Intersect, Vol 18, No 3 (2025)

Artificial Intelligence in Healthcare: Early Pancreatic Cancer Detection Using Urinary Biomarkers

Pavlos Martinis
International School of Lausanne

Abstract

Pancreatic cancer is one of the deadliest malignancies due to its late-stage diagnosis and lack of effective early detection tools. Existing detection and screening methods currently fail to identify the tumor at its early, more treatable stages, contributing to persistently low survival rates and necessitating alternative approaches. However, in recent times, machine learning (ML), which is a branch of artificial intelligence (AI), has shown immense promise in the field, potentially enhancing early cancer detection by identifying minute and subtle patterns in clinical data. This study explores the application of machine learning and deep learning in the prediction of pancreatic cancer, using notably as input a set of patient urinary and blood biomarkers identified in previous studies as potentially promising for early detection of pancreatic cancer. The goal, after all, of this study is to predict the presence of the disease *before* it is diagnosed. Four classification models (Neural Network, Decision Tree, Random Forest, and K-Nearest Neighbors) were implemented to analyze the data features, classifying individuals as healthy, having benign hepatobiliary disease, or having pancreatic cancer. To further improve prediction reliability, a Multiplicative Weight Update (MWU) method was applied to dynamically adjust the influence of each model based on their testing performance, finally forming an overall more robust and accurate program. The integration of four distinct classification models, in tandem with the MWU method, distinguishes this research from previous studies and enhances its predictive performance. Given the varying concentrations of biomarkers associated with different pancreatic conditions, the use of multiple diverse models to capture both linear and complex non-linear patterns in the biomarker data was particularly important, something prior studies relying on individual models rarely achieved. As a result, the final prediction accuracy was significantly improved. The results demonstrate high accuracies for most models, with the Decision Tree achieving the highest predictive accuracy of **98.7%**. These results highlight the potential of AI-driven diagnostic tools in improving early pancreatic cancer detection.

Introduction

Pancreatic cancer is one of the most lethal tumors, partially due to its difficulty of early detection and the absence of existing detection tools in the industry. By the time symptoms of this malignant cancer appear, the disease is far too advanced, drastically reducing therapeutic options and resulting in dismal survival rates. Historical data shows that the 10-year survival rate for pancreatic cancer has remained at a meager 1% both in 1971 and 2011, showing no improvement despite advancements in cancer research and treatment methods. On the other hand, other cancers such as testicular, skin, breast, and prostate, have seen dramatic increases in their long-term survival rates over the same period (Ali, 2016). Such static survival outcomes urgently call for an innovative solution, and with the constant advancements of AI-driven methodologies, machine learning can most certainly be effectively leveraged to help advance this cause, and detect malignancy at more treatable stages, hence potentially altering and improving its historically grim prognosis and survival rates.

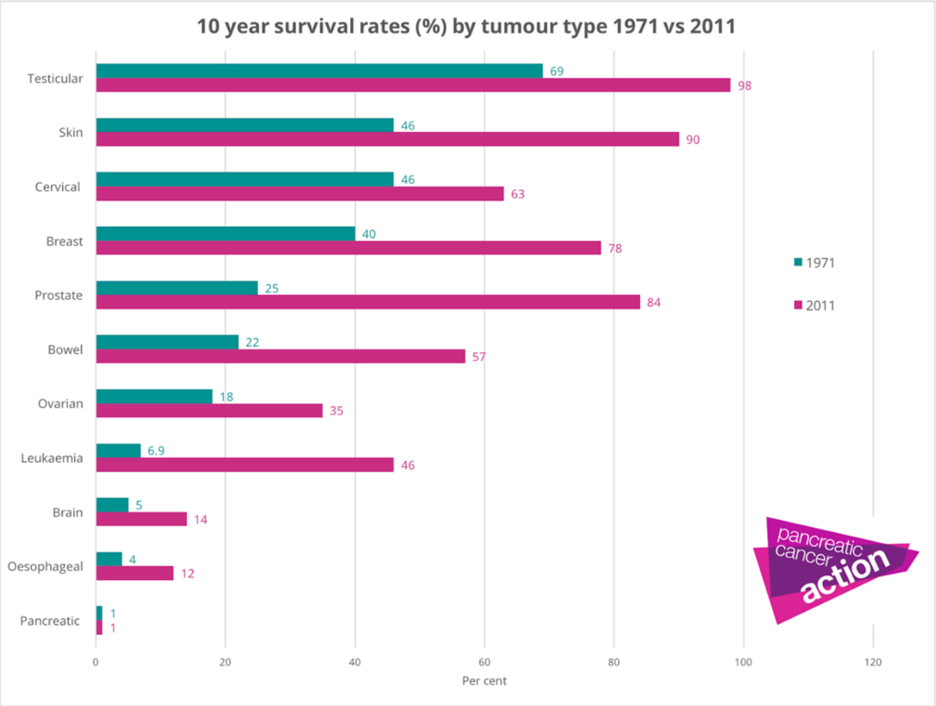


FIGURE 1: 10-year survival rates of different cancers in 1971 vs 2011 (Ali, 2016).

AI in healthcare

Artificial intelligence and machine learning, a branch of artificial intelligence, have already and are yet to further revolutionize the medical and pharmaceutical sectors, offering faster, more efficient, and more accurate diagnosis and treatment for potentially life-threatening diseases. By employing a variety of statistical, probabilistic and optimization techniques, AI technologies offer the possibility to analyze and process vast amounts of clinical data, from medical records to medical imaging, as well as identify patterns that would not be easily identifiable by humans. In this way, they enable early and precise disease diagnosis through analysis of subtle changes in patients' vital signs, medical imaging, histopathology slides or biometrics, and propose efficient personalized treatments, based on combining patient outcomes with massive datasets of clinical data. This predictive capability of AI in both medical diagnostics and treatment is transforming the healthcare landscape by leading not only to better patient outcomes but also to significantly reduced healthcare costs.

AI in oncology

AI advancements have also gained a lot of significance in the realm of oncology by demonstrating their immense potential to enhance cancer diagnostic accuracy, improve early-stage detection rates and suggest the most effective tailor-made treatments. A PubMed search in May of 2022 of machine learning cross referenced with cancer revealed around 26,000 citations, more than 60% of these being published in the past five years, evincing the rapid expansion of the use of AI in cancer care. There have been numerous indicative studies in this field.

A study by Shaikh and Rao (2021) leveraged machine learning to spot minute and precise patterns in histopathological data by using various models. Notably, artificial neural networks (ANN), support vector machines (SVM), and decision trees (DT), were implemented to classify patients into high or low risk categories.

Islam et al. (2022) focused on breast cancer prediction by comparing the performance of DTs, random forests (RF), extreme gradient boosting (XGBoost), Naïve Bayes (NB), and more, to determine the most effective algorithm for classifying breast cancer using newly collected datasets. Their results showed that the RF and XGBoost achieved the highest accuracy of 94%, demonstrating the effectiveness of an ensemble of ML models to improve predictive performance.

Chip M. Lynch et al. (2017) focused on predicting lung cancer patient survival times by applying linear regression (LR), DTs, Gradient Boosting Machines (GBM), SVMs, and a custom ensemble model to analyze attributes such as tumor grade, size, gender, age, and stage, thus treating the survival predictions as a continuous target, rather than a classification problem. Among their set of models, GBM performed the best with a Root Mean Squared Error (RMSE) of 15.32, while the DTs struggled, likely due to limited discrete outputs.

Panayiotis Petousis et al. (2016) assisted in informing decisions about lung cancer screening by developing and evaluating Dynamic Bayesian

Networks (DBNs) and leveraging longitudinal data for enhanced decision making. The team incorporated factors such as demographics, smoking history, cancer risk factors, and LDCT screening outcomes. Their study showed that DBNs outperformed logistic regression (LR) and NB, evincing strong predictive accuracy and reliability in identifying high-risk lung cancer patients.

AI cancer prediction models typically utilize a variety of data features to help detect the disease at its earliest possible stage, basing decisions purely on objective facts by comparing past and present cases. Such a prediction classifier model can play a huge role in the healthcare industry, with its potential use as a quick, real-time predictor, helping not only make an educated prediction on a diagnosis, but also potentially correcting doctors' accidental and systematic errors (Conger, 2025).

AI in early pancreatic cancer detection

In the field of pancreatic cancer early detection, AI-assisted diagnostic techniques are also gaining more attention, particularly in image-based detection, where AI tools can help in identifying pancreatic lesions in Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) images, which would be challenging to recognize or quantify by the human eye. However, such tests are not usually performed when there are no apparent symptoms, both due to financial reasons and risks associated with these tests. This becomes an issue in the case of pancreatic cancer, for which symptoms do not show until the late-stage phase of the illness. Therefore, the use of appropriate endogenous blood or urine biomarkers could be an essential aspect of the early diagnosis of pancreatic cancer, especially in high-risk populations which could regularly be monitored for these biomarkers.

Pancreatic cancer detection based on biomarkers is facing some challenges: pancreatic tumors are highly heterogeneous between individuals; singular biomarkers do not have high enough sensitivity, and there are currently no biomarkers validated for early detection of PDAC. Nevertheless, past studies have pointed to some certain biomarkers which show some potential to be included in a robust set of high-specificity biomarkers. These biomarkers could be further analyzed by AI for their association with pancreatic cancer, and, therefore, included in a routine test for early diagnosis of the disease. Some of these promising biomarkers are the proteins LYVE-1, REG1B, and TFF1, creatinine and plasma CA19-9 found in urine samples (Huang et al., 2022).

Objective of this research

This research paper aims to address the pressing challenge outlined to predict pancreatic cancer using past case data of urinary biomarkers, by developing four machine learning classifier models—Neural Network

(NN), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN). The NN helps establish a benchmark for leveraging nonlinear dependencies in the data, as is later explored, providing a ‘standard.’ The other three models help evaluate and strengthen the reliability of the results and enhance the robustness of the predictions. However, solely relying on any of the four fundamental models can introduce doubts and limitations, therefore a Multiplicative Weight Update (MWU) system was developed to combine and aggregate the predictions of individual classifiers and enhance overall accuracy. The MWU combines the strengths of each individual model, assigning weights to each based on their testing data performance, and iteratively updating said weights to improve the final prediction accuracy. In turn, the MWU allows the most statistically reliable predictions to have a greater toll on the final prediction, creating a more robust and accurate classification system.

Results

A variety of scores were achieved by each model. Evaluation of the results was based solely on model accuracy for the following reason: the task involved a three-class classification problem (classes 0, 1, and 2), with the dataset being fully balanced (each class comprising one-third of the data). Consequently, evaluation metrics such as precision, recall, or F1 score were not considered, as they are typically applied in binary classification tasks with imbalanced datasets.

For the Neural Network, the following hyperparameter combinations were used in attempt to obtain the highest testing accuracy:

```
hidden_layer_options = [(128, 64, 32), (500, 250, 150, 2), (256, 128,
64)]
learning_rate_options = [0.001, 0.01, 0.1]
alpha_options = [0.0001, 0.001, 0.01]
```

The best outcome was achieved by the hyperparameters below, reaching a testing accuracy of 85.1%

```
'hidden_layer_sizes': (128, 64, 32)
'learning_rate_init': 0.1
'alpha': 0.0001
'epochs_trained': 24
```

For the Decision Tree, the max_depth was optimized with a step of two, and the most accurate test accuracy came out at 98.7%, with a max_depth of 20.

For the Random Forrest, different values for the number of estimators and the max_depth were experimented, and the best performance was achieved when n_estimators was 100 and max_depth was 22, reaching 96.2% testing accuracy.

Finally, K-Nearest Neighbor's prediction accuracy was tested with different values for the number of neighbors, the highest accuracy reaching 98.1% at 5 neighbors.

Using the MWU, each model was assigned an initial weight of 1, but the weight was dynamically adjusted based on each distinctive model's performance. Below are the final weights of each model:

Model	Weight
NN	165 (22.45%)
DT	192 (26.12%)
RF	187 (25.44%)
KNN	191 (25.99%)

TABLE 1: Table showing the different weights of each model at the end of testing.

Materials and method

Data set & features

The data set used was imported from Kaggle as a CSV file (Davis, 2021). Specifically, by importing the Kaggle library, the data set could then be accessed using just its URL link.

The dataset consists of 590 individuals, with key features including four urinary proteomic biomarkers: LYVE1, REG1B, TFF1, and creatinine (Debernardi et al., 2020). While LYVE1, REG1B, and TFF1 serve as potential biomarkers for pancreatic cancer, creatinine is used for normalization to account for variations in urine concentration (Yip-Schneider et al., 2020).

Along with urinary biomarkers, the model also uses other features from the dataset to improve its accuracy and reliability, including the sex of the patients, their age, and their Plasma CA19-9.

Firstly, the lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1) biomarker is a glycoprotein found mostly in lymphatic endothelial cells. LYVE1 is commonly associated with lymphatic vessel function, however recent studies have explored its potential as a biomarker in various diseases, including cancer (Jackson, 2018). In the context of pancreatic cancer, LYVE1 has been identified as a non-invasive biomarker due to its changed expression in early-stage malignancies. For more information about LYVE1, see Appendix A.1. Figure 2 demonstrates the very strong correlation between elevated LYVE1 levels and PDAC, showing a p value of 0.0006.

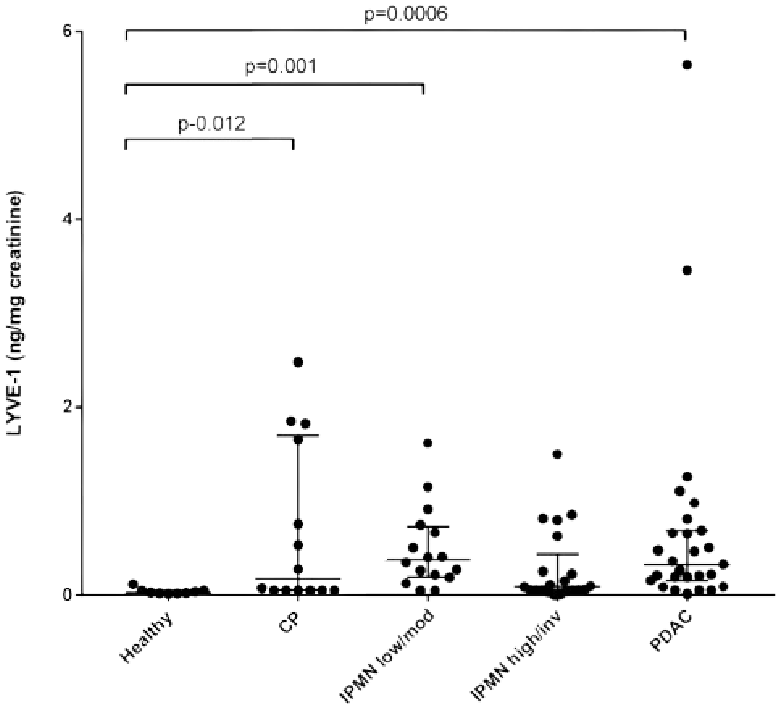


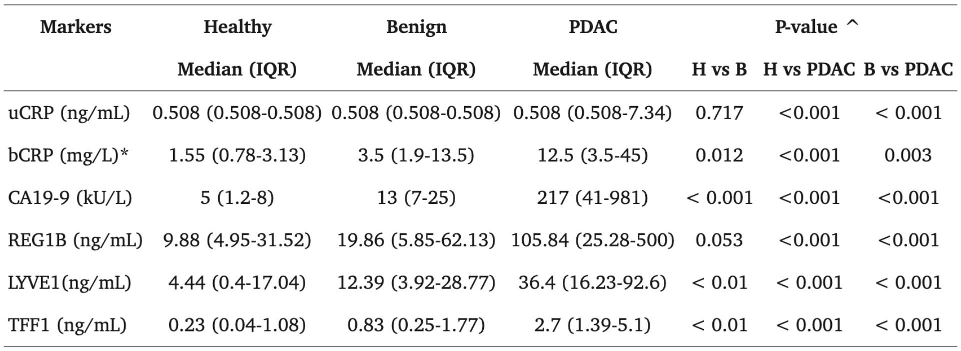
FIGURE 2. Graph showing LYVE-1 levels in differently diagnosed patients (Yip-Schneider et al., 2020).

Additionally, a study by Ali N et al in 2024 showed that elevated REG1B levels in blood and especially urine correlate with early-stage PDAC, likely because of the cancer's impact on pancreatic tissue, triggering regenerative and inflammatory responses that lead to higher REG1B secretion. For more information about REG1B, see Appendix A.2.

Furthermore, TFF1, being a small secretory protein, also exhibits elevated levels in PDAC patients. Ali N et al.'s study has suggested that increased urinary TFF1 levels are associated with early stage PDAC, likely due the cancer's influence on the gastrointestinal environment and epithelial cell turnover, causing secretion of TFF1. For more information about TFF1, see Appendix A.3.

Moreover, in PDAC patients, CA 19-9 becomes heightened because of increased tumor cell secretion and impaired clearance caused by biliary obstruction, a common circumstance of PDAC. Although high CA 19-9 levels can also be observed in non-cancerous pancreatic conditions, steadily and consistently elevated levels are strong associated with PDAC. For more information about CA 19-9, see Appendix A.4.

Table 2 shows the biomarker concentrations in different clinical groups (healthy, non-cancerous condition, and PDAC), demonstrating that PDAC patients exhibit significantly elevated biomarker levels compared to other groups.



Markers	Healthy	Benign	PDAC	P-value ^		
	Median (IQR)	Median (IQR)	Median (IQR)	H vs B	H vs PDAC	B vs PDAC
uCRP (ng/mL)	0.508 (0.508-0.508)	0.508 (0.508-0.508)	0.508 (0.508-7.34)	0.717	<0.001	< 0.001
bCRP (mg/L)*	1.55 (0.78-3.13)	3.5 (1.9-13.5)	12.5 (3.5-45)	0.012	<0.001	0.003
CA19-9 (kU/L)	5 (1.2-8)	13 (7-25)	217 (41-981)	< 0.001	<0.001	<0.001
REG1B (ng/mL)	9.88 (4.95-31.52)	19.86 (5.85-62.13)	105.84 (25.28-500)	0.053	<0.001	<0.001
LYVE1(ng/mL)	4.44 (0.4-17.04)	12.39 (3.92-28.77)	36.4 (16.23-92.6)	< 0.01	< 0.001	< 0.001
TFF1 (ng/mL)	0.23 (0.04-1.08)	0.83 (0.25-1.77)	2.7 (1.39-5.1)	< 0.01	< 0.001	< 0.001

Table 2: Biomarker concentrations across different groups (Ali et al., 2024).

Finally, creatinine is a microfluidic waste product formed by the breakdown of creatine phosphate in muscles. In the field of pancreatic cancer, creatinine is not a direct urinary biomarker, but it is often measured to standardize biomarker concentrations, such as TFF1, REG1B, AND LYVE1. Normalizing typical biomarker levels against creatinine helps achieve more accurate comparisons between patients, as differences in biomarker levels due to kidney activity are accounted for and standardized (Yip-Schneider et al., 2020). For more information about creatinine, see Appendix A.5.

By analyzing these features, the models predict the diagnosis of a patient, which returns either 1 (healthy sample), 2 (benign hepatobiliary disease - non-cancerous pancreas condition), or 3 (pancreatic cancer disease).

However, certain features used as input for the model's training often had null values, meaning some values were missing. Since the model cannot train itself on non-existent data, all null values had to be replaced by a value of 0. Additionally, the sklearn model can only be trained on number values. Therefore, to include the sex as a feature, all "F" and "M" values were replaced by -1 and 1 for female and male, respectively.

By operating at the cross-section of computational and medical domains, this study offers valuable insights for clinicians by enabling the discovery of novel patterns in biomarker data that may not have been previously apparent. Moreover, models implemented in the study, notably DTs and RFs, have strong interpretability, allowing clinicians to understand the rationale behind ML predictions. This system helps support more informed medical decisions and helps avoid the downfalls of relying on a 'black box' of a healthcare system.

Splitting the data

When creating a machine learning model, the entire data set must be separated into training and testing data. When using data to train the model, it's difficult to know whether the model is truly learning or just memorizing the training data given (or otherwise called overfitting). Overfitting occurs when the model adapts itself and clings too closely to patterns in the training data, therefore struggling with unseen data. This is why the entire dataset must be split into training and testing sets, and as so the model's performance can truly be evaluated using data which it has never encountered before.

The data was split using the train_test_split function from sklearn, which randomly put 33% of the data towards testing purposes, and 66% of the data towards training purposes.

Building the Neural Network

A Neural Network consists of three components: the input layers (data inputted), the hidden layers, and one output layer. The aim of a neural network is to use parameters within the hidden layers in order to capture nonlinear dependencies between the input and output. This particular model uses such nonlinear dependencies in urinary biomarkers to predict and output a patient's diagnosis.

Lots of different structures were experimented with to find the optimal topology that would result in the highest test accuracy, the final of which was a network 4 hidden layers (with depths of 256, 128, and 64, respectively), and as usual one output layer (diagnosis).

For more information about Neural Networks, see Appendix B.1.

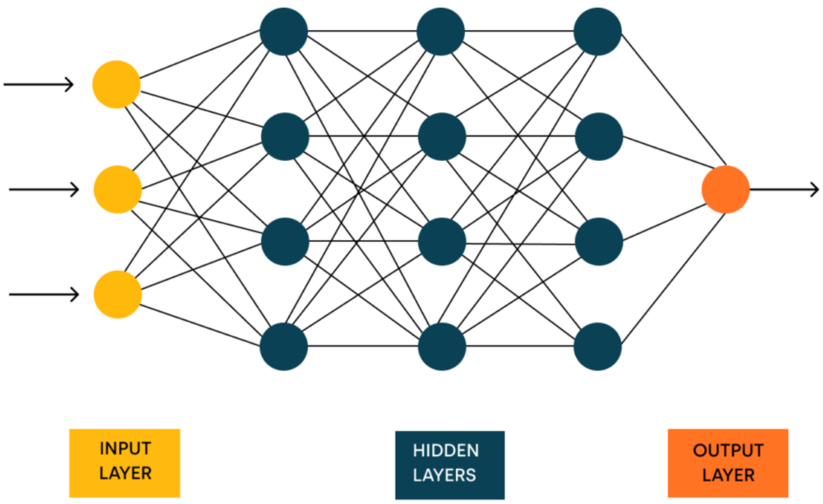


FIGURE 3: Visual demonstration of a simple Neural Network (Rojewska, 2023).

Building Decision Tree

A decision tree is a machine learning model that relies on a sequence of nested "if-else" statements to make predictions. These statements act as decision nodes, where the model evaluates certain conditions and criteria, and branches the input data into different paths depending on the outcome of each branch. At the end of every path is a leaf node, which provides the model's final prediction.

For more information about Decision Trees, see Appendix B.2.

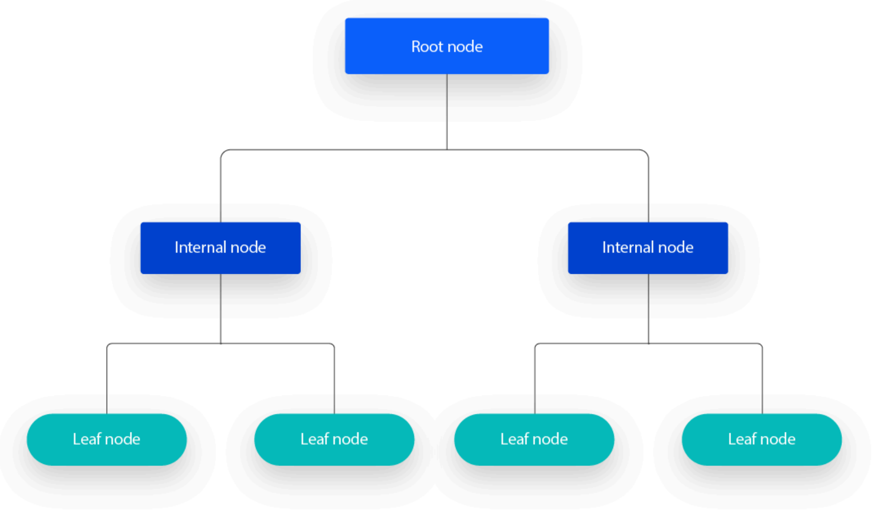


FIGURE 4: Visual demonstration of a simple Decision Tree (IBM, 2022).

Building K neighbors

A K-Nearest Neighbor (KNN) model operates by predicting the label of a data point by finding the majority class of its K closes neighbors in the feature space. To calculate the majority class, the KNN relies on a distance metric, such as Euclidean, to measure the similarities between two points. The KNN has a variety of hyperparameters to help increase its accuracy.

For more information about KNN models, see Appendix B.3.

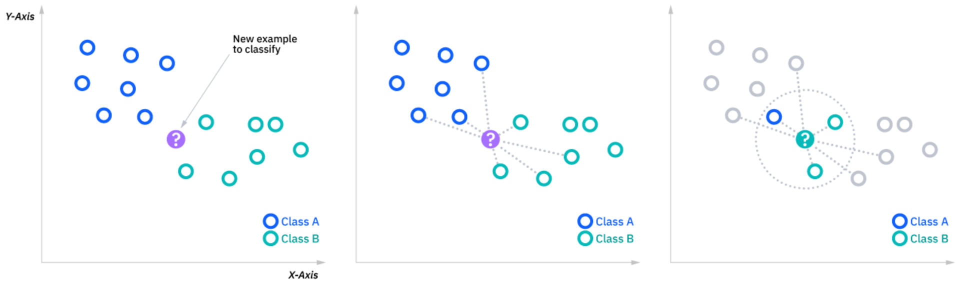


Figure 5: Visual demonstration of a KNN (IBM, 2025).

Building random forest

A random forest is fundamentally a collection of multiple decision trees, helping improve the predictions' overall robustness and accuracy. Decision trees come with a high risk of overfitting if their given depth is too high; therefore, by aggregating the predictions of many decision trees into one ensemble, the majority vote is likely to be more accurate.

For more information about Random Forests, see Appendix B.4.

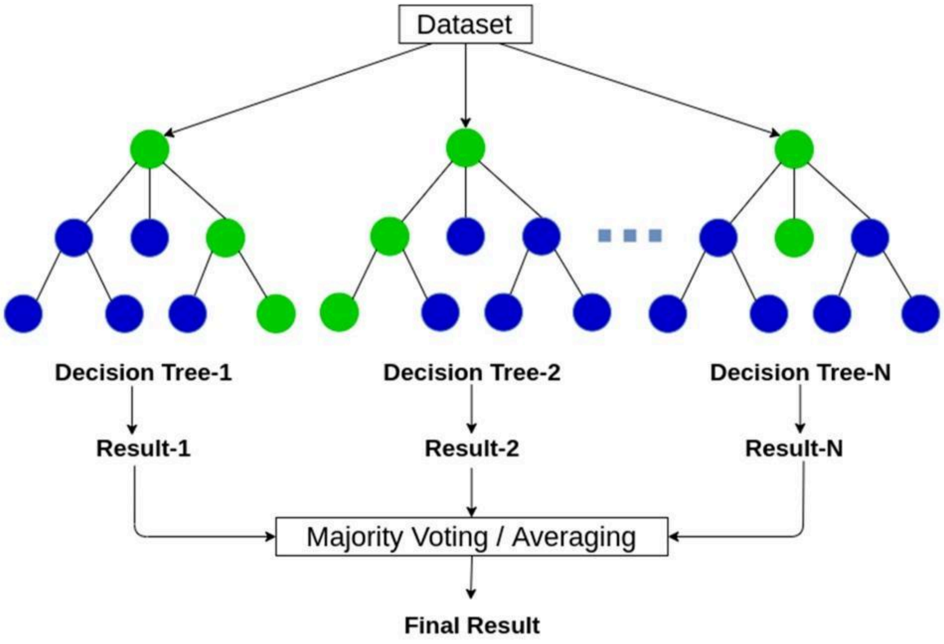


Figure 6: Visual demonstration of a Random Forres (Brital, 2021).

Multiplicative weight update

The Multiplicative Weight Update (MWU) serves as the final stage of the study. By employing this method, the program is able to combine the multiple classifier models programmed. Rather than treating all models equally, however, the MWU dynamically adjusts each model's influence based on its accuracy, ensuring that more reliable models contribute more significantly to the final prediction.

Initially, each model is assigned an equal weight of 1, but as predictions on the large testing data are made, correct classifications increase a model's weight, while misclassification keeps it unchanged. This adaptive process allows the MWU system to adjust its weighting strategy over time, prioritizing the best performing models. Consequently, models that consistently generated high accuracies become more influential in the final prediction, whereas those with recurrent misclassifications dynamically lose impact relative to other models.

In the end, the final classification is determined using a weighted majority vote, where each model's vote is weighted by its performance-based 'score'. By utilizing a weighted majority vote, the MWU system optimizes the decision-making process by undergoing accuracy-driven adjustments rather than arbitrarily keeping all the models' weights equal. As a result, this method ensures that the final classification prediction reflects the collective strengths of all models, in turn enhancing the robustness and reliability of the program.

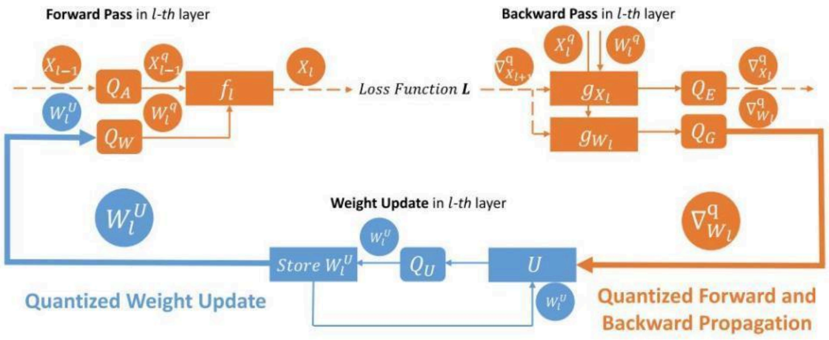


FIGURE 7: Illustration of Multiplicative Weight Update in model training with logarithmic number system (Anandkumar et al., 2022).

While MWU is a method widely used in machine learning applications, its use in the medical field has been relatively limited. There are however papers, such as that published by Chawla S. in 2020, which leverage an MWU-style framework to scale large LP relaxations in networked domains. Although somewhat experimental, MWU frameworks have been leveraged in many fields, and could potentially present significant advantages by combining models, each contributing its unique strengths.

Figure 8 illuminates how the algorithm is integrated: features taken in as input are passed separately into each of four models – DT, NN, RF, KNN – before these are weighed as part of the MWU to output a final predicted diagnosis.

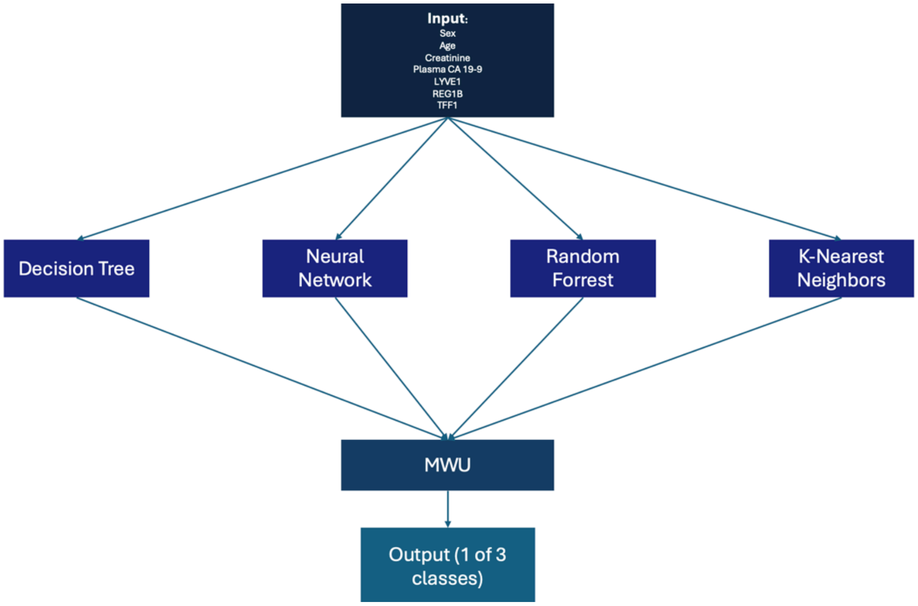


Figure 8: ML Flow diagram of Pancreatic Cancer Prediction model.

Conclusion

Methodology and key findings

This study aimed to address the challenge imposed by the difficulty of early pancreatic cancer detection by developing a machine learning based classification model that predicts a patient's diagnosis using biological indicators in bodily fluids, which have been previously identified as potentially promising for early detection of the disease. To accomplish this, four models were implemented and compared based on their predictive accuracy and performance: NN, DT, RF, and KNN. The models take patient age, sex, and urinary biomarker levels, to output a predicted classification of the patient as either healthy, having benign hepatobiliary disease, or pancreatic cancer. Taking into consideration the fact that each model had limitations, a Multiplicative Weight Update (MWU) method was applied to dynamically adjust each model's influence on the final prediction based on their accuracy, producing more robust and reliable final predictions. The results demonstrate with very high predictive accuracy the potential of AI-driven diagnostic tools in assisting early pancreatic cancer detection based on urine and blood biomarkers, helping potentially enhance the currently grim survival rates of pancreatic cancer.

Implications for practice

By applying the findings of this study to real-time practices, the promising and accessible set of urine and blood biomarkers used (LYVE1, REG1B, TFF1, Creatinine, Plasma CA19-9) could serve as a routine screening tool for the early detection of pancreatic cancer in high-risk individuals. These individuals would be identified based on factors such as their medical and family history, age and sex. Depending on the results of this initial screening, the healthcare system could then prioritize certain patients for additional diagnostic procedures, such as imaging tests, that are more costly and carry some risk, to detect the certain presence of the disease. This approach would enhance early detection of PDAC and therefore treatment effectiveness, while minimizing the financial burden on the healthcare system and the additional risks to the patients. If such an approach were to be implemented in healthcare, there would be a need to establish a framework defining the specific characteristics of high-risk individuals as well as the threshold of the biomarkers which would trigger the performance of additional diagnostic procedures. This would, in turn, probably require further research.

Limitations and future direction

The use of the four distinct classification models, along with the MWU method, distinguishes this research from previous studies on pancreatic cancer biomarkers and helps enhance its predictive performance. However, a limitation of the study is its inadequate sample size of just 590 individuals, relative to the global prevalence of pancreatic cancer. For context, in the United States alone, an estimated 107,988 people were living with pancreatic cancer in 2022 (NIH, 2011). This limited sample may have influenced the study's findings, particularly if the data reflects gene mutations which are specific in certain populations, thus limiting the ability to fully generalize the results. Future studies on early detection of pancreatic cancer should therefore address this issue by increasing the sample size.

Ethical considerations

In this study, we ensured that the dataset was balanced by sex (50% male, 50% female) to minimize gender-related bias. Additionally, the dataset does not contain any features that directly identify individual participants, hence protecting privacy and data security.

Expanding the horizons of AI in healthcare and its ethical implications

This study also prompts the consideration of how every model mentioned in this research can be used in a wider spectrum. For instance, one could explore the implementation of AI to make predictions about a different factor, such as classifying patients based on their cancer risk. On a more global scale, machine learning could be exploited by the healthcare and pharmaceutical industries in a variety of applications. With remarkable capabilities in pattern spotting, such algorithms could be leveraged to advance research on gene mutations, optimize vaccine development, or enhance personalized medicine by tailoring treatments to individual patient profiles based on genetic and biomarker data. The list of applications in healthcare and pharmaceuticals is extensive.

Despite AI's unparalleled advantages in healthcare, it's ethical implications in clinical fields should not be ignored. There are several concerns such as the risk that an AI algorithm may include bias towards a gender or race, because of heterogeneity between a dataset representing a given cancer population and other patients. Another concern is the need for researchers and healthcare organizations to protect data for patient privacy. In addition, healthcare systems should ensure equitable access for all patients to the benefits of AI-driven tools (Hantel et al., 2022). To alleviate these concerns, the development of standards and processes for AI's ethical development and application in healthcare is of utmost importance.

References:

Ali. (2016, February 5). 10 year survival rates improving for most cancers but sadly not pancreatic cancer. Pancreatic Cancer Action. https://pancreaticcanceraction.org/news/10-year-survival-rates-improving-for-most-cancers-but-sadly-not-pancreatic-cancer/

Ali, N., Debernardi, S., Kurotova, E., Tajbakhsh, J., Gupta, N. K., Pandol, S. J., Wilson, P., Pereira, S. P., Greenhalf, B., Blyuss, O., & Crnogorac-Jurcevic, T. (2024). Evaluation of urinary C-reactive protein as an early detection biomarker for pancreatic ductal adenocarcinoma. Frontiers in Oncology, 14, 1450326.

Barth, S. (2020). Artificial Intelligence (AI) in Healthcare & Medical Field. ForeSee Medical. https://www.foreseemed.com/artificial-intelligence-in-healthcare

Brital, A. (2021, September 20). Random Forest Algorithm Explained . Anas Brital. https://anasbrital98.github.io/blog/2021/Random-Forest/

Cancer of the Pancreas - Cancer Stat Facts. (n.d.). SEER. Retrieved June 19, 2025, from https://seer.cancer.gov/statfacts/html/pancreas.html

Conger, K. (2025, January 8). Unique Stanford Medicine-designed AI predicts cancer prognoses, responses to treatment. https://med.stanford.edu/news/all-news/2025/01/ai-cancer-prognosis.html

Debernardi, S., O'Brien, H., Algahmdi, A. S., Malats, N., Stewart, G. D., Plješa-Ercegovac, M., Costello, E., Greenhalf, W., Saad, A., Roberts, R., Ney, A., Pereira, S. P., Kocher, H. M., Duffy, S., Blyuss, O., & Crnogorac-Jurcevic, T. (2020). Urinary biomarkers for pancreatic cancer [Data set]. In Can a simple urine test detect one of the deadliest cancers? https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer eBioMedicine Contributors. (2022). Emerging biomarkers for early diagnosis of pancreatic cancer. EBioMedicine, 79(104064), 104064.

Hantel, A., Clancy, D. D., Kehl, K. L., Marron, J. M., Van Allen, E. M., & Abel, G. A. (2022). A process framework for ethically deploying artificial intelligence in oncology. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 40(34), 3907–3911.

Huang, B., Huang, H., Zhang, S., Zhang, D., Shi, Q., Liu, J., & Guo, J. (2022). Artificial intelligence in pancreatic cancer. Theranostics, 12(16), 6931–6954.

IBM. (2024, December 19). What is Overfitting? IBM. https://www.ibm.com/think/topics/overfitting

IBM. (2025, January 22). What is a Decision Tree? https://www.ibm.com/think/topics/decision-trees

IBM Contributors. (2025, February 12). What is the k-nearest neighbors algorithm? IBM. https://www.ibm.com/think/topics/knn

Islam, T., Kundu, A., Khan, N. I., Bonik, C. C., Akter, F., & Islam, M. J. (2022). Machine learning approaches to predict breast cancer: Bangladesh perspective. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2206.14972

Jackson, D. G. (2018). Hyaluronan in the lymphatics: The key role of the hyaluronan receptor LYVE-1 in leucocyte trafficking. Matrix Biology: Journal of the International Society for Matrix Biology, 78–79, 219–235.

Karar, M. E., El-Fishawy, N., & Radad, M. (2023). Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks. Journal of Biological Engineering, 17(1), 28.

Kohara, Y., Yasuda, S., Nagai, M., Nakamura, K., Matsuo, Y., Terai, T., Doi, S., Sakata, T., & Sho, M. (2024). Prognostic significance of creatine kinase in resected pancreatic cancer. Journal of Hepato-Biliary-Pancreatic Sciences, 31(12), 906–916.

Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. International Journal of Medical Informatics, 108, 1–8.

NIH Contributors. (2011). SEER Incidence Data, 1975-2021. SEER. https://seer.cancer.gov/data/index.html

NIH Contributors. (2019). NLST - The Cancer Data Access System. https://cdas.cancer.gov/nlst/

Page, P. (2023, September 13). A Novel Pancreatic Cancer Early Diagnostic Technology. https://blog.crownbio.com/a-novel-pancreatic-cancer-early-diagnostic-technology

Petousis, P., Han, S. X., Aberle, D., & Bui, A. A. T. (2016). Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network. Artificial Intelligence in Medicine, 72, 42–55.

Rojewska, K. (2023, July 20). What are Neural Networks and What are Their Applications? Qtravel.Ai. https://www.qtravel.ai/blog/what-are-neural-networks-and-what-are-their-applications/

Samir, S., El-Ashry, M., Soliman, W., & Hassan, M. (2024). Urinary biomarkers analysis as a diagnostic tool for early detection of pancreatic adenocarcinoma: Molecular quantification approach. Computational Biology and Chemistry, 112(108171), 108171.

SciKit Contributors. (2011). KNeighborsClassifier. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier.kneighbors

SciKit Contributors. (2012). RandomForestClassifier. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Shaikh, F. J., & Rao, D. S. (2022). Prediction of Cancer Disease using Machine learning Approach. Materials Today: Proceedings, 50, 40–47.

Yip-Schneider, M. T., Soufi, M., Carr, R. A., Flick, K. F., Wu, H., Colgate, C. L., & Schmidt, C. M. (2020). Performance of candidate urinary biomarkers for pancreatic cancer - Correlation with pancreatic cyst malignant progression? American Journal of Surgery, 219(3), 492–495.

Zhao, J., Dai, S., Venkatesan, R., Zimmer, B., Ali, M., Liu, M.-Y., Khailany, B., Dally, W., & Anandkumar, A. (2022). LNS-Madam: Low-Precision Training in Logarithmic Number System Using Multiplicative Weight Update. Research Invidia. https://research.nvidia.com/publication/2022-12_lns-madam-low-precision-training-logarithmic-number-system-using-multiplicative

Appendix A:

Urinary Biomarker details:

1. **LYVE1:** The lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1) biomarker is a glycoprotein found mostly in lymphatic endothelial cells. LYVE1 is commonly associated with lymphatic vessel function, however recent studies have explored its potential as a biomarker in various diseases, including cancer (Jackson, 2018). In the context of pancreatic cancer, LYVE1 has been

identified as a non-invasive biomarker due to its changed expression in early-stage malignancies. More specifically, researchers have found that urinary LYVE1 levels are elevated in patients with pancreatic ductal adenocarcinoma (PDAC) - the most common and aggressive type of pancreatic cancer - compared to healthy individuals (Yip-Schneider et al., 2020). This is likely due to the cancer's impact on the lymphatic system, which in turn leads to increased secretion of LYVE1 into bodily fluids such as urine. Although not enough on its own, the measurement of urinary LYVE1 levels provides a potential non-invasive approach for early detection.

2. **REG1B:** The regenerating islet-derived protein 1 beta (REG1B) biomarker is a secretory protein primarily expressed in the pancreas and gastrointestinal tract. A study by Ali N et al in 2024 showed that elevated REG1B levels in blood and especially urine correlate with early-stage PDAC, likely because of the cancer's impact on pancreatic tissue, triggering regenerative and inflammatory responses that lead to higher REG1B secretion.

   It belongs to a family of proteins (REG) which are typically involved in tissue regeneration, cell proliferation, and inflammation. Particularly, REG1B plays role in maintaining pancreatic function, particularly in response to injury or stress. Looking at its role in pancreatic cancer detection, REG1B can act as a potential biomarker due to its increased expression in tumor cells.

3. **TFF1:** The Trefoil Factor 1 (TFF1) biomarker is a small secretory protein, expressed mostly in the mucosal lining of the gastrointestinal tract, which plays a role in mucosal protection, repair, and cell migration. TFF1 can act as a potential indicator of pancreatic cancer due to its altered expression in tumor cells. Ali N et al.'s study has suggested that increased urinary TFF1 levels are associated with early stage PDAC, likely due the cancer's influence on the gastrointestinal environment and epithelial cell turnover, causing secretion of TFF1.

4. **Plasma CA 19-9:** Plasma CA 19-9, while not a urinary biomarker, is a blood-based tumor marker which is commonly used as a means of pancreatic cancer detection. Plasma CA 19-9 measures the blood plasma levels of CA 19-9 monoclonal antibody, a glycoprotein produced by pancreatic ductal epithelial cells. In PDAC patients, CA 19-9 becomes heightened because of increased tumor cell secretion and impaired clearance caused by biliary obstruction, a common circumstance of PDAC. Although high CA

19-9 levels can also be observed in non-cancerous pancreatic conditions, steadily and consistently elevated levels are strong associated with PDAC.

5. **Creatinine:** Finally, creatinine is a microfluidic waste product formed by the breakdown of creatine phosphate in muscles. It gets filtered out by the kidneys and excreted in urine, therefore acting as a widely recognized marker for kidney health. In the field of pancreatic cancer, creatinine is not a direct urinary biomarker, but it is often measured to standardize biomarker concentrations, such as TFF1, REG1B, AND LYVE1. Taking into consideration the fact that urine dilution can vary based on hydration and kidney function, normalizing typical biomarker levels against creatinine helps achieve more accurate comparisons between patients, as differences in biomarker levels due to kidney activity are accounted for and standardized (Yip-Schneider et al., 2020).

Appendix B:

ML Model details:

1. **Neural Network:** A Neural Network consists of three components: the input layers (data inputted), the hidden layers, and one output layer. The aim of a neural network is to use parameters within the hidden layers to capture nonlinear dependencies between the input and output. This model uses such nonlinear dependencies in urinary biomarkers to predict and output a patient's diagnosis.

   Lots of different structures were experimented with to find the optimal topology that would result in the highest test accuracy, the final of which was a network 4 hidden layers (with depths of 256, 128, and 64, respectively), and as usual one output layer (diagnosis).

   There were hyperparameters that could also be taken into consideration when creating this model, as each could affect the model's testing accuracy. The 'learning rate' hyperparameter, for example, determines how aggressively the network changes its weights (within hidden layers) during its training. With a high learning rate, the weights would be changed drastically during the training process, and vice versa with a low training rate. This model uses a constant learning rate which does not change throughout the training phase.

2. **Decision Tree:** A decision tree is a machine learning model that relies on a sequence of nested "if-else" statements to make predictions. These statements act as decision nodes, where the model evaluates certain conditions and criteria, and branches the input data into different paths depending on the outcome of each branch. At the end of every path is a leaf node, which provides the model's final prediction.

   In this model, the maximum depth of the tree, which determines the maximum number of decision nodes along any path before reaching a leaf node, was carefully tuned to maximize the testing accuracy. Experimentation with various maximum depths was conducted, varying from a depth of 1 to 20, because, although a deeper tree can better capture complex patterns in the data, it also risks overfitting to the training data. On the other hand, a tree too shallow might generalize better but could underfit the data, leading to lower accuracy.

   Finally, after testing all depths in said range, a decision of keeping a maximum depth of 17 was reached as this achieved the highest test accuracy without overfitting to the training data.

3. **K-Nearest Neighbors:** A K-Nearest Neighbor (KNN) model operates by predicting the label of a data point by finding the majority class of its K closes neighbors in the feature space. To calculate the majority class, the KNN relies on a distance metric, such as Euclidean, to measure the similarities between two points. The KNN has a variety of hyperparameters to help increase its accuracy:

   Most importantly, the n neighbors parameter determines the number of nearest data points considered when making a prediction, with a higher value leading to closer decision boundaries but also potentially reducing the model’s sensitivity to different local patterns.

   After testing several values for n neighbors, measuring the four closest data points achieves the highest test accuracy.

4. **Random Forrest:** A random forest is fundamentally a collection of multiple decision trees, helping improve the predictions’ overall robustness and accuracy. Decision trees come with a high risk of overfitting if their given depth is too high; therefore, by aggregating the predictions of many decision trees into one ensemble, the majority vote is likely to be more accurate.

The model takes several hyperparameters, notably max depth and n estimators. Like the decision tree, max depth controls the depth of each decision tree, and n estimators controls how many different trees make up the random forest.