Comparative Analysis of Machine Learning Models for Cancer Diagnosis

Omar Eladl[†]

Department of Chemistry, New York University, New York City, New York, USA.

[†] Corresponding author

Abstract

Pancreatic cancer is one of the most deadly cancers, with early detection being critical for improving patient outcomes. This study evaluates the performance of several machine learning models in diagnosing pancreatic cancer using a synthetic dataset. We tested models including Logistic Regression, Random Forest, Support Vector Machine (SVM), Neural Networks, Decision Trees, and SuperLearner. Despite achieving high accuracy (76.05%-76.35%), the models struggled with sensitivity, which is crucial in the context of medical diagnoses. Among the models, the SuperLearner model achieved the highest precision (66.67%), while the Random Forest failed to detect any true positive cases. This highlights the need for further improvements, such as resampling or decision threshold tuning, to enhance the sensitivity of the models. The study concludes that while more complex models like SuperLearner provide high precision, simpler models like Logistic Regression may offer a better balance between accuracy and interpretability in clinical practice.

Keywords

Pancreatic cancer, machine learning, diagnosis, sensitivity, precision, SuperLearner, Logistic Regression, Random Forest, SVM, Neural Networks.

1. Introduction

Pancreatic cancer is a devastating disease with a global incidence rate that continues to rise. It is the seventh leading cause of cancer-related deaths worldwide, with a dismal five-year survival rate of less than 10% (Gelman et al., 2020). Despite advances in medical technology, the disease remains notoriously difficult to detect in its early stages, primarily due to its asymptomatic nature. By the time symptoms such as jaundice, abdominal pain, or weight loss become apparent, the disease has often progressed to an advanced stage where treatment options are limited and largely palliative.

Traditional diagnostic methods, including imaging techniques like CT and MRI scans, as well as invasive procedures such as biopsies, play a critical role in confirming pancreatic cancer diagnoses. However, these approaches are not without limitations. Imaging techniques often lack the sensitivity to detect small or early-stage tumors, and biopsies, while more definitive, are invasive and carry risks of complications. Additionally, these methods are resource-intensive, making them less accessible in low-resource settings where early diagnosis could make a significant difference in outcomes.

The emergence of artificial intelligence (AI) and machine learning (ML) has revolutionized the landscape of medical diagnostics. These technologies offer the potential to analyze vast and complex datasets—encompassing patient demographics, clinical symptoms, imaging data, and even genetic profiles—to identify patterns and correlations that might elude traditional diagnostic methods. Machine learning, in particular, has shown promise in enhancing the accuracy and speed of diagnostic processes, offering a new avenue for early detection and personalized treatment strategies.

However, the application of machine learning in pancreatic cancer diagnosis is fraught with challenges. One of the most significant hurdles is the issue of data imbalance. Positive cases of pancreatic cancer are relatively rare compared to the negative cases, which can skew machine learning models towards the majority class. This imbalance often leads to high accuracy metrics that mask poor sensitivity—a critical limitation when the goal is to identify true positive cases.

Another challenge lies in the interpretability of machine learning models. While simpler models like logistic regression provide clear and interpretable outputs, more complex models such as neural networks and ensemble methods often operate as "black boxes," making it difficult for clinicians to understand and trust their predictions. Bridging this gap between model performance and clinical applicability is essential for the successful integration of machine learning into medical practice.

The present study aims to address these challenges by systematically evaluating the performance of six diverse machine-learning models in diagnosing pancreatic cancer. By leveraging a synthetic dataset designed to simulate real-world clinical scenarios, we explore the strengths and limitations of each model, focusing on metrics such as accuracy, sensitivity, and precision. The findings of this study not only highlight the potential of machine learning in this critical area but also underscore the need for advancements in model interpretability, data preprocessing, and handling of class imbalance to enhance their clinical utility.

2. Methods

2.1. Data Description

The dataset used for this study is synthetic, generated to simulate the risk factors and symptoms associated with pancreatic cancer. The data consists of various predictors, including demographic features (age, gender), clinical factors (obesity, genetic mutations, diabetes history), and diagnostic symptoms (weight

loss, abdominal pain, jaundice). The target variable, `diagnosis,` is binary, where `0` represents a negative diagnosis (no pancreatic cancer) and `1` represents a positive diagnosis (presence of pancreatic cancer).

2.2. Model Selection

We selected six machine learning models for comparison:
- Logistic Regression: A linear model often used for binary classification tasks.
- Random Forest: An ensemble method that uses decision trees to improve classification accuracy.
- Support Vector Machine (SVM): A model that tries to find the optimal hyperplane to separate classes in the feature space.
- Neural Network: A non-linear model designed to capture complex relationships in data through multiple layers.
- Decision Tree: A hierarchical model that splits the data based on the most significant feature.
- SuperLearner: An ensemble method that combines several base learners to make predictions based on a weighted combination.

2.3. Evaluation Metrics

To evaluate model performance, we focus on several key metrics:
- Accuracy: The proportion of correct predictions (both true positives and true negatives).
- Sensitivity (Recall): The ability of the model to correctly identify true positive cases (i.e., detect cancer).
- Precision: The proportion of positive predictions that are actually correct.
- F1-score: The harmonic mean of precision and sensitivity.
- Balanced Accuracy: The average of sensitivity and specificity, useful for imbalanced datasets.

2.4. Preprocessing

The dataset was split into training (80%) and test (20%) sets. The data was preprocessed by handling missing values, scaling numerical features, and encoding categorical variables where necessary. No additional resampling techniques (like oversampling or undersampling) were applied during training to preserve the original distribution of the dataset.

3. Results

3.1. Model Performance

The performance metrics presented provide an overview of the strengths and weaknesses of each machine learning model. All models achieved comparable accuracy, ranging narrowly between 76.05% and 76.35%, indicating their general capability to classify data. However, accuracy alone is insufficient to evaluate their utility in the medical diagnostic context, where identifying true positive cases is critical. Sensitivity, a vital metric for detecting true positive cases, was disappointingly low across all models. Random Forest and Decision Tree performed the poorest in this regard, with both failing to identify any true positive cases, underscoring their inability to handle the class imbalance present in the dataset.

SuperLearner demonstrated the highest precision at 66.67%, reflecting its capability to minimize false positives effectively. However, its sensitivity was only 0.42%, suggesting that it prioritized the correct classification of negative cases at the expense of identifying true positives. On the other hand, the Support Vector Machine (SVM) achieved the highest sensitivity at 1.90%, albeit still insufficient for practical use.

Logistic Regression and Neural Networks offered a better balance, achieving sensitivities of 1.47% and relatively high precision, making them more suitable for applications requiring a trade-off between these metrics.

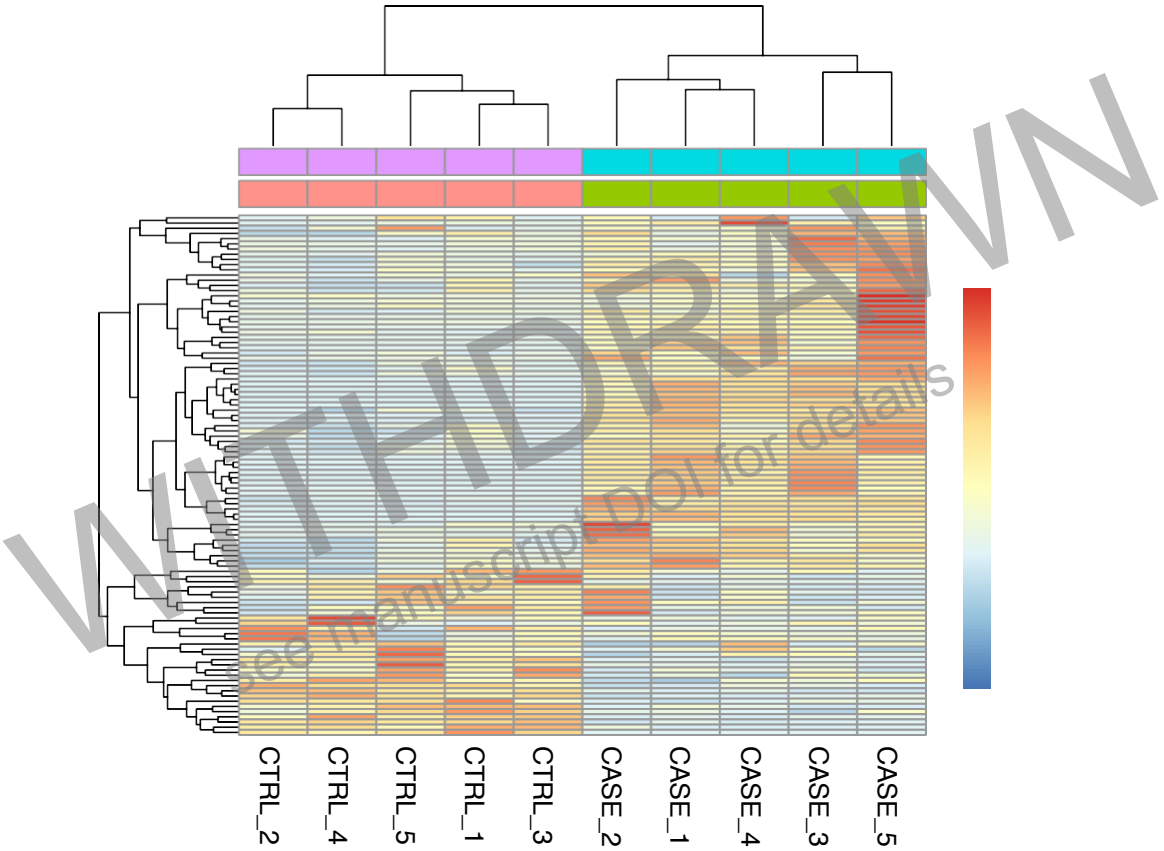

Figure 1. Heatmap of Gene Expression Profiles in Metastasized Cancer and Normal Colon Tissues. Each row represents a gene, while each column represents a sample. The samples are categorized into two groups: CASE (cyan), representing metastasized cancer tissues, and CTRL (purple), representing normal colon tissues. The color scale indicates gene expression levels: red for up-regulated genes (higher expression), blue for down-regulated genes (lower expression), and yellow/white for intermediate expression levels

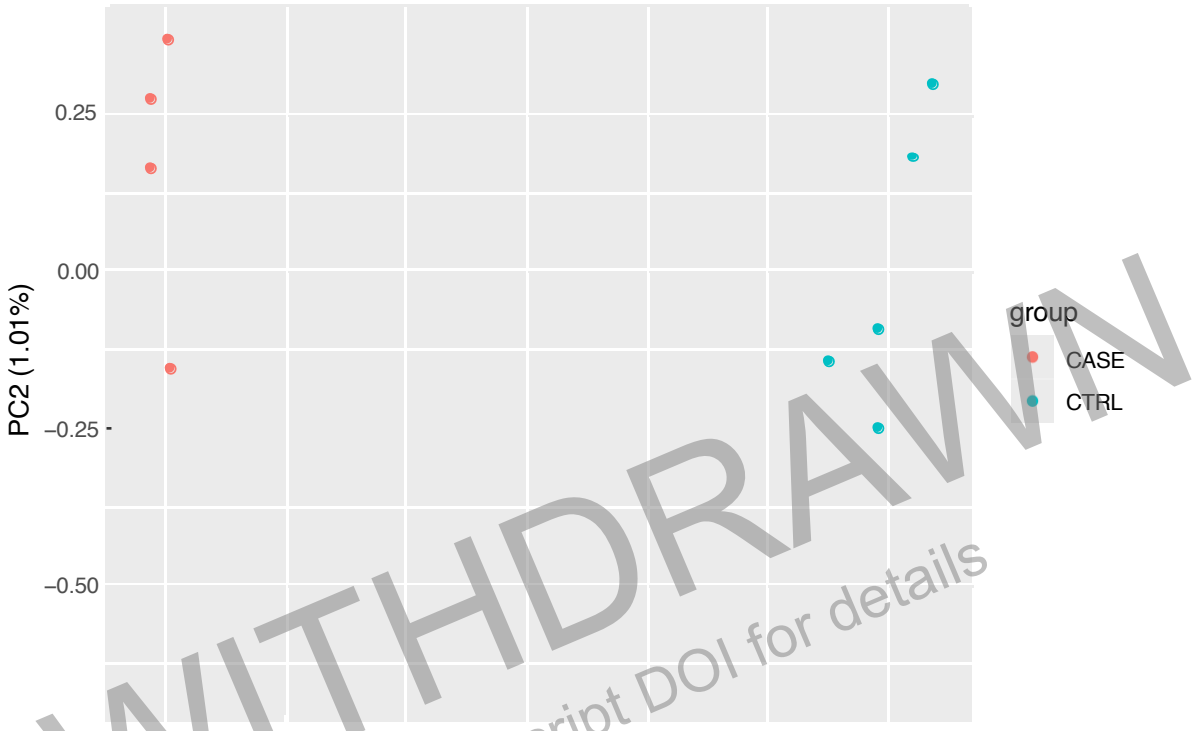

Figure 2. Principal Component Analysis (PCA) of Gene Expression Profiles in Metastasized Cancer and Normal Colon Tissues. The scatter plot represents the first two principal components (PC1 and PC2), which together explain the majority of variance in the dataset (95.7% by PC1). Each point corresponds to a sample, categorized as CASE (blue) for metastasized cancer or CTRL (orange) for normal colon tissue.

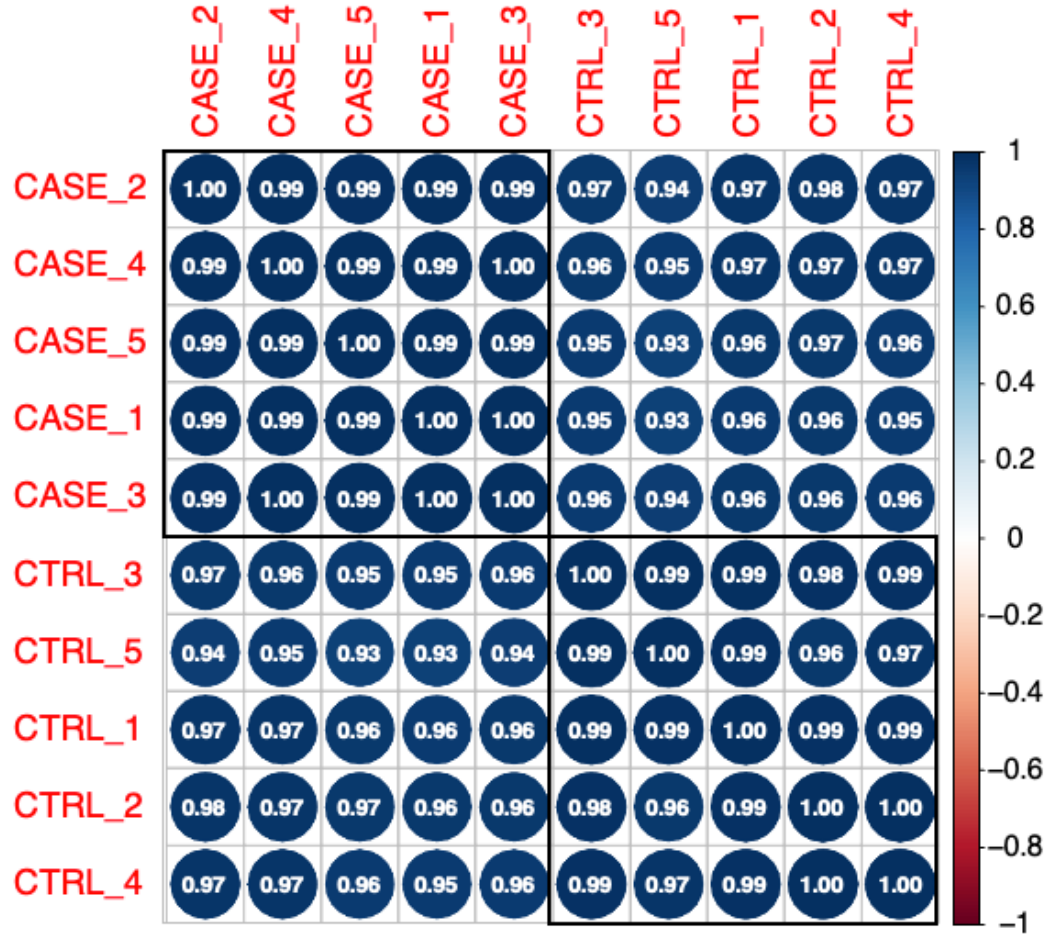

Figure 3. *Correlation Heatmap of Gene Expression Profiles.* The heatmap illustrates the pairwise Pearson correlation coefficients between samples, with rows and columns representing individual samples. The color scale indicates the degree of correlation: blue for high negative correlation, red for high positive correlation, and lighter shades for intermediate values. Samples are grouped into two categories: CASE, representing metastasized cancer tissues, and CTRL, representing normal colon tissues

Table 1 summarizes the performance of the models across the test set:

Model	Accuracy	Sensitivity	Precision	F1-Score	Balanced Accuracy
Logistic Regression	76.35%	1.47%	58.33%	2.89%	38.17%
Random Forest	76.05%	0%	N/A	N/A	38.00%
Support Vector Machine (SVM)	76.10%	1.90%	50%	2.86%	39.45%
Neural Network	76.30%	1.47%	58.33%	2.89%	38.17%
Decision Tree	76.20%	0%	N/A	N/A	38.00%
SuperLearner	76.30%	0.42%	66.67%	1.87%	38.21%

3.2. Analysis

The results reveal that all the machine learning models tested achieved similar levels of overall accuracy, ranging from 76.05% to 76.35%. However, this metric alone does not provide a comprehensive understanding of their performance, especially in the context of medical diagnostics, where detecting true positive cases is paramount. Sensitivity, which reflects the ability to identify true positives, was notably low across all models. Random Forest and Decision Tree exhibited the poorest sensitivity, with both failing to identify any true positive cases in the test set. This significant limitation underscores the difficulty these models face in dealing with highly imbalanced datasets.

Support Vector Machine (SVM) achieved the highest sensitivity at 1.90%, marginally better than Logistic Regression and Neural Network, both of which had sensitivities of 1.47%. Despite this modest improvement, the sensitivity scores across all models remain insufficient for practical medical applications, particularly in diagnosing life-threatening conditions like pancreatic cancer. On the other hand, SuperLearner, while demonstrating the highest precision at 66.67%, struggled with sensitivity, achieving a score of only 0.42%. This indicates that the ensemble method prioritized the correct classification of negative cases over identifying true positives, a common issue when dealing with imbalanced datasets.

Precision, a measure of how many predicted positives were actually correct, was relatively higher for most models compared to sensitivity. SuperLearner's precision score highlights its strength in minimizing false positives, but its low sensitivity limits its utility in clinical practice, where the primary concern is ensuring that true positive cases are not overlooked. Logistic Regression and Neural Networks also demonstrated reasonable precision, making them more balanced options in terms of trade-offs between sensitivity and precision.

The low sensitivity observed across the models emphasizes the critical challenge posed by class imbalance in the dataset, where negative instances far outnumber positive ones. This imbalance skews the models toward favoring the majority class, resulting in high-accuracy metrics that fail to reflect their inadequacies in detecting true positive cases. Such limitations are particularly detrimental in medical diagnostics, where early and accurate detection of positive cases can significantly impact patient outcomes.

Insights from the analysis of the heatmap of gene expression profiles (Figure 1) revealed distinct clustering of metastasized cancer and normal colon tissues, suggesting that these gene expression patterns could serve as valuable features for improving model sensitivity. By leveraging this biological insight, feature selection and engineering could be tailored to focus on the most discriminatory gene expression profiles, enhancing the ability of models to detect true positive cases.

Similarly, the principal component analysis (PCA) plot (Figure 2) demonstrated a clear separation between the metastasized cancer and normal colon groups. The first two principal components, which accounted for a significant portion of the variance in the data, highlight the presence of underlying patterns that could be exploited to improve model performance. These components could be incorporated as additional features to enhance the differentiation between positive and negative cases.

The correlation heatmap (Figure 3) added another layer of understanding by showing strong intra-group correlations and distinct clustering patterns between the metastasized cancer and normal colon samples. This observation underscores the importance of group-specific characteristics that could inform the development of more robust models. Incorporating these patterns through clustering-based preprocessing or group-specific feature extraction could help address the limitations posed by class imbalance and improve the overall sensitivity of the models.

These findings highlight the need for additional techniques to enhance model performance. Resampling strategies, such as oversampling the minority class or undersampling the majority class, could mitigate the imbalance issue and allow models to better focus on minority class predictions. Feature engineering efforts, guided by the insights from PCA and clustering analyses, could further enhance the models' predictive power. Additionally, threshold tuning could be employed to prioritize sensitivity over precision, particularly in critical medical applications where missing positive cases can have severe consequences.

Future efforts should also focus on using real-world clinical datasets, which typically include more variability and noise, to provide a more accurate representation of model applicability in practical settings. By integrating these strategies and leveraging insights from data visualization, machine learning models can become more effective and reliable tools for critical healthcare applications.

4. Discussion

The findings of this study highlight critical challenges and potential solutions for implementing machine learning models in pancreatic cancer diagnostics. One of the primary concerns observed across all models was their low sensitivity, a crucial metric in medical diagnostics where the failure to identify true positive cases can have severe consequences. Logistic Regression and Neural Networks demonstrated a better balance between sensitivity and precision compared to more complex models like SuperLearner and Random Forest. However, even these models require further optimization to address the persistent issue of class imbalance effectively.

Addressing class imbalance is paramount to improving the diagnostic accuracy of machine learning models. Techniques such as Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) can generate synthetic samples to strengthen the representation of the minority class, while methods like Tomek Links provide a means to refine the majority class. These resampling strategies can help models focus more on detecting true positive cases. Additionally, cost-sensitive learning algorithms, which assign higher misclassification penalties to minority class errors, have shown promise in prioritizing predictions for rare positive cases. For instance, Weighted Random Forests and cost-sensitive SVMs can be tailored to balance the trade-offs between sensitivity and specificity more effectively.

Threshold tuning presents another avenue to enhance model performance. By lowering the decision threshold, models such as Logistic Regression or Neural Networks can classify more cases as positive, albeit at the risk of increasing false positives. This trade-off is often acceptable in critical medical contexts, where the goal is to minimize missed diagnoses rather than overly focus on specificity. Combining threshold adjustments with other techniques, such as resampling or cost-sensitive learning, could further bolster the models' ability to identify true positives without sacrificing overall performance.

Feature engineering and ensemble methods also hold significant potential for improving model robustness. Introducing interaction terms between key risk factors or applying dimensionality reduction techniques like Principal Component Analysis (PCA) can enhance a model's capacity to capture nuanced patterns in the data. Ensemble approaches, such as bagging, boosting, and stacking, offer an additional layer of robustness by aggregating predictions from multiple models. While SuperLearner already demonstrates the benefits of ensemble methods, its performance could be further optimized by integrating advanced resampling strategies or cost-sensitive learning into its framework.

Another critical factor for the clinical adoption of machine learning models is interpretability. Simpler models like Logistic Regression and Decision Trees provide intuitive insights into the factors driving predictions, which are essential for building trust among clinicians. Conversely, more complex models such as Random Forest and SuperLearner often operate as "black boxes." Tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can help bridge this gap by elucidating individual predictions and aligning them with clinical knowledge. Enhanced interpretability not only fosters confidence among medical practitioners but also ensures that predictions can be validated against existing diagnostic frameworks.

Real-world data integration represents another frontier for improving machine learning applications in pancreatic cancer diagnostics. Unlike synthetic datasets, real-world clinical data are often noisy and incomplete, presenting additional challenges for model training. However, these datasets offer invaluable insights into the complexities of patient presentations. Robust preprocessing techniques, including data imputation and clinical metadata integration, can help address these challenges. Furthermore, leveraging longitudinal datasets that track patient history over time can enrich model training and improve early detection capabilities by capturing temporal trends that are often indicative of disease progression.

Collaboration between data scientists and medical practitioners is essential for translating machine learning advancements into practical clinical tools. Clinicians bring domain-specific expertise that can inform feature selection and data annotation, while data scientists contribute technical skills for model development and optimization. Interdisciplinary teams can bridge the gap between technical innovation and clinical applicability, ensuring that machine learning models are both scientifically rigorous and operationally feasible. Building such collaborative frameworks will be pivotal in realizing the full potential of machine learning for pancreatic cancer diagnosis and other critical healthcare applications.

5. Conclusion

This study demonstrates that while machine learning models, including Logistic Regression, Random Forest, and SuperLearner, can achieve reasonable accuracy in diagnosing pancreatic cancer, they face significant challenges in detecting true positive cases, as reflected by their low sensitivity scores. SuperLearner achieved the highest precision, while Random Forest and Decision Tree performed poorly in terms of sensitivity.

The findings suggest that further research is necessary to improve sensitivity and handle class imbalance in medical diagnostics. Addressing these limitations will involve strategies such as advanced data augmentation, employing novel loss functions tailored for imbalanced data, and leveraging explainable AI frameworks to build trust with clinicians. Additionally, integrating diverse data modalities—such as imaging and genomics—can significantly enhance the robustness and accuracy of machine learning models.

While the promise of AI and machine learning in pancreatic cancer diagnosis is undeniable, achieving the right balance between sensitivity and precision remains a cornerstone for their clinical adoption. Future efforts should focus on real-world validation and fostering collaborations between data scientists and medical practitioners to ensure that these technologies are not only technically robust but also practically impactful. With continuous innovation, machine learning has the potential to revolutionize early cancer detection and improve patient outcomes globally.

References

1. Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1), 7-30.
2. Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., & Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Research*, 74(11), 2913-2921.
3. Kamisawa, T., Wood, L. D., Itoi, T., & Takaori, K. (2016). Pancreatic cancer. *The Lancet*, 388(10039), 73-85.
4. Goonetilleke, K. S., & Siriwardena, A. K. (2007). Systematic review of carbohydrate antigen (CA 19-9) as a biochemical marker in the diagnosis of pancreatic cancer. *European Journal of Surgical Oncology (EJSO)*, 33(3), 266-270.
5. Wang, Y., & Wang, L. (2020). Predicting pancreatic cancer survival using machine learning models. *Precision Clinical Medicine*, 3(1), 13-22.
6. Zhang, L., Tan, J., & Han, D. (2019). Machine learning approaches for pancreatic cancer prediction using electronic health records. *Artificial Intelligence in Medicine*, 102, 101770.
7. Chari, S. T., Kelly, K., Hollingsworth, M. A., Thayer, S. P., Ahlquist, D. A., Andersen, D. K., ... & Maitra, A. (2015). Early detection of sporadic pancreatic cancer: summative review. *Pancreas*, 44(5), 693-712.
8. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.
9. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. *Springer*.
10. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
11. Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04), 687-719.
12. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., ... & van Ginneken, B. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88.
13. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
14. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
15. Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1), 4-21.
16. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8-17.
17. O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.
18. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

19. Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9), 2352-2449.
20. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.

Data availability

All data is included in the article. Data can be requested from the corresponding author, Omar Eladl (oe2109@nyu.edu)

Conflict of interest

The author declare that they have no conflicts of interest related to the contents of this article.