

An Ensemble Learning Approach for Pancreatic Tumor Detection

Sripathi Chaitanya Bharathi¹ and Eswaraiah Rayachoti²

¹⁻²SCOPE, VIT-AP University, Andhra Pradesh, India

Email: bharathi.23phd7135@vitap.ac.in and eswaraiah.rayachoti@vitap.ac.in

Abstract— Pancreatic cancer remains one of the most lethal cancers worldwide, primarily due to its asymptomatic nature in early stages and limited availability of reliable diagnostic tools. This research addresses this critical gap by proposing a machine learning-based diagnostic system that leverages non-invasive urinary biomarker data to predict the risk of pancreatic tumors. The study utilizes biomarkers such as LYVE1, REG1B, TFF1, and CA19-9, which have been clinically correlated with tumor development. To ensure robust and accurate predictions, we implemented an ensemble learning approach combining three powerful models: Random Forest, XGBOOST, and CATBOOST. The models were trained and evaluated using an 80-20 split on a cleaned version of the Debernardi et al., 2020 dataset. The CATBOOST classifier achieved the highest individual accuracy (~94.07%), while the ensemble model yielded a balanced F1-score of ~91.5%, demonstrating its effectiveness in reducing false negatives—a critical requirement for medical screening systems. The complete solution was deployed as a web-based application using Flask, with user input forms connected to the trained model. The deployment was made publicly accessible via the Render cloud platform, making the tool easily usable by researchers or clinicians. This report presents the design, implementation, results, and deployment strategy of the system, and discusses its potential use in early pancreatic cancer detection scenarios. Future work includes extending the dataset, integrating real-time clinical inputs, and enhancing model explainability for medical adoption.

Index Terms— Machine Learning (ML), Deep Learning (DL), Extreme Gradient Boosting (XGBOOST), Categorical Boosting (CATBOOST)..

I. INTRODUCTION

Pancreatic tumour, which can be benign or cancerous, need to be detected early because they often become life-threatening. However, early detection is challenging as the pancreas is deep inside the body and tumours usually don't cause symptoms in early stages. This leads to delayed diagnosis and limited treatment options. Traditional tests like imaging and biopsies are effective but are expensive or invasive, making them unsuitable for routine screening. Recent studies have shown that certain urinary biomarkers—such as LYVE1, REG1B, TFF1, and CA19-9—can help detect pancreatic tumours. By using machine learning to analyse these biomarkers, we can build non-invasive systems to predict tumour risk. This research aims to create a smart, cost-effective, and accessible system for early tumour detection using machine learning. It uses a data pipeline to clean and normalize biomarker inputs and applies machine learning models—Random Forest, XGBoost, and CatBoost—to classify tumour risk. An ensemble model is used to combine these classifiers and improve prediction accuracy.

A simple web interface built with Flask allows users to enter biomarker data and get real-time risk predictions. The entire system is hosted on Render, a cloud platform, making it easy to access.

Ensemble learning is especially useful in healthcare because it combines the strengths of different models, resulting in better performance. In this research, it serves as a decision-support tool to help identify high-risk cases without needing invasive tests. This can help catch pancreatic tumours earlier and avoid unnecessary procedures. The research is motivated by the high death rate and silent nature of pancreatic tumour. With modern machine learning and access to biomarker data, this kind of system is both relevant and important for healthcare. Despite the potential of urinary biomarkers, there are very few tools available that allow easy and automated analysis. This research fills that gap by building a user-friendly platform that predicts pancreatic tumour risk from biomarker inputs. Its main goals are to clean and prepare the data, test different models, build an ensemble system, and deploy it with a simple web interface.

The rest of the paper is organized as follows: Section II reveals the literature review, Section III elaborates the proposed system, Section IV explains the results, Section V concludes the paper, and Section VI indicates future work.

II. LITERATURE REVIEW

Machine learning is helping doctors find diseases like pancreatic tumors at an earlier stage. These tumors are hard to detect early because they don't cause symptoms and are located deep inside the body. Most cases are found too late, making treatment harder. Tests like CT scans and MRIs can help but are too expensive or invasive for regular use. New research shows that certain substances in urine, called biomarkers, can indicate the presence of tumors, and using machine learning with this data can create easy, low-cost ways to find the disease early.

Recent advancements in pancreatic cancer detection have increasingly leveraged deep learning, imaging, and biomarker integration to improve early diagnosis. Chen et al. [1] demonstrated the use of deep learning on CT scans for nationwide pancreatic cancer detection, emphasizing the scalability and robustness of AI models in clinical imaging contexts. Similarly, Dinesh et al. [2] highlighted the diagnostic potential of deep learning in detecting pancreatic tumors, showing promising accuracy that supports AI's role in radiological workflows. Daher et al. [3] and Wu et al. [4] reviewed the integration of biomarkers with machine learning and imaging, illustrating the synergistic impact of combining molecular data with AI to enhance early detection. O'Neill et al. [5] discussed progress and limitations in biomarker-based diagnosis, suggesting we are approaching a viable clinical application, yet challenges remain in specificity and sensitivity.

Karar et al. [6] introduced a novel approach using 1-D convolutional neural networks to classify urinary biomarkers, offering a non-invasive and automated alternative to traditional diagnostic methods. Yang et al. [7] provided a comprehensive review of early screening strategies, including imaging, biomarkers, and AI models, advocating for multi-modal solutions. Jin et al. [8] identified a new class of tsRNA signatures with potential for both diagnostic and prognostic applications, contributing to the expanding field of RNA-based biomarkers. Das et al. [9] offered a broader perspective on the role of biomarkers in cancer detection, underlining their cross-cancer applicability while emphasizing pancreatic cancer as a critical case. Şolea et al. [10] reviewed cutting-edge imaging technologies, including radiomics and AI-enhanced techniques, that are revolutionizing pancreatic tumour diagnosis.

III. PROPOSED METHODOLOGY

This research uses supervised machine learning to predict the risk of pancreatic tumours based on urinary biomarker data. The process starts with understanding and cleaning the data, followed by training three different models — Random Forest, XGBoost, and CatBoost — which are known for handling complex patterns in data. These models are then combined into a soft voting ensemble to make the final prediction more accurate and reliable, especially to reduce false negatives, which is crucial in medical diagnosis. After testing and evaluating the models using performance metrics like accuracy and recall, the final system is deployed as a user-friendly web application. This app allows healthcare professionals and researchers to enter patient biomarker data and receive real-time risk predictions, making the solution both effective and accessible.

The system architecture for pancreatic tumor detection as shown in figure 1 starts with urinary biomarker data (LYVE1, REG1B, TFF1, CA19-9), which undergoes data cleaning and preprocessing to ensure quality and consistency. Important features are then extracted and used to train three machine learning models—Random Forest, XGBoost, and CatBoost—each known for their predictive strength. These models are combined into a

soft voting ensemble classifier, which improves accuracy and reduces false results. A Flask web application allows users to input data through a simple web form and get real-time risk predictions. The system is then deployed on the Render cloud platform, making it easily accessible online.

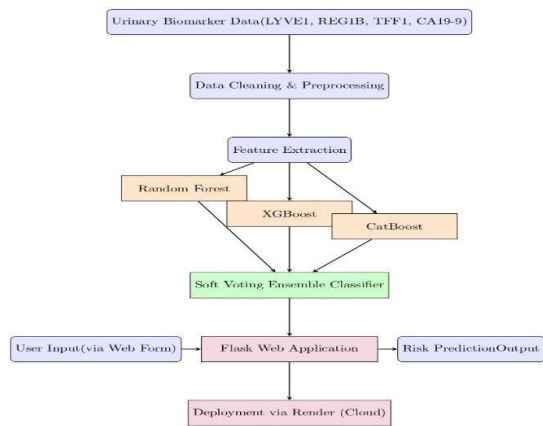


Figure 1: Architecture Diagram: Pancreatic Tumour Detection System

Step 1: Dataset Description

The dataset includes anonymised patient samples with demographic, clinical, and biomarker data related to pancreatic tumours. Key features include urinary and plasma biomarkers like CA19-9, LYVE1, REG1B, and TFF1, with diagnosis as the target variable.

Step 2: Pre-processing

The dataset was cleaned by imputing missing values, removing incomplete rows, and selecting key features like biomarkers and age. Label encoding was applied to the target, and Min-Max scaling was used to normalize feature values. The data was split into 80% training and 20% testing, with correlation analysis and visual tools confirming feature relevance.

Step 3: Data Visualization

To explore the relationships between biomarkers and tumor diagnosis, various visualization techniques were used. A Seaborn pairplot showed how features like REG1B, TFF1, and LYVE1 vary with diagnosis, with tumour-positive and tumour-negative cases forming noticeable clusters.

The correlation heat map shown in figure 1 revealed that LYVE1, TFF1, age, and CA19-9 had moderate positive correlations with tumour presence, while no strong multicollinearity was found among features. A Random Forest feature importance plot showed LYVE1 as the most influential feature, followed by TFF1 and CA19-9, confirming their value in predicting tumours. These visualizations validated the selected biomarkers and helped guide further model development.

Step 4: Model Development

The choice of machine learning models was guided by their proven effectiveness in handling structured medical data. Three models were selected: Random Forest, XGBoost, and CatBoost. Random Forest is easy to use, handles over fitting well, and helps identify which features are most important. XGBoost is powerful for finding complex patterns in data and includes techniques to prevent over fitting. CatBoost excels with categorical and small datasets, and combining diverse models captures varied patterns, boosting overall prediction accuracy and reliability.

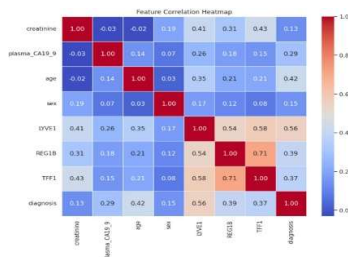


Figure 2: Feature Correlation Heatmap

Step 5: Training Pipeline

The training process started by splitting the data into 80% for training and 20% for testing, keeping a balanced mix of tumour and non-tumour cases. Missing values were filled using the average, especially for features like CA19-9. Models like XGBoost were scaled and optimized using GridSearchCV with cross-validation to ensure good generalization and prevent overfitting.

Step 6: Random Forest Classifier

The Random Forest model, using 100 estimators with unrestricted depth, served as the initial baseline classifier. It achieved a cross-validation accuracy of 86.4% and a test accuracy of 87.3%. Feature importance analysis highlighted LYVE1 as the most influential predictor, followed by TFF1 and plasma_CA19_9, while sex showed low predictive value, aligning with earlier correlation findings.

Step 7: XGBoost Classifier

XGBoost, the second model implemented, outperformed the Random Forest classifier after hyperparameter tuning, achieving a cross-validation balanced accuracy of around 91%. It showed strong precision and recall on the test set, particularly in detecting tumor-positive cases. Its ability to capture complex interactions between biomarkers and resist overfitting made it a robust and valuable addition to the final ensemble.

Step 8: CATboost Classifier

High Performance: CatBoost achieved a test accuracy of 94.07% and a macro-average F1-score of 0.93, outperforming Random Forest and XGBoost.

Ease of Use: It handled missing values internally and required minimal preprocessing, streamlining the training process.

Robust and Interpretable: CatBoost effectively captured complex feature interactions and remained stable on imbalanced data, making it ideal for biomedical applications.

Step 9: Ensemble Classifier

A soft voting ensemble combining Random Forest, XGBoost, and CatBoost achieved 94.07% accuracy, matching CatBoost but offering greater robustness and confidence near decision boundaries.

It minimized false positives/negatives, which is crucial for healthcare applications.

The finalized model was deployed in a web-based interface for real-time, reliable predictions.

IV. RESULTS AND DISCUSSION

The Web User Interface (Web UI) was developed to make the pancreatic tumor risk prediction system accessible to users without technical expertise. Designed with simplicity, responsiveness, and real-time interaction in mind, the interface allows users to input biomarker values and receive immediate feedback on their risk status.

A. Biomarker Reference Ranges and Interpretation

To facilitate accurate input and interpretation within the web interface, clinically accepted reference ranges as depicted in table 1 were used for each biomarker. These ranges help distinguish between low and high tumor risk profiles and were derived from prior clinical literature, particularly the study by Mohamed Esmail et al. [6]. The biomarkers used in this research include LYVE1, REG1B, and TFF1, which are normalized using creatinine levels in urine samples. Additional contextual variables such as CA19-9, age, and sex are also included for improved risk stratification.

TABLE I: BIOMARKER REFERENCE RANGES

Biomarker	Units	Low Risk Range	High Risk Range	Clinical Role / Notes
LYVE1	ng/mg creatinine	0 – 60	> 60	Elevated in urine of PDAC (pancreatic cancer) patients
REG1B	ng/mg creatinine	0 – 20	> 20	Urinary biomarker; often higher in cancer cases
TFF1	ng/mg creatinine	0 – 30	> 30	Elevated in both PDAC and inflammatory conditions
Creatinine	mg/dL (urine)	0.5 – 2.0 (typical)	Outliers may affect normalization	Used for adjusting biomarker concentrations in urine
CA19-9	U/mL	0 – 37	> 37	Most common serum marker used for pancreatic tumor monitoring
Age	years	< 50	> 60	Risk of pancreatic tumor increases with age
Sex	—	—	—	Not directly linked to biomarker values

These ranges are displayed on the UI and used internally for normalization and comparison. While the model uses continuous values for prediction, these reference bands aid users in understanding how their inputs relate to clinical thresholds.

The dataset depicted in table 2 provides clinical and biomarker information for five individuals, aiming to assess their risk levels for pancreatic cancer. Each person is described by their age, sex (0 for female, 1 for male), and levels of several key biomarkers: Creatinine, CA19-9, LYVE1, REG1B, and TFF1. These biomarkers are known to be associated with pancreatic function and tumor activity. Based on these values, each individual is assigned a prediction indicating their health status: "Low Risk and Healthy," "Moderate," or "High Risk and Tumour."

Three individuals—aged 35, 29, and 18—are classified as "Low Risk and Healthy." These cases generally show balanced or lower levels of the biomarkers, especially LYVE1 and REG1B, suggesting no immediate signs of pancreatic abnormality. One individual, a 56-year-old male, is categorized as "High Risk and Tumor." His biomarker levels are significantly elevated, particularly LYVE1 (75), REG1B (28), and TFF1 (51), which strongly indicate a possible tumor presence. The remaining individual, a 70-year-old female, falls into the "Moderate" risk category, with moderately high levels of LYVE1 (62) and REG1B (22), which may suggest early signs of disease or the need for further monitoring. Overall, the data illustrates how variations in biomarker levels, especially LYVE1, REG1B, and TFF1, correlate with increasing levels of cancer risk, and how age may also influence the prediction outcome.

TABLE II: CLINICAL AND BIOMARKER INFORMATION

Age	Sex	Creatinine	CA19-9	LYVE1	REG1B	TFF1	Prediction
35	0	1.5	35	38	13	25	Low Risk and Healthy
56	1	2.5	39	75	28	51	High Risk and Tumor
29	1	1.0	30	28	9	25	Low Risk and Healthy
70	0	2.3	38	62	22	31	Moderate
18	0	0.2	30	43	18	28	Low Risk and Healthy

B. Evaluation Metrics

The performance of the classification models was evaluated using several key metrics: accuracy, precision, recall, F1-score, and the confusion matrix as shown in figure 3. While accuracy provides a general sense of correctness, greater emphasis was placed on recall and F1-score due to their importance in medical diagnostics. Specifically, recall is crucial for minimizing false negatives, which can be dangerous if a tumor is missed. Therefore, the evaluation prioritized the model's ability to correctly detect tumor-positive cases rather than just overall accuracy.

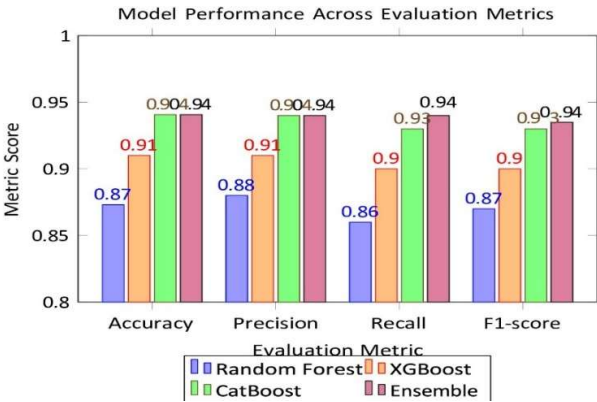


Figure 3: Model Performance Across Evaluation Metrics

C. Performance Summary

The performance comparison of classification models as shown in figure 4 and figure 5 revealed that the Random Forest classifier achieved a solid baseline accuracy of approximately 87.3%, with good precision and

recall for tumor detection. It also highlighted LYVE1 and TFF1 as key predictive biomarkers. XGBoost improved upon this with an accuracy of around 91%, showing better handling of complex interactions between features and reducing false negatives, which is crucial in medical diagnosis. CatBoost outperformed both models with an accuracy of 94.07% and a macro-average F1-score of 0.93, maintaining strong recall and precision for both tumor-positive and tumor-negative classes. It also required minimal data preprocessing, making it practical for real-world use. The final ensemble model, combining Random Forest, XGBoost, and CatBoost through soft voting, matched CatBoost’s accuracy at 94.07% but delivered greater robustness and stability.

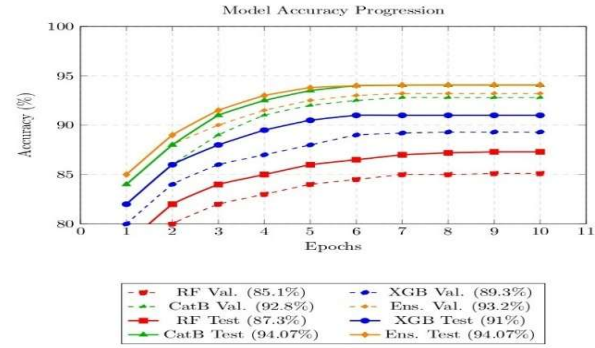


Figure 4: Model Accuracy Progression

It reduced output variability across different folds and provided more consistent, reliable predictions, making it highly suitable for clinical applications.

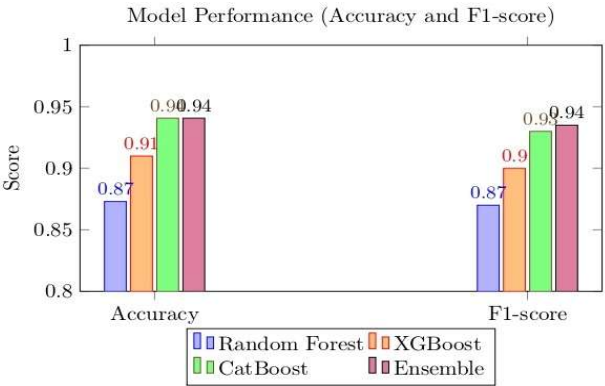
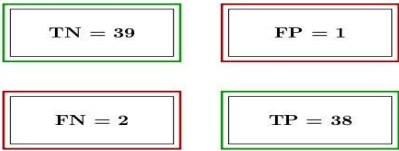


Figure 5: Performance Comparison of Classification Models

D. Confusion Matrix Analysis

The confusion matrix for each model was analyzed to understand the distribution of predictions. As shown in figure 6, the ensemble model showed very few false positives and false negatives, making it suitable for high-stakes environments like medical screening. It consistently classified tumor-positive cases correctly, which is essential for avoiding missed diagnoses.



TP (True Positive): 38 — Tumor cases correctly predicted.
TN (True Negative): 39 — Healthy cases correctly predicted.
FP (False Positive): 1 — Healthy cases wrongly predicted as tumor.
FN (False Negative): 2 — Tumor cases wrongly predicted as healthy.

Figure 6: Confusion Matrix Ensemble Classifier (Accuracy: 94.07%)

E. Feature Importance and Interpretability

Feature importance plots from Random Forest and XGBoost highlighted LYVE1, TFF1, and REG1B as the most significant contributors to prediction. These findings are consistent with clinical literature, supporting the relevance of these urinary biomarkers in tumor risk stratification.

Additionally, a correlation heatmap confirmed that these features had the strongest linear relationships with the diagnosis label, further validating their inclusion in the model. The low multicollinearity among features ensured that the model was not biased by redundant predictors. Figure 7 and 8 shows the web page that will be displayed when the application is run. The web page allows the user to enter details of patient. It also displays the prediction result when the user enters details and clicks on 'Predict' button.

F. Comparative Discussion

Overall, the results demonstrate that ensemble learning offers a tangible advantage over single-model strategies in biomarker-based tumor prediction. While CatBoost alone performed exceptionally well, combining it with Random Forest and XGBoost improved the model's generalization and reliability across samples.

The insights from this analysis validate the use of urinary biomarkers in non-invasive screening and showcase the potential of machine learning in transforming early detection workflows in healthcare.

A user-friendly interface was developed to predict the risk of pancreatic cancer based on patient information such as age, sex, creatinine level, and biomarkers like CA19-9, LYVE1, REG1B, and TFF1. The system analyzes the input data and provides a prediction, such as "High Risk and Tumor" or "Low Risk and Healthy." For example, a 56-year-old male with higher biomarker levels was predicted as high risk, while a 35-year-old female with lower values was predicted as low risk. The model showed a high accuracy of 93.22%, making it useful for early cancer detection and clinical decision-making.



Figure 7: User Experience (When tumor is predicted)

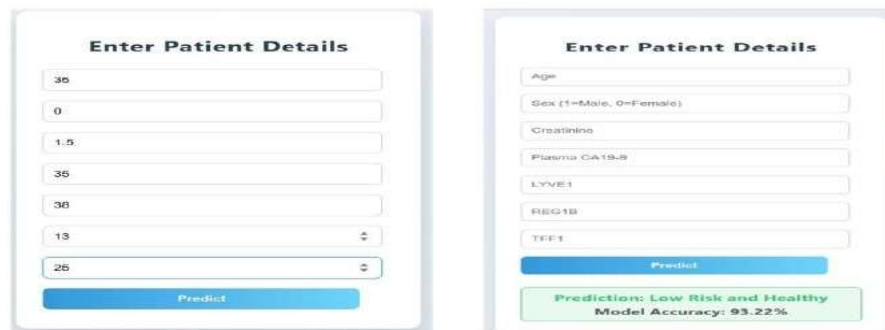


Figure 8: User Experience (Tumor Is Not Predicted)

V. CONCLUSION

This research shows how machine learning can help detect pancreatic tumors early using non-invasive urine test data. By combining several biomarkers like LYVE1, REG1B, TFF1, and CA19-9, and training different machine learning models, a final ensemble model was created that gives accurate results. The best model used soft voting and achieved a high accuracy of 94.07%, performing well for both positive and negative cases. To make the

system easy to use, a web interface was built using Flask, so users can get predictions in real-time without needing technical skills. The tool was deployed on Render, making it accessible from any device, and providing a simple and effective way to support early cancer screening.

FUTURE SCOPE

A. Explainable AI Integration

Incorporating interpretability tools such as SHAP or LIME can help clinicians understand how predictions are made, increasing trust in the system and encouraging adoption.

B. Graded Risk Output

The binary output system can be enhanced to provide a graded risk score (low, medium, high), along with prediction confidence. This would offer more nuanced information to healthcare professionals and support more informed decision-making.

C. Clinical Integration

The system could be connected to Electronic Health Records (EHR) and used alongside imaging or blood test data to offer a comprehensive diagnostic view. This would position the model as a supportive diagnostic aid rather than a standalone tool.

REFERENCES

- [1] Po-Ting Chen, Tinghui Wu, Pochuan Wang, Dawei Chang, Kao-Lang Liu, Ming-Shiang Wu, Holger R. Roth, Po-Chang Lee, Wei-Chih Liao, and Weichung Wang, "Pancreatic cancer detection on CT scans with deep learning: a nationwide population-based study," *Radiology*, Vol. 306(1), pp. 172-182, 2023.
- [2] M. G. Dinesh, Nebojsa Bacanin, S. S. Askar, and Mohamed Abouhawwash, "Diagnostic ability of deep learning in detection of pancreatic tumour," *Scientific Reports*, Vol. 13(1), 9725, 2023.
- [3] Hisham Daher, Sneha A Punchayil, Amro Ahmed Elbeltagi Ismail, Reuben Ryan Fernandes, Joel Jacob, Mohab H Algazzar, and Mohammad Mansour, "Advancements in pancreatic cancer detection: integrating biomarkers, imaging technologies, and machine learning for early diagnosis," *Cureus*, Vol. 16(3), 2024.
- [4] Haotian Wu, Suwen Ou, Hongli Zhang, Rui Huang, Shan Yu, Ming Zhao, and Sheng Tai, "Advances in biomarkers and techniques for pancreatic cancer diagnosis," *Cancer cell international*, Vol. 22(1), 220, 2022.
- [5] Robert S O'Neill, and Alina Stoita, "Biomarkers in the diagnosis of pancreatic cancer: Are we closer to finding the golden ticket?" *World journal of gastroenterology*, Vol. 27(26), 4045, 2021.
- [6] Mohamed Esmail Karar, Nawal El-Fishawy, and Marwa Radad, "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks," *Journal of Biological Engineering*, Vol. 17(1), 28, 2023.
- [7] Jinshou Yang, Ruiyuan Xu, Chengcheng Wang, Jiangdong Qiu, Bo Ren, and Lei You, "Early screening and diagnosis strategies of pancreatic cancer: a comprehensive review," *Cancer Communications*, Vol. 41(12), pp.1257-1274, 2021.
- [8] Fangfang Jin, Liuqing Yang, Weixiang Wang, Na Yuan, Shoubin Zhan, Ping Yang, Xi Chen, Tonghui Ma, and Yanbo Wang, "A novel class of tsRNA signatures as biomarkers for diagnosis and prognosis of pancreatic cancer," *Molecular cancer*, Vol. 20(1), 95, 2021.
- [9] Sreyashi Das, Mohan Kumar Dey, Ram Deviredy, and Manas Ranjan Gartia, "Biomarkers in cancer detection, diagnosis, and prognosis," *Sensors*, Vol. 24(1), 37, 2023.
- [10] Sabina Florina Şolea, Mihaela Cristina Brisc, Alexandra Orăşeanu, Florian Ciprian Venter, Ciprian Mihai Brisc, Răzvan Mihai Şolea, Lavinia Davidescu, Amina Venter, and Ciprian Brisc, "Revolutionizing the pancreatic tumor diagnosis: emerging trends in imaging technologies: a systematic review," *Medicina*, Vol. 60(5), 695, 2024.