Evaluating Hyperparameter Tuned Machine Learning Classifiers for Early-Stage Pancreatic Cancer Detection via Urinary Biomarkers

Nivrith Ananth Iyer

Abstract

Pancreatic cancer accounts for only 3% of all cancers in the United States but remains one of the deadliest due to its asymptomatic progression and late-stage diagnosis. Early detection improves recovery rates by up to 44%, underscoring the importance of sensitive diagnostic tools. Current methods, such as blood tests, need more accuracy for early detection. This study investigates urinary biomarkers—creatinine, LYVE1, REG1B, and TFF1—as promising alternatives. A 590-sample dataset from the Spanish National Cancer Research Center was evaluated for classification accuracy using hyperparameter-tuned machine learning models, including XGBoost, LightGBM, Random Forest, Support Vector Machine, and 1D CNN-LSTM. Results showed XGBoost and LightGBM models achieving 91% accuracy, outperforming other classifiers. A discrepancy with a reported 97% accuracy for the 1D CNN-LSTM model in prior studies suggests parameter and dataset size differences. These findings support the potential of urinary biomarkers for early pancreatic cancer detection and highlight the efficacy of gradient-boosting models. Future work will explore larger datasets and the development of a urine-sample-based diagnostic device.

Keywords: **Pancreatic Cancer, Urinary Biomarkers, Early Detection, AI in Medicine, Diagnostic Tools**

Introduction

Pancreatic cancer, despite comprising about 3% of all cancers in the US, is a very deadly disease. It can be treated if the tumor is dePancreatic cancer is one of the most lethal cancers, with a five-year survival rate below 10%. This is largely attributed to the lack of effective early diagnostic tools, as most cases remain asymptomatic until metastasis. Early detection significantly improves patient outcomes, with recovery rates reaching up to 44%. Urinary biomarkers have emerged as a promising alternative for early detection due to their non-invasive nature and biological relevance. Key biomarkers, including creatinine, LYVE1, REG1B, and TFF1, show abnormal levels in the presence of pancreatic cancer. This study focuses on evaluating these biomarkers using machine learning classifiers to enhance diagnostic accuracy. The research objectives are twofold: (1) to assess the performance of various machine learning models for classifying pancreatic cancer using urinary biomarkers and (2) to address the limitations of existing methods by leveraging hyperparameter tuning for improved model accuracy and reliability.

Literature Review

A research paper published by Plos Medicine features the dataset used for this analysis. The title of this research paper is, "A Combination of Urinary Biomarker Panel and PancRISK Score for Earlier Detection of Pancreatic Cancer: A Case-control Study". This paper explains the practicality of using urinary biomarkers to detect early-stage pancreatic cancer. The study developed the PancRISK model, achieving a ROC rate of over 90%. Another related study, "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks, " explores the performances of a proposed 1-D CNN + LSTM model. The proposed CNN model achieved a 97% accuracy score and an AUC curve score of 98%. We will test this model out, bringing into question the specific parameters used for this model.

Methods

1. Dataset Acquisition

The dataset, sourced from the Spanish National Cancer Research Center, consists of 590 samples, with features representing urinary biomarker levels and binary labels indicating cancer presence. The data was preprocessed using standardization techniques to enhance model compatibility.

2. Setting up Google Colab

Pair plots and box plots were generated to analyze data trends. Inconsistent class distributions in pair plots were corrected by adjusting outlier weights during model training. Figures were color-coded consistently for clarity.

3. Loading the Dataset into Google Colab and Creating a Data Visualization File

Five classifiers were evaluated: XGBoost, LightGBM, Random Forest, Support Vector Machine (SVM), and 1D CNN-LSTM. Data was split into 70% training, 15% validation, and 15% testing sets. Hyperparameter tuning was performed using grid search to optimize model performance.

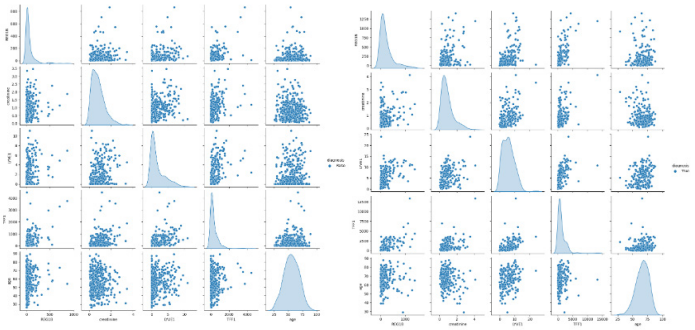


Fig. 2. The pairplot on the left showcases the abundance of biomarkers that people with early-stage pancreatic cancer possess. The pair plot on the right contrasts this with the abundance of biomarkers from people with early pancreatic cancer. Telling a difference between the values of the two pair plots is difficult for a model due to a lack of outliers, so adding separability into the dataset will allow the model to distinguish between the two groups, making its results more accurate.

3. Model Training File

Split the dataset into training and testing sets to ensure the models have sufficient data for both training and validation. Classification models, including 1D CNN-LSTM, gradient boosting, and regression models, should be tested on the dataset. The performance of these models will be improved through hyperparameter tuning, a process in which specific parameters dictating model performance are adjusted to align with the dataset's trends and correlations.

4. Model Evaluation

Once the models are trained, their performance must be evaluated using key metrics such as accuracy, precision, F1 score, and recall. These evaluations will ensure a model's accuracy, how accurate its accuracy is, and how consistent it is. ROC curves should also be generated to compare the true positive rate against the false positive rate, providing more insight toward the accuracy of a model.

Results

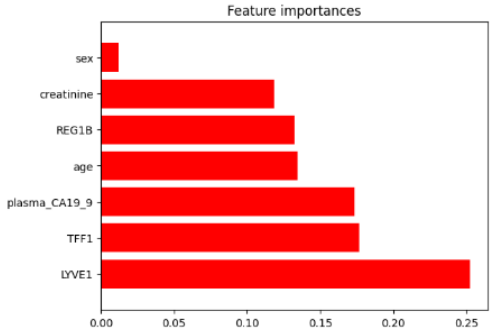
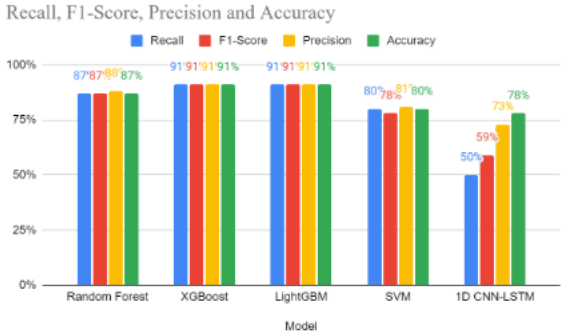


Fig. 3. This graph shows that the urinary biomarker, LYVE1, is most crucial for detecting early-stage pancreatic cancer.



Model	Recall	F1-Score	Precision	Accuracy
Random Forest	87%	87%	88%	87%
XGBoost	91%	91%	91%	91%
LightGBM	91%	91%	91%	91%
SVM	80%	78%	81%	80%
1D CNN-LSTM	50%	59%	73%	78%

Fig. 4. The data show that the model with the highest performing accuracy and reliability is LightGBM. LightGBM had a 91% accuracy rate, equivalent to that of XGBoost, but LightGBM scored a higher precision, recall, and F1-Score by a couple of hundredths of a decimal.

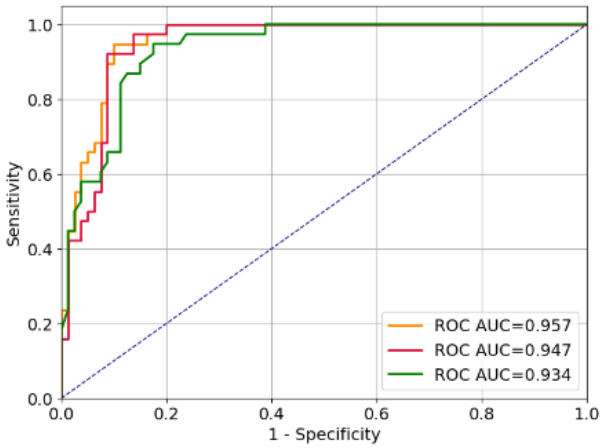


Fig. 5. LightGBM is the orange line, XGBoost is the red line, and a random forest classifier model is the green line. LightGBM scored the highest ROC AUC curve, meaning it has the most reliable accuracy.

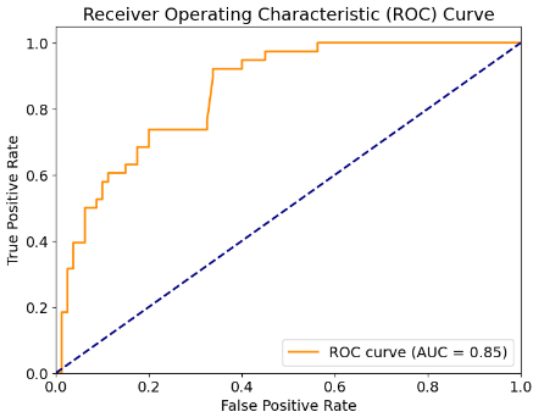


Fig. 6. This ROC AUC curve is for the 1D CNN-LSTM model, which achieved a 70% accuracy rate and an 85% ROC AUC score.

Discussion

The findings of this study highlight the efficacy of urinary biomarkers, combined with hyperparameter-tuned machine learning models, for early pancreatic cancer detection. LightGBM and XGBoost models achieved the highest accuracy (91%), demonstrating their reliability for this classification task. These results confirm the suitability of gradient-boosting methods for structured medical datasets due to their ability to capture complex feature interactions. However, while the Random Forest model also performed well, its capacity to handle both classification and regression tasks does not directly imply superiority for this specific problem. Its slightly lower accuracy (87%) underscores its limitations compared to gradient-boosting approaches. Contrary to prior studies, the 1D CNN-LSTM model underperformed with a 78% accuracy rate, raising questions about the dataset size and hyperparameter configurations used. Standardizing these parameters in future analyses will be essential to validate its potential. The 1D CNN-LSTM's deviation also highlights the importance of reproducibility in machine learning research, as models are highly sensitive to data preprocessing and tuning. The slight edge of LightGBM over XGBoost in this study should be interpreted cautiously. If the

performance gap refers to computational efficiency, processing time must be treated as a separate metric, independent of accuracy. Further experiments could quantify this trade-off to clarify model selection criteria for clinical implementation. The feature importance analysis underscored the relevance of LYVE1, REG1B, and TFF1, particularly LYVE1, as key biomarkers. Their biological significance supports their prioritization in diagnostic applications. However, dataset limitations, including a relatively small sample size and class imbalance, may have influenced the models' predictive power. Balancing techniques and larger datasets could enhance model validity and generalizability. This study emphasizes the promise of integrating machine learning with urinary biomarker analysis for non-invasive diagnostics. Future work should focus on expanding datasets, standardizing model parameters, and developing practical diagnostic tools to transition from research to clinical practice.

Conclusion

This study demonstrates the potential of urinary biomarkers, analyzed through hyperparameter-tuned machine learning models, for early pancreatic cancer detection. LightGBM and XGBoost emerged as the most reliable models, achieving 91% accuracy, with LYVE1 identified as the most critical biomarker. These findings reinforce the value of integrating advanced computational techniques with biomarker analysis to improve diagnostic precision. However, this work is not without limitations. The dataset size was relatively small, which may affect the generalizability of the results. Additionally, class imbalance in the data could have influenced the evaluation metrics. Expanding the dataset, incorporating more diverse features, and optimizing model training and prediction times are essential for enhancing model performance and practical applicability. Future research should explore combining urinary biomarker analysis with imaging-based methods, such as X-ray or MRI, to further improve detection accuracy. The development of hardware capable of real-time biomarker-based diagnostics would also bridge the gap between laboratory research and clinical implementation. By addressing these challenges, this work lays the foundation for non-invasive, scalable, and accurate diagnostic solutions for pancreatic cancer, ultimately aiming to reduce mortality rates and improve patient outcomes.

References

Capilitan, S. (2022, September 8). Employers: What are the turnaround times for drug testing? *Chane Solutions*, www.chanesolutions.com/blog/2021/08/14/turnaround-times-drug-testing.

Debernardi, S., Blyuss, O., Rycyk, D., Srivastava, K., Jeon, C. Y., Cai, H., Cai, Q., Shu, X., & Crnogorok-Jurcevic, T. (2022, September 15). Urine biomarkers enable pancreatic cancer detection up to 2 years before diagnosis. *International Journal of Cancer*, 152(4), 769-80. https://doi.org/10.1002/ijc.34287.

Debernardi, S., O'Brien, H., Algahmdi, A. S., Malats, N., Stewart, G. D., Pljesa-Ercegovac, M., Costello, E., Greenhalf, W., Saad, A., Roberts, R., Ney, A., Pereira, S. P., Kocher, H. M., Duffy, S., Blyuss, O., & Crnogorak-Jurcevic, T. (2020, December 10). A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study. *PLOS Medicine*, 17(12). https://doi.org/10.1371/journal.pmed.1003489.

"Facts About Pancreatic Cancer." American Cancer Society, www.cancer.org/cancer/types/pancreatic-cancer/about/key-statistics.html.

Fernandez-del Castillo, C., Jimenez, R. E., & Murphy, J. E. (2023, December 1). Supportive care of the patient with locally advanced or metastatic exocrine pancreatic cancer. *UpToDate*. Retrieved August 26, 2024, from https://www.uptodate.com/contents/supportive-care-of-the-patient-with-locally-advanced-or-met astatic-exocrine-pancreatic-cancer

Hoffman, M. Pancreatic cancer diagnosis and early detection. (2024, February 14). *WebMD*, www.webmd.com/cancer/pancreatic-cancer/pancreatic-cancer-diagnosis.

Karar, M. E., El-Fishawy, L., & Radad, M. (2023, April) Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks. *Journal of Biological Engineering*, 17(1), https://doi.org/10.1186/s13036-023-00340-0.

LightGBM (Light Gradient Boosting Machine). (2024, July 4). *GeeksforGeeks*,, www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine.

Crnogorak-Jurcevic, T. New tests for rarly detection of pancreatic cancer offer significant hope. *Queen Mary University of London Research*. www.qmul.ac.uk/research/featured-research/new-tests-for-early-detection-of-pancreatic-cancer-offer-significant-hope.

O'Neill, R. S., & Stoita, A. (2021, July 14) Biomarkers in the diagnosis of pancreatic cancer: Are we closer to finding the golden ticket? *World Journal of Gastroenterology*, 27(26), 4045-87. https://doi.org/10.3748/wjg.v27.i26.4045.

Radon, T. P., Massat, N. J., Jones, R. S., Alrawashdeh, W., Dumartin, L., Ennis, D., Duffy, S. W., Kocher, H. M., Pereira, S. P., Guarner, L., Murta-Nascimento, C., Real, F. X., Malats, N., Neoptolemos, J. P., Costello, E., Greenhalf, W., Lemoine, N. R., & Crnogorac-Jurčević, T. (2015, April). Identification of a three-biomarker panel in urine for early detection of pancreatic adenocarcinoma. *Clinical Cancer Research,* 21(15), 3512–3521. https://doi.org/10.1158/1078-0432.ccr-14-2467

Random forest. (2023, November 22) *Corporate Finance Institute*. https://corporatefinanceinstitute.com/resources/data-science/random-forest

Tan, D. J., Crnogorac-Jurčević, T., Massat, N. J., Jones, R. S., Alrawashdeh, W., Dumartin, L., Ennis, D., Duffy, S. W., Kocher, H. M., Pereira, S. P., Guarner, L., Murta-Nascimento, C., Real, F. X., Malats, N., Neoptolemos, J. P., Costello, E., Greenhalf, W., & Lemoine, N. R. (2020, December 11). Intron retention is a robust marker of intertumoral heterogeneity in pancreatic ductal adenocarcinoma. *NPJ Genomic Medicine,* 5: 55. https://doi.org/10.1038/s41525-020-00159-4

What is XGBoost? *NVIDIA Data Science Glossary*, www.nvidia.com/en-us/glossary/data-science/xgboost.

Zhou, B., Xu, J., Cheng, Y., Gao, J., Hu, S., Wang, L., & Zhan, H. (2017, July 15). Early detection of pancreatic cancer: Where are we now and where are we going? *International Journal of Cancer*, 141(2), 231–241.https://doi.org/10.1002/ijc.30670