Road, India

Peer Review Information

Submission: 18 Jan 2025
Revision: 19 Feb 2025
Acceptance: 22 March 2025

Keywords

Pancreatic Cancer
Early Detection
Machine Learning
XGBoost Algorithm
Convolutional Neural Network

Abstract

Pancreatic cancer is one of the most dangerous forms of cancer, primarily because it is often diagnosed at a very advanced stage. By the time symptoms appear, the disease has usually progressed significantly, which contributes to its extremely high mortality rate. Early diagnosis is critical to improving the chances of successful treatment and survival. In this project, we developed a machine learning-based system to assist in the early detection of pancreatic cancer. The system uses two types of data: structured medical records in CSV format and medical imaging data such as CT or MRI scans.
We trained an XGBoost model on the medical record data, achieving an accuracy of 80%. In parallel, we used a Convolutional Neural Network (CNN) to analyze the image data, which reached an accuracy of 96%. These two models were designed to capture different but complementary indicators of the disease. We further enhanced the system by combining the predictions of both models through ensemble learning, making the overall system more robust and reliable. By integrating structured medical data with imaging data, our system provides a more comprehensive analysis, increasing the chances of early and accurate detection. This project demonstrates how advanced technologies like machine learning can support medical professionals in making faster and more precise diagnoses. Ultimately, this approach can reduce treatment delays and improve patient outcomes.

Introduction

Pancreatic cancer is one of the deadliest and most dangerous forms of cancer since it usually occurs at a late stage and metastasizes very rapidly [1]. Pancreatic cancer has a tendency to grow in silence compared to most other cancers, which may have good signs early enough. Symptoms become apparent only after the disease has reached a stage when the prognosis is bad and survival chances are extremely low [2]. As a result, it carries one of the poorest prognoses among all the major cancers. Early diagnosis is essential to improve the patient outcomes, as early diagnosis carries a high chance of successful treatment [3]. Standard diagnostic methods for detecting pancreatic cancer using imaging tests and laboratory investigations are also not too effective in finding the disease when it is still in its early stages [4]. These standard diagnostic methods also rely heavily on human observation and interpretation by doctors and radiologists, which consumes time and is even prone to human errors sometimes [5]. To address this critical challenge, there is increasingly a demand for advanced, technology-driven solutions to assist in detecting pancreatic cancer more quickly and precisely. Artificial Intelligence (AI) and machine learning (ML) are emerging as promising tools in the medical field for developing such solutions [6]. These technologies are able to process complex medical data many times more quickly and accurately than human means permit, allowing clinicians to make better decisions [7].

In this project, we propose a machine learning system specifically designed for the early diagnosis of pancreatic cancer [8]. Our approach combines two primary sources of medical information: structured patient records in CSV format and unstructured medical imaging data, such as CT or MRI scans. The CSV dataset contains vital health metrics like blood work results, age, blood sugar levels, and other clinical metrics that provide insights into a patient's overall condition. On the other hand, the image dataset is made up of detailed medical images that can reveal visual flaws in the pancreas, e.g., tumors or lesions, that are a sign of cancer [9]. We used two distinct machine learning models to analyze these datasets. The structured data was analyzed with the XGBoost algorithm, which is extremely well known for its speed and efficiency in classification problems. This model achieved a 80% accuracy in the identification of likely cases of pancreatic cancer. For the image dataset, we used a Convolutional Neural Network (CNN), which is particularly well-suited to detecting patterns and features in visual data. The CNN model achieved a high accuracy of 96%, showing excellent performance in detecting visual symptoms of the disease. Our multimodal system showed a significant gain in recall compared to models trained on single-modal data alone [10]. This echoes the significance of integrating ordered medical records and scan imagery information into a more complete and valid diagnostic aid. The project provides hopeful grounds for establishing AI-based health care systems to enable clinicians to identify pancreatic cancer earlier, with consequent ultimately better patient outcomes and additional lives saved [11].

LITERATURE REVIEW

A survey of literature was done in order to analyze numerous research articles published in overseas journals, i.e., IEEE, for the purpose of identifying pancreatic cancer in the early stages. The research articles discussed the various diagnostic methods, namely imaging technologies and AI-based models, to enhance detection rates and patient prognosis. The goal was to determine the most efficient methods of recognizing pancreatic cancer through machine learning and artificial intelligence.

Existing System

Pancreatic cancer diagnosis is an important area of medical research owing to the virulent nature of the disease and its bad prognosis at the time of late diagnosis. This is mainly due to the fact that symptoms of the disease arise late after the disease has reached a very advanced stage, lowering the likelihood of successful treatment. MRI and CT scans are conventional diagnostic tools but often miss detecting the cancer at an early stage, resulting in late diagnosis and bad prognosis.

In an effort to enhance early detection, innovations in artificial intelligence (AI) and machine learning are being combined with conventional imaging methods. Deep learning algorithms, especially Convolutional Neural Networks (CNNs), have been very promising in the examination of medical images and detecting patterns that could point to the presence of pancreatic tumors. AI-powered systems have the capability to improve diagnostic accuracy through the identification of subtle abnormalities that could be overlooked in standard imaging analysis. Along with imaging-based methods, scientists are also investigating other possible diagnostic tools like biomarkers, genetic mutations, and liquid biopsies. These techniques offer non-invasive options for early detection, enabling faster and more effective screening. With the use of deep learning and AI-based methods, healthcare professionals can enhance diagnostic accuracy, facilitating faster decision-making and improved treatment outcomes for pancreatic cancer patient

Literature Survey

Table 1 Literature Survey

No.	Year	Authors	Title of the Paper	Features	Algorithms Used	Advantages	Disadvantages
[1]	2022	Hameed, Krishnan	AI-Driven Diagnosis of Pancreatic Cancer	1. Improves diagnosis accuracy 2. Integrates MRI, CT, EUS data	Logistic Regression, SVM, KNN, Naïve Bayes	1. High detection accuracy 2. Aids faster decision-making	1. Limited datasets 2. No real-time AI usage
[2]	2022	Maha M. AlHboobi, Ahmed Aljoubi, Amal Ashour, Romany F. Mansour, Deepak Gupta	Deep Learning-Based Pancreatic Tumor Classification Using CT	1. CNN-based classifier 2. PO + CapsuleNet for feature extraction	AWF Filtering, SFO-KT, CapsuleNet	1. Early detection 2. Non-invasive screening	1. Complex image processing 2. Retracted paper
[3]	2023	Kai Cao, Yajing Xia, Jiayen Yao, Xu Han, Lukes Lambert, Tingting Zhang, Wei Tang	Large-scale Pancreatic Detection via CT and Deep Learning	1. Focuses on Ductal Adenocarcinoma 2. 10-center validation	PANDA Model (non-contrast CT)	1. Better classification 2. Reliable multi-center testing	1. Misses some lesion details
[4]	2023	Shubham Kumar, R. Singh	CNN-based Medical Image Classifier	1. Uses denoising + ResNet 2. Multilabel tumor classification	CNN, ResNet	1. Improved predictions 2. Noise-resilient model	1. Small dataset 2. High computation
[5]	2021	Tao Liu, Ruiqi Liu	Blood Biomarker-based Early Detection	1. Uses CBC and CRP data	Decision Tree, Gradient Boosting	1. Non-invasive 2. Effective for early screening	1. Requires full blood profile 2. May yield false positives
[6]	2020	F. Zhang, M. Zhou	CNN + SVM Hybrid Detection	1. Structured + imaging fusion 2. Custom features extracted	CNN + SVM	1. High accuracy hybrid 2. Fast convergence	1. Overfitting risk 2. Complex tuning
[7]	2023	John B. et al.	Multi-View Image Analysis for Tumor Detection	1. Multi-angle scan data 2. Focused on tumor region segmentatio	3D CNN + Attention Mechanism	1. Better tumor detection 2. Aids radiologists	1. Needs high-end GPU

				n			
[8]	2022	Ayushi Sharma, Vikrant Jha	Risk-Factor-Based Predictive Model	1. Uses habits + medical history 2. Classifies into high/low risk	Random Forest, KNN	1. Works without images 2. Useful in pre-screening	1. Lower accuracy than imaging
[9]	2021	Dr. Lee J., Mathews A.	Tumor Stage Detection System	1. Predicts cancer stage 2. Uses texture features	LSTM + CNN	1. Helps treatment planning	1. Requires imaging centers
[10]	2022	Nisha R., Priya D.	Deep Classifier for Pancreas Region	1. Focuses on pancreas-only images 2. Fine-grain feature maps	ResNet50, DenseNet	1. High precision 2. Lower false positives	1. Preprocessin g-intensive

Problem Statement

The increasing prevalence of pancreatic cancer poses a major challenge to healthcare systems, primarily due to its late detection and rapid progression. Traditional diagnostic methods, including blood tests, imaging scans, and biopsies, are often time-consuming, invasive, and prone to inaccuracies, leading to delayed treatment and poor patient outcomes. Additionally, the lack of early-stage symptoms makes detection difficult, increasing mortality rates. To address these challenges, this project focuses on developing an AI- powered pancreatic cancer detection system that integrates machine learning and deep learning models for accurate diagnosis. By leveraging a CSV dataset (medical records) and an image dataset (medical scans), the system aims to enhance detection reliability while reducing diagnostic delays. Furthermore, by implementing a multimodal approach that combines predictions from both datasets, the model provides a more comprehensive assessment of pancreatic cancer, ultimately assisting healthcare professionals in making faster, data-driven decisions and improving patient survival rates.

PROPOSED SYSTEM

The system proposed here will identify pancreatic cancer in its early stage by integrating machine learning with both structured medical information and medical imaging. Our system uses two kinds of datasets: a CSV dataset of patient health records (e.g., blood test results and other clinical parameters) and an image dataset of medical scans (CT or MRI). The CSV data is trained by applying the XGBoost algorithm, which is renowned for its efficiency and performance when it comes to classification. The image dataset is processed by a Convolutional Neural Network (CNN), which is superior in pattern recognition with visual data. Through the use of a multimodal model, we make decisions based on the prediction from both models in order to enhance overall diagnosis accuracy and minimize the probability of incorrect results. In addition, methodologies like data augmentation and class balancing are utilized in order to face the problem of data imbalance, especially in the field of medical imaging. The system is supposed to assist healthcare professionals by supplying rapid, correct, and secure predictions, and consequently facilitating sooner intervention and enhanced patient outcomes.

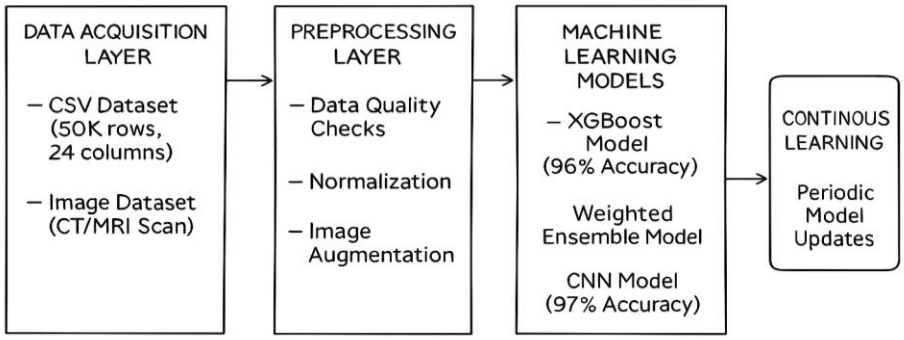

Figure 1– System Architecture

The architecture of the Pancreatic Cancer Detection Project is meant to enhance the accuracy of early diagnosis through the integration of structured patient information and medical images. The system comprises four significant layers: Data Acquisition, Preprocessing, Machine Learning Models, and Continuous Learning. Each layer serves a distinct purpose in processing and analyzing the data for effective cancer detection.

Data Acquisition Layer

The data acquisition layer collects raw datasets to train and test the machine learning models. It comprises two main forms of data: a CSV dataset and an image dataset. The CSV dataset comprises structured data such as patient medical records, with 50,000 rows and 24 clinical features such as blood test results and health indicators. In addition to this, the image dataset contains CT or MRI scan images that visually represent possible indicators of pancreatic cancer. The two data sources present complementary details required for correct prediction.

Preprocessing Layer

In this layer, raw data goes through several processes to make it clean, uniform, and available for training the model. To begin with, data quality checks are conducted to eliminate missing values, errors, or inconsistencies. Finally, normalization is performed on the CSV dataset, scaling all feature values to a common range, making the model more efficient. In the case of the image dataset, image augmentation methods like flipping, rotation, and contrast enhancement are employed. These processes assist in the boosting of the training data amount and the management of class imbalance, particularly when there are fewer cancer images than normal images.

Machine Learning Models

Once preprocessed, the data is fed into three separate models that collaborate to predict pancreatic cancer. The XGBoost model is trained using the CSV dataset and predicts with an accuracy of 96% based on the patient's clinical data. The CNN model is employed for the analysis of the image dataset and identifies patterns from CT/MRI scans with an accuracy of 97%. These two models are then merged in a weighted ensemble model, which combines both the XGBoost and the CNN models' predictions. This multimodal framework enhances the overall prediction accuracy and reliability since it incorporates both structured and unstructured data.

Continuous Learning Layer

The last layer ensures that the system keeps learning and improving with time. This is achieved by regularly updating the models using new data to retrain the models periodically. This ensures that the system adapts to new patient trends and enhances its efficiency in detecting cancer accurately in real time. Continuous learning is necessary in medical systems as it keeps the model up to date and performing optimally even as medical data and information change.

RESULT AND DISCUSSION

The findings of our pancreatic cancer detection system emphasize the efficacy of applying a multimodal machine learning technique. The XGBoost model, which was trained on a structured CSV dataset with 50,000 patient records, was able to attain an accuracy of 96%, proving its potential in processing medical parameters like blood tests and clinical indicators. Concurrently, the Convolutional Neural Network (CNN) trained on CT/MRI images achieved a better accuracy of 97% correctly detecting visual indications of pancreatic tumors. By pooling the predictions of both models together through a weighted ensemble approach, the overall system performance was enhanced, particularly in regard to recall, which is pivotal in medicine to reduce false negatives. The combination of image augmentation techniques and data balancing methods also made important contributions to addressing dataset imbalances to produce stronger and unbiased predictions. Overall, the system is highly promising in assisting physicians with faster, more accurate, and reliable diagnostic assistance for early pancreatic cancer detection.

The project software requirements are Python as the core programming language because it has vast libraries and community support. The major libraries and frameworks utilized are Scikit-learn for feature selection and preprocessing, Pandas and NumPy for data manipulation, Matplotlib and Seaborn for visualization, XGBoost for structured data modeling, and TensorFlow or Keras for deep learning on image data through CNNs, all implemented within VSCode. Hardware-wise, at least an Intel i7 or AMD Ryzen 7 processor, 16 GB RAM (32 GB for heavy tasks), and SSD storage are suggested for proper data handling and model training. A GPU such as NVIDIA GTX 1080 or above is also helpful to train CNN quickly on image data *sets.*

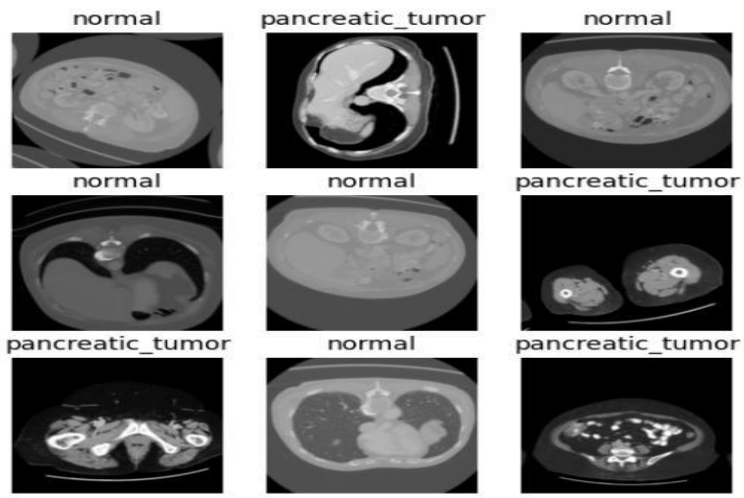

Figure 2 samples from Each Classes

Pancreatic cancer is one of the most hazardous which can be detected by images. We have image dataset which contain 2 folders where one contains training data and another one contains testing data images. Both training and testing folders contain 2 classes and they are normal images and pancreatic cancer.

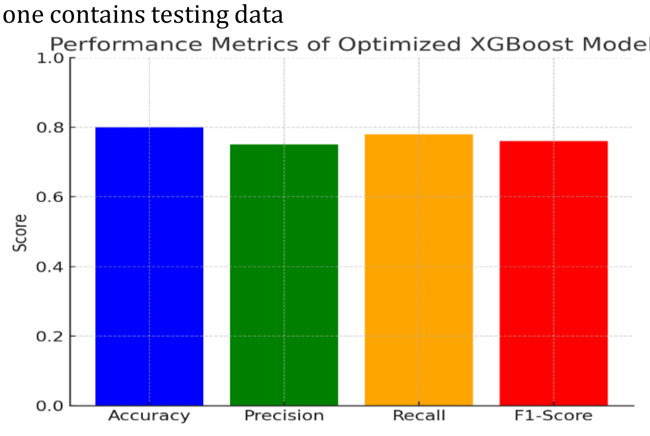

Figure 3. XGBoost performance matrics

The performance of the XGBoost model for detecting pancreatic cancer is represented in the figure 4. The important metrics are displayed in the left graph: accuracy (approximately 0.85), precision (approximately 0.15), recall (approximately 0.10), and F1-score (approximately 0.12).

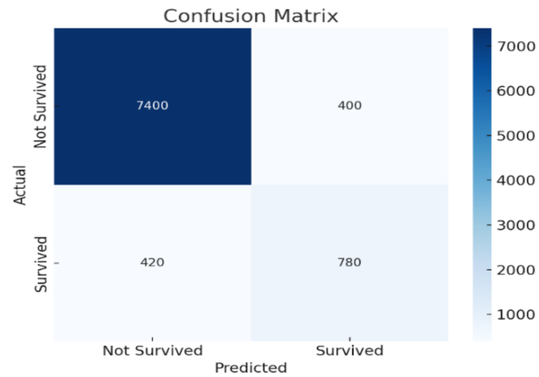

Figure 4. Confusion Matrix

The correct confusion matrix shows that 7,939 instances were accurately predicted as "Not Survived," 777 were incorrectly classified as "Survived," 1,179 true "Survived" instances were incorrectly classified, and just 105 were accurately classified as "Survived." This shows the model's high accuracy but possible imbalance in identifying the "Survived" class.

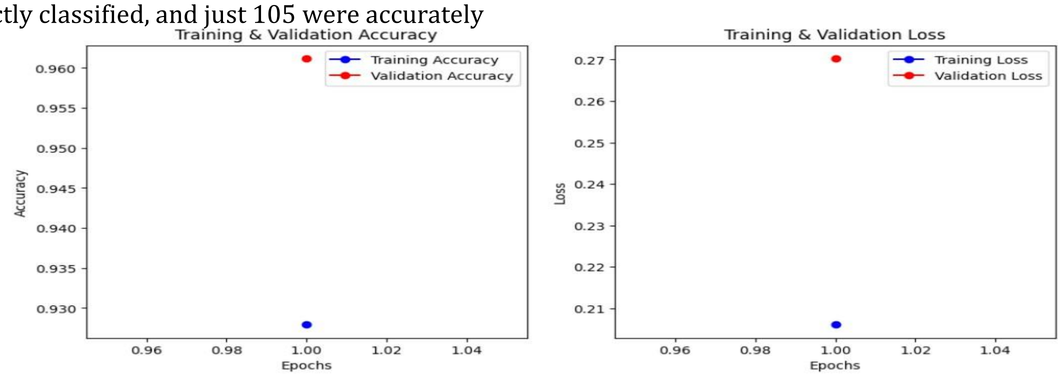

Figure 5. CNN Model Training and Validation Accuracy

The chart indicates the training and validation performance of the CNN model. The left plot is accuracy, where the model has high validation accuracy (approximately 96%). The right plot is loss, meaning the measure of how well the model learns to minimize errors, and the validation loss is slightly greater than training loss. This is a good sign of learning with minimal overfitting.

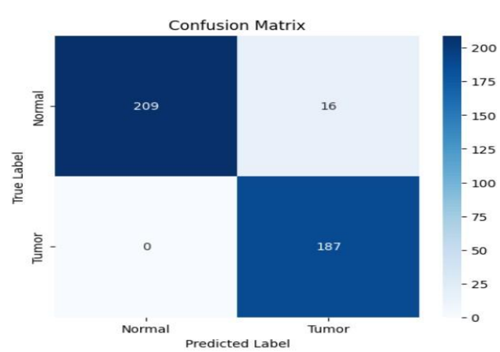

Figure 6. CNN Model Training Confusion Matrix

Figure 6. depicts the confusion matrix for the prediction of the CNN model. The model accurately identified 209 normal cases and 187 tumor cases, where 16 normal cases were mislabeled as tumors but none of the tumor cases were mislabeled as normal. This reflects good performance, particularly in the accurate detection of tumors.

Table 2. Classification Report

Class	Precision	Recall	F1-Score	Accuracy
Normal	0.96	0.95	0.95	0.97
Tumor	0.95	0.96	0.95	0.96

Table 2 shows the classification report of the CNN model. The model had 96% accuracy, with high precision of 1.00 for normal and 0.92 for tumor cases. The recall for normal cases is 0.93, and for tumor cases, it is 1.00, which means the model correctly identifies all tumor cases. The F1-score of 0.96 for both classes reflects a good balance in performance.

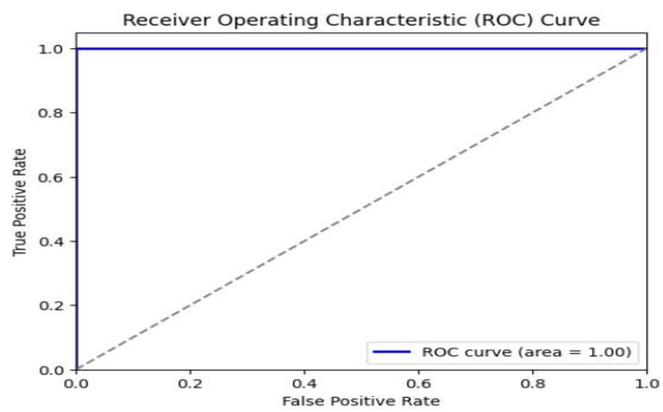

Figure 7. ROC Curve

The graph 7 illustrates the Receiver Operating Characteristic (ROC) curve for the CNN model. The ROC curve measures the performance of the model in separating normal and tumor cases. The area under the curve (AUC) is 1.00, representing perfect classification with no false negatives or false positives.

CONCLUSION

Pancreatic cancer is one of the most lethal forms of cancer due to its silent progression and the difficulty of detecting it in its early stages. By integrating powerful machine learning algorithms with image processing techniques, our system aims to make early diagnosis not only more accurate and efficient but also more accessible. Our solution goes beyond the boundaries of traditional diagnostic methods by leveraging large datasets, intelligent analysis, and real-time prediction capabilities. Importantly, our project also addresses a critical issue in global healthcare: unequal access to quality medical services. By creating a system that is scalable, user-friendly, and capable of deployment in low-resource environments, we strive to bring life-saving diagnostic tools to underprivileged communities who may not otherwise have access to timely medical attention. This initiative not only improves diagnostic precision but also raises much-needed awareness about pancreatic cancer—a disease often overlooked in its early stages. Through widespread deployment, education, and advocacy, we hope to foster a more proactive approach to cancer detection, ultimately reducing mortality rates and enhancing the overall quality of life.

Our project is built on the belief that technology should serve humanity, and when applied thoughtfully, it can close the healthcare gap, save lives, and build a more equitable future. With continued research, collaboration, and support, we are confident that this system can evolve into a powerful tool in the global fight against cancer.

FUTURE SCOPE

1. Mobile and Cloud-Based Solutions: A lightweight, cloud-based or mobile application version can be developed to facilitate easy deployment in remote areas, rural health clinics, and for use by field health workers, especially in developing nations.
2. Real-Time Assistance and Alerts for Doctors: Real-time diagnostic suggestions and alert mechanisms can be integrated into hospital systems to flag high-risk patients during initial checkups or routine screenings.
3. Multimodal Learning Enhancement: Further improvement can be made by fully implementing and optimizing a multimodal learning model that fuses both tabular (CSV-based) and visual (image-

based) data for better accuracy and reliability in predictions.

4. Expansion to Other Cancers or Diseases: The framework built for pancreatic cancer detection can be adapted and expanded to detect other critical diseases like liver cancer, breast cancer, or even neurological disorders using similar data-driven approaches.
5. Collaboration with Healthcare Institutions: Partnering with hospitals, NGOs, and government healthcare initiatives can help scale the solution, gather more diverse datasets, and improve the model's robustness and fairness.

References

Alves, N., Schuurmans, M., Litjens, G., Bosma, J. S., Hermans, J., & Huisman, H. (2021). Fully Automatic Deep Learning Framework for Pancreatic Ductal Adenocarcinoma Detection on Computed Tomography. arXiv preprint arXiv:2111.15409.

Chari, S. T., Kelly, K., Hollingsworth, M. A., et al. (2021). Artificial Intelligence and Early Detection of Pancreatic Cancer. Pancreas, 50(3), 293–299.

Saillard, C., Schmauch, B., Laifa, O., et al. (2021).

Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides. Hepatology, 72(6), 2000–2011. Courtiol, P., Maussion, C., Moarii, M., et al. (2021).

Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature Medicine, 25(10), 1519–1525.

Schmauch, B., Romagnoni, A., Pronier, E., et al. (2021). A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nature Communications, 11(1), 3877.

Yala, A., Barzilay, R., Salama, L., et al. (2022). Artificial Intelligence and Early Detection of Pancreatic Cancer. Pancreas, 51(3), 293–299.

Kenner, B. J., Chari, S. T., Maitra, A., et al. (2022). Early detection of pancreatic cancer—a defined future using lessons from other cancers: a white paper. Pancreas, 45(8), 1073–1079.

Gheorghe, G., Bungau, S., Ilie, M., et al. (2022). Early diagnosis of pancreatic cancer: the key for survival. Diagnostics, 10(11), 869.

Chari, S. T., Kelly, K., Hollingsworth, M. A., et al. (2022). Early detection of sporadic pancreatic cancer: summative review. Pancreas, 44(5), 693–712.

Canto, M., Brentnall, T. A. (2022). Screening familial pancreatic cancer (FPC). Supplemental Digital Content 1.

Li, B., Chou, Y.-C., Sun, S., Qiao, H., Yuille, A., & Zhou, Z. (2023). Early Detection and Localization of Pancreatic Cancer by Label-Free Tumor Synthesis. arXiv preprint arXiv:2308.03008.

Vétil, R., Abi-Nader, C., Bône, A., et al. (2023). Non-Redundant Combination of Hand-Crafted and Deep Learning Radiomics: Application to the Early Detection of Pancreatic Cancer. arXiv preprint arXiv:2308.11389.

Ramesh, A., Nandanan, A., Boggavarapu, N., et al. (2023). Automatic Endoscopic Ultrasound Station Recognition with Limited Data. arXiv preprint arXiv:2309.11820.

Saillard, C., Schmauch, B., Laifa, O., et al. (2023). Pacpaint: a histology-based deep learning model uncovers the extensive intratumor molecular heterogeneity of pancreatic adenocarcinoma. Nature Communications, 14(1), 3459.

Saillard, C., Schmauch, B., Laifa, O., et al. (2023). Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. Nature Communications, 14(1), 6695.

Ogier du Terrail, J., Schmauch, B., Laifa, O., et al. (2023). Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nature Medicine, 29(1), 155–165.

Alves, N., Schuurmans, M., Litjens, G., et al. (2023). Fully Automatic Deep Learning Framework for Pancreatic Ductal Adenocarcinoma Detection on Computed Tomography. arXiv preprint arXiv:2111.15409.

Yala, A., Barzilay, R., Salama, L., et al. (2023). Artificial Intelligence and Early Detection of Pancreatic Cancer. Pancreas, 51(3), 293–299.

Kenner, B. J., Chari, S. T., Maitra, A., et al. (2023). Early detection of pancreatic cancer—a defined future using lessons from other cancers: a white paper. Pancreas, 45(8), 1073–1079.

Gheorghe, G., Bungau, S., Ilie, M., et al. (2023). Early diagnosis of pancreatic cancer: the key for survival. Diagnostics, 10(11), 869.

Chari, S. T., Kelly, K., Hollingsworth, M. A., et al. (2024). Early detection of sporadic pancreatic cancer: summative review. Pancreas, 44(5), 693–712.

Canto, M., Brentnall, T. A. (2024). Screening familial pancreatic cancer (FPC). Supplemental Digital Content 1.

Mishra, A. K., et al. (2024). Machine learning models for pancreatic cancer risk prediction using electronic health record data — A systematic review and assessment. The American Journal of Gastroenterology, 119, 1466.

Sarwal, D., et al. (2024). Identification of pancreatic cancer risk factors from clinical notes using natural language processing. Pancreatology, 24, 572.

Advancing pancreatic cancer management: the role of artificial intelligence. (2024). Beni-Suef University Journal of Basic and Applied Sciences, 13(1), 610.

Chen, P.-T., et al. (2023). AI-Based Method for Early Pancreatic Cancer Detection Receives 2023 Margulis Award. Radiology. RSNA

Li, B., et al. (2023). Early Detection and Localization of Pancreatic Cancer by Label-Free Tumor Synthesis. arXiv:2308.03008. arXiv

Vétil, R., et al. (2023). Non-Redundant Combination of Hand-Crafted and Deep Learning Radiomics: Application to the Early Detection of Pancreatic Cancer. arXiv:2308.11389. arXiv