Leveraging lncRNA Expression Profiles for Machine Learning-Based Prediction of Lymph Node Involvement in PancreaticAdenocarcinoma (PAAD)

A.Mawada MohammadAli FadlAllah	Researcher , bioinformatics
A. Abdelrahman Hamza Abdelmoneim	Researcher , bioinformatics
A. Moaaz Mohammed Saadaldin	Researcher , bioinformatics
Dr. Mohamed mamoun abdulelrahim	alzaiem alazhri university- information technology department

Abstract:

Pancreatic adenocarcinoma (PAAD) remains one of the most lethal malignancies, with the presence of lymph node metastasis (LNM) serving as a critical prognostic factor that heavily influences treatment decisions; however, current preoperative diagnostic methods often lack the precision needed for reliable risk stratification. This study evaluated the effectiveness of machine learning models trained on long non-coding RNA (lncRNA) expression profiles to predict lymph node metastasis in PAAD patients by analyzing RNA-seq and clinical data from the TCGA-PAAD cohort to identify lncRNAs associated with lymph node status, addressing class imbalance using the Synthetic Minority Oversampling Technique (SMOTE), and training Logistic Regression, Random Forest, and XGBoost models to classify patients as node-positive (N1) or node-negative (N0), with performance assessed via five-fold cross-validation and hold-out testing using ROC-AUC and accuracy metrics. Logistic Regression achieving an ROC-AUC of 1.000, Random Forest 0.998, and XGBoost 0.946, while four lncRNAs—AC096920.1, AC116003.3, CNTFR-AS1, and AC093734.1—consistently emerged as the top predictors and showed significant differential expression between N0 and N1 groups ($p < 0.05$). These findings indicate that machine learning models based on lncRNA expression can accurately predict nodal metastasis in PAAD, and the identified four-lncRNA panel holds strong potential as a biomarker signature for preoperative risk stratification, although external validation and functional studies will be essential for future clinical translation.

Keywords:PAAD, Machine Learning, lncRNAs, lymph node metastasis (LNM), TCGA

استخدام ملفات تعبير الـ lncRNA في نماذج التعلم الآلي للتنبؤ بغزو العقد اللمفاوية في السرطان الغدي للبنكرياس(PAAD)

أ.مودة محمد علي فضل الله - طالبة ماجستير – الجامعة الوطنية
أ. عبد الرحمن حمزة عبد المنعم – طالب ماجستير – الجامعة الوطنية
أ. معاذ محمد سعد الدين – طالب ماجستير- الجامعة الوطنية
د. محمد المامون عبد الرحيم خضر – قسم تقنية المعلومات – جامعة الزعيم الأزهري

المستخلص:

يُعد سرطان البنكرياس الغدي (PAAD) من الأورام الخبيثة الأكثر فتكاً، حيث يُمثل وجود النقائل اللمفاوية (LNM) مؤشراً تنبؤياً حاسماً يؤثر بشكل كبير على قرارات العلاج، إلا أن الطرق التشخيصية الحالية غالباً ما تفتقر إلى الدقة الكافية للتصنيف الخطر قبل الجراحة. هدفت هذه الدراسة إلى تقييم فعالية نماذج التعلم الآلي المدربة على ملفات تعبير الـ lncRNA في التنبؤ بوجود النقائل اللمفاوية لدى مرضىPAAD ، وذلك من خلال تحليل بيانات RNA-seq والمعلومات السريرية من مجموعة TCGA-PAAD لتحديد الـ lncRNA المرتبطة بحالة العقد اللمفاوية، مع معالجة عدم التوازن في الفئات باستخدام تقنيةSMOTE ، ثم تدريب نماذجLogistic Regression و Random Forestو XGBoost لتصنيف المرضى إلى مجموعتي N0 وN1، وتقييم الأداء باستخدام التحقق المتقاطع بخمسة أضعاف والاختبار المستقل مع قياسات ROC-AUC والدقة. ، حيث حقق Logistic Regression قيمة ROC-AUC بلغت 1.000، و0.998 Random Forest، وXGBoost 0.946، مع ظهور أربعةlncRNA (AC096920.1 ، AC116003.3، CNTFR-AS1، AC093734.1) باستمرار كأهم المتنبئات وعبر عنها بشكل مختلف بشكل معنوي بين المجموعتين.($p < 0.05$) تُظهر هذه الدراسة أن نماذج التعلم الآلي المعتمدة على تعبير الـ lncRNA قادرة على التنبؤ بدقة عالية بالنقائل اللمفاوية فيPAAD ، وأن اللوحة المكونة من هذه الأربعة lncRNA تُعد مرشحة واعدة كتوقيع بيوماركري للتصنيف الخطر قبل الجراحة، على أن تتطلب الترجمة السريرية إجراء التحقق الخارجي والدراسات الوظيفية الإضافية.

الكلمات المفتاحية : سرطان الغدة النخامية, تعلم الآلة , , الانبثاث العقدي الليمفاوي, معهد الجينوم السرطاني

Introduction:

Pancreatic adenocarcinoma is a lethal cancer with increasing global incidence and mortality. It accounts for 1.8% of all cancers and 4.6% of cancer deaths worldwide (1). Highest incidence rates were noted in North America and Western Europe, with lower rates in Asia and Africa (2). Some risk factors for PAAD include advancing age, smoking, obesity, diabetes, chronic pancreatitis, and genetic predisposition (3,4). The disease often presents with late-stage symptoms, limiting surgical options to about 20% of cases (2). Despite its relative infrequency, pancreatic cancer poses a significant clinical, societal, and economic burden (1). Future projections suggest a nearly twofold

increase in pancreatic cancer deaths by 2060 (1). Lymph node metastasis (LNM) is a critical factor influencing prognosis in pancreatic cancer patients (5,6). The complex lymphatic drainage of the pancreas often leads to distant LNM, including para-aortic lymph nodes (PALN), particularly in advanced tumors (6). PALN involvement is significantly associated with arterial and perineural invasion (6). Surgical strategy depends on the pattern of PALN involvement and how it relates to other lymph node groupings (7). The lymph node ratio (LNR) is becoming a valuable predictor of survival after resection, outperforming other parameters (8,9). LNR, along with the number and extent of LN involvement, significantly impacts prognosis and can complement existing staging systems (10). It's interesting to note that patients with direct tumour expansion including lymph nodes have comparable survival rates to those with node-negative disease, casting doubt on the conventional wisdom regarding the influence of LNM on prognosis (11). These results highlight the intricacy of lymphatic dissemination in pancreatic cancer and the necessity of vigorous multi-modality treatment to enhance long-term survival (6). Long non-coding RNAs, or lncRNAs, are RNA transcripts longer than 200 nucleotides that do not contain significant open reading frames (12). Despite comprising a large portion of the human genome's transcriptional output, most lncRNAs have unknown functions (13). Nevertheless, studies have shown that lncRNAs are essential for a number of biological functions, such as the control of gene expression at the epigenetic, transcriptional, and post-transcriptional levels (14). Long non-coding RNAs (lncRNAs) are emerging as crucial players in cancer biology. These non-protein-coding transcripts, function as both tumor suppressors and oncogenes (15,16). LncRNAs regulate various cellular processes, including cell proliferation, apoptosis, and gene expression at epigenetic, transcriptional, and post-transcriptional levels (15,16). They interact with DNA, proteins, and other RNAs, influencing chromatin organization and directing ribonucleoprotein complexes (15,17). Genome-wide association studies have identified numerous lncRNAs associated with various cancer types, with their alterations promoting tumorigenesis and metastasis (18). The tissue-specific expression patterns of lncRNAs make them promising biomarkers and therapeutic targets for cancer diagnosis and treatment (18). Ongoing research focuses on unraveling the complex gene expression networks involving lncRNAs in cancer development (17). In cancer research, machine learning (ML) has become a potent tool, especially for prognosis, diagnosis, and prediction. With an emphasis on ovarian, prostate, and breast cancers, machine learning (ML) approaches such as decision trees, support vector machines, and artificial neural networks have been used to treat a variety of cancer types (19,20). These methods have demonstrated the ability to improve cancer detection accuracy by 15-25% and assist in

early diagnosis, which is crucial for patient outcomes(19,21). Personalised Medicine techniques have the potenital to enhance patient survuval and quality of life. This can be achieved by integrating machine learning (ML) and clinical data (20,21). Despite growing interest in the role of long non-coding RNAs (lncRNAs) in cancer, their specific functions and clinical relevance in pancreatic adenocarcinoma (PAAD) remain poorly explored. This lack of comprehensive investigation has hindered the development of lncRNA-based biomarkers and therapeutic targets in PAAD. Addressing this gap is essential for advancing our understanding of PAAD pathogenesis and for uncovering novel molecular tools that can improve diagnosis, prognosis, and treatment planning. Using information from TCGA-PAAD, this work attempts to create and assess machine learning models based on lncRNA expression profiles to forecast lymph node involvement in pancreatic adenocarcinoma. The study aims to enhance preoperative risk classification and assist clinical decision-making by finding patterns linked to nodal metastasis.

Literature Review

Recent advancements in molecular biology and machine learning have significantly deepened our understanding of pancreatic cancer (PCa) progression and improved the ability to predict clinical outcomes. Among these developments, long non-coding RNAs (lncRNAs) have emerged as pivotal regulators of key oncogenic processes, including tumor proliferation, invasion, epithelial-mesenchymal transition (EMT), angiogenesis, and resistance to chemotherapy. Furthermore, the integration of multi-omics datasets with advanced machine learning algorithms offers substantial promise for enhancing diagnostic precision and prognostic modeling in PCa.

Mo and colleagues (2023) performed a thorough review of lncRNAs in pancreatic cancer, which demonstrated their essential role in tumor proliferation and invasion, EMT, angiogenesis, and chemotherapy resistance. The research established that certain elevated lncRNA expression levels correlated with more severe disease progression and metastasis and reduced patient survival rates. The study recognized that small sample sizes and methodological inconsistencies create obstacles for clinical applications. The authors proposed using expression data in predictive models while suggesting that bioinformatics and machine learning approaches could connect molecular research to clinical applications.(22)

Li and colleagues (2022) developed a multi-omics machine learning system to predict pancreatic adenocarcinoma (PAAD) recurrence and metastasis through analysis of TCGA transcriptomic and microbiome data. They identified ten bacterial markers that showed the best ability to distinguish between metastatic and non-metastatic cases. The model used transcriptomic features including lncRNAs, miRNAs, and

mRNAs and showed strong predictive capabilities with lncRNAs achieving a high accuracy rate (AUC 0.791. This demonstrates that disease progression depends on the combination of microbial and lncRNA profiles. However, the study did not investigate lymph node involvement as a specific factor and did not create lncRNA-based models for nodal prediction. The study demonstrates machine learning potential for biomarker-based prognosis and supports additional development of lncRNA models to predict lymphatic spread in PAAD.(23)

A research by Alsharoh (2023) used TCGA RNA-sequencing data to discover metastatic pancreatic cancer-related differentially expressed long non-coding RNAs (lncRNAs). Alshroah performed differential expression analysis to find lncRNAs that contribute to disease progression. He trained four machine learning algorithms to distinguish between metastatic and non-metastatic cases. The models demonstrated good predictive accuracy (AUC 0.75 for the Random Forest Model). These findings suggest that these lncRNAs could serve as diagnostic and prognostic biomarkers. The research investigated metastasis in general terms, but its results apply to pancreatic adenocarcinoma (PAAD) lymph node involvement. The biological functions of the discovered lncRNAs need experimental verification. The preprint study demonstrates an effective approach but needs additional peer-reviewed validation to become fully established.(24)

A pivotal study by Chen et al. (2020) defined the lncRNA LINP1 as a key oncogenic mediator of pancreatic cancer. Their results indicated that the expression of LINP1 was greatly amplified in pancreatic tumor tissues as compared to control normal tissues near the tumors. Positive correlations of high LINP1 expression were also found to be correlated with higher frequency of distant metastasis as well as shorter overall survival of PCa patients. Mechanistically, the study revealed that LINP1 promotes the proliferation, invasion, and migration of PCa cells by directly interacting and “sponging” the microRNA-491-3p (miR-491-3p). The result inhibits the anti-tumor activity of miR-491-3p, and hence encourages the cancerous progression of the cancer.(25)

Widening the broad scene of the landscape of the lncRNAs of pancreatic cancer, the review of Ghafouri-Fard et al. (2021) summarized dozens of lncRNAs as oncogenes and/or tumor suppressors. The review highlights the point that oncogenic lncRNAs like MACC1-AS1, LINC00462, and UCA1 are upregulated and related to adverse survival. Conversely, lncRNAs as tumor suppressors are exhibited by MEG3, GAS5, and LINC00261. The review repeats the ceRNA hypothesis as the prime mode of action, enumerating dozens of lncRNA-miRNA pairs acting to regulate critical cancer-related pathways like the TGF-β/SMAD, PI3K/AKT, and MAPK/ERK cascades.(26)

Significance of ncRNAs also extends to their potential as biomarkers for predicting clinical outcomes, such as lymph node metastasis, which is one critical determinant of the treatment method and prognosis of the patient. Zhang et al. (2021) have developed and authenticated lymph node metastasis predictive models for nine different cancer types, including pancreatic adenocarcinoma, based on the expression signatures of mRNAs, miRNAs, and lncRNAs, gleaned from The Cancer Genome Atlas (TCGA). It has been suggested by their study that such classifiers built out of these molecular signatures might distinguish metastasis very precisely (with greater than 80% accuracy). For pancreatic cancer alone, the study has identified the minimum signature of just two lncRNA biomarkers that might efficiently predict risk of metastasis. What is involved here is the potential for using a narrow, effective set of ncRNA biomarkers for building powerful, minimally-invasive diagnosis tools for aiding clinical decision-making(27).

Fang et al. study focused mainly was on developing a 24-gene risk score model in order to predict the overall survival in patients with pancreatic adenocarcinoma. The machine based model, showed high accuracy in distinguishing between high- and low-risk patients, particularly highlighting lymph node metastasis as a key prognostic factor. The study utilized data from TCGA and GEO databases, and analyzed it using WGCNA and LASSO regression tools to pinpoint genes closely tied to lymph node involvement. Moreover the model proved capable of predicting one to three year survival outcomes(AUC : 0.81 to 0.92) with impressive precision, confirming its clinical value for PAAD prognosis(28).

Building on the previous theme of machine learning-based biomarker discovery, Zhang et al. study introduces CRlncRC, an innovative tool created for detecting cancer-related lncRNAs by relying on multiple features such a genomic, epigenetic, and expression data. Although not specifically designed for pancreatic cancer, the work highlights the effectiveness of machine learning in analyzing the involvement of lncRNA in cancer. Random forest was the best classifier in this study, outperforming other models in accurately detecting cancer-associated lncRNAs. This contributes broadly to the growing body of smart tools facilitating biomarker discovery in oncology(29).

Working on the same line of using machine learning and lncRNAs for prognosis, Ma et al. research focuses specifically on pancreatic cancer. The researchers in this study use random survival forest and cox regression models, to construct an 8-lncRNA signature, to effectively stratify high-risk PAAD patients and predict their survival outcomes. Similar to Fang et al. study, data from public databases such as TCGA, Cbioportal, and CCLE were harnessed to create and validate this model, which also achieved similar

accuracy (AUC up to 0.90). Beyond prognosis, the research also identified potential therapeutic strategies for high-risk patients, highlighting the clinical promise of lncRNA-based models in drug designing field(30).

Method:

1.Dataset Description and Preprocessing

For this study, two primary datasets were sourced from the UCSC Xena platform, both pertaining to pancreatic adenocarcinoma (PAAD) patients from The Cancer Genome Atlas (TCGA).

1. Clinical Dataset:

Clinical data for pancreatic adenocarcinoma were obtained from The Cancer Genome Atlas (TCGA) Pancreatic Adenocarcinoma (PAAD) cohort through the UCSC Xena Browser (https://tcga-xena-hub.s3.us-east 1.amazonaws.com/download/TCGA.PAAD.sampleMap%2FPAAD_clinicalMatrix) (31). This dataset contains curated clinical information for 196 TCGA-PAAD patients, including demographic variables (age, sex), pathological features (tumor stage, grade, lymph node involvement), and survival outcomes. Data were originally generated by The Cancer Genome Atlas Research Network and are publicly available through the Genomic Data Commons (GDC).

For the present analysis, only the lymph node involvement variable was extracted from the clinical dataset. Samples lacking this information were excluded, and only those with clearly defined lymph node status were retained for downstream analysis.

2. Gene Expression Dataset:

Gene expression data for pancreatic adenocarcinoma were obtained from The Cancer Genome Atlas (TCGA) Pancreatic Adenocarcinoma (PAAD) cohort via the GDC Xena Hub (https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-PAAD.star_tpm.tsv.gz) was used (31). This dataset provides transcript-per-million (TPM) normalized RNA-sequencing data generated using the STAR aligner as part of the GDC RNA-seq pipeline. It includes expression profiles for all annotated genes across 183 pancreatic adenocarcinoma samples obtained from TCGA.

The data were originally derived from the Genomic Data Commons (GDC) and subsequently processed by the UCSC Xena team for standardized normalization and integration.

Gene Annotation and Selection

The expression dataset included all annotated genes in the TCGA-PAAD cohort, encompassing protein-coding genes, pseudogenes, long non-coding RNAs (lncRNAs), and other transcript types. Gene annotation was performed using the GENCODE gene annotation database (release v36) (https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_36/gencode.v36.

annotation.gtf.gz), which provides comprehensive classification of human genes based on Ensembl gene IDs and biotypes (32).
Annotation and filtering were performed using Python (version 3.13.3). Expression data were merged with the GENCODE annotation file to map each Ensembl gene ID to its corresponding gene type (e.g., *protein_coding*, *lncRNA*, *pseudogene*, *miRNA*). The annotated dataset was then exported into separate files for each gene category. For the present analysis, only long non-coding RNA (lncRNA) expression data were retained for downstream analysis, while other gene types were excluded.

3. Data Integration:

To integrate clinical and molecular data, the TCGA-PAAD lncRNA expression dataset was merged with the clinical dataset containing lymph node involvement using the patient identifier (barcode) as the key. Only samples present in both datasets were retained for analysis to ensure consistency between clinical and molecular information.
Lymph node status, originally annotated as N0 (no regional lymph node metastasis) and N1 (presence of regional lymph node metastasis), was converted into a binary variable for downstream analysis: N0 = 0 and N1 = 1. This allowed the lymph node involvement to be treated as a binary outcome in subsequent analyses of lncRNA expression associations.

4. Class Imbalance Handling:

Initial class distribution showed an imbalance, with approximately 28% of patients
classified as N0 (LN-negative) and 72% as N1 (LN-positive). To address this and prevent
model bias, the dataset was balanced using Synthetic Minority Oversampling Technique (SMOTE) (33). This technique synthetically generates new samples for the
minority class based on its existing data distribution, resulting in a balanced dataset suitable
for machine learning analysis.
The final processed dataset consisted of balanced class labels with well-defined outcome variable: LN_Binary (0 = LN-negative, 1 = LN-positive) 127 each, high-dimensional lncRNA expression profiles including lncRNA name and TPM

Analytical Algorithms

Three machine-learning algorithms were implemented to compare performance and optimize predictive accuracy in identifying lncRNAs associated with lymph node involvement. These included XGboost Logistic Regression, Random. Each model was trained on the same dataset, and performance was assessed using accuracy, precision, recall, and area under

the ROC curve (AUC).

XGBoost algorithm

XGBoost, (Extreme Gradient Boosting), is a widely used machine learning algorithm in data mining. It's highly valued due to its speed, accuracy, and the ability to work well with large, various and sometimes incomplete datasets. XGBoost has become one of the standard tools in fields like healthcare analytics, financial forecasting, motion detection and customer behavior prediction.

The model works by building many small decision trees, each tree learning from the errors of the previous ones (34). The final prediction for each data point is a combination of all these trees. This idea is captured by the equation:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

The equations basically means that the predicted value ŷi is the total sum of the outputs from all the individual trees fk, applied to the input xi. Each tree contributes a small part to the final answer, gradually improving the overall accuracy. Moreover by employing parallel and distributed computing and using out-of-core computation, scientists are able to process hundred millions of examples on a desktop in relatively short time(34).

While XGBoost is powerful, it has its limitations. It can be computationally demanding and requires careful tuning to avoid overfitting. Moreover, its complexity also makes it more difficult to interpret than simpler models. However, due to its high performance, it has become indispensable tool in modern data mining workflows with several practical applications including detecting COVID-19 from chest X-rays (35) and predicting credit card fraud with high precision(36) among many others.

The model was trained using the eval_metric='logloss' to handle binary classification. Native XGBoost feature importance scores were visualized. SHAP values were computed using shap.TreeExplainer, which natively supports XGBoost, to identify key lncRNAs influencing predictions

1. Logistic Regression (LR)

Logistic regression (LR) is a statistical modeling technique broadly employed to predict binary outcomes (e.g., disorder presence/absence, treatment success/failure) based totally on predictor variables (37, 38). As an extension of linear regression, LR addresses the issue of modeling chances by way of transforming the linear predictor into a bounded variety [0, 1] using the logit hyperlink feature (39). This approach lets in researchers to quantify

relationships among danger elements (e.g., age, biomarkers) and effects while adjusting for confounders, making it quintessential for clinical choice aid and epidemiological research (40). Its mathematical basis, rooted within the logistic feature advanced via Verhulst in the nineteenth century, gives sturdy probabilistic outputs that are each interpretable and computationally green (Boateng and Abaye, 2019 (38).

Instead of predicting the outcome directly, logistic regression predicts the probability that the outcome will occur. This is done by modeling the log-odds (also called the logit) of the probability, which ensures that the predicted values always remain between 0 and 1 (38).

The main equation of logistic regression can be expressed as:

$$ln(P\ /\ (1-P))\ =\ \beta_0\ +\ \beta_1 X_1\ +\ \beta_2 X_2 + \ldots + \beta_k X_k$$

Here, P is the probability of the event occurring, X_1, X_2,....., XK are the predictor variables, β_0 is the intercept, and β_1,…,β_k are the regression coefficients that describe the effect of each predictor. The equation can also be rearranged to give the probability directly:

$$P\ =\ e\text{^}(\beta_0\ +\ \beta_1 X_1 + \ldots + \beta_k X_k)\ /\ (1\ +\ e\text{^}(\beta_0\ +\ \beta_1 X_1 + \ldots + \beta_k X_k))$$

In logistic regression, results are often interpreted using odds and odds ratios (OR). The odds are calculated as P/1−P , while the OR is e^{β}. The OR indicates how much the odds of the event change when a predictor increases by one unit. An OR greater than 1 means the odds increase, an OR less than 1 means the odds decrease, and an OR equal to 1 means there is no change (37). Logistic regression assumes that the outcome is binary, that the observations are independent, and that continuous predictors have a linear relationship with the log-odds of the outcome (37). Moreover, the model should be correctly specified by including relevant predictors whi le avoiding unnecessary variables, to ensure accurate and reliable results (38).

One of LR's strengths is its flexibility it works with both continuous and categorical variables, and it's particularly valuable in clinical and public health studies where adjusting for confounders is crucial (40). It can also handle specialized cases, such as zero-inflated datasets often found in disease modeling (41).

The method assumes; a linear relationship between predictors and the log-odds of the outcome, Independence of observations and no severe multicollinearity among predictors Evaluating a logistic regression model typically involves checking calibration (e.g., Hosmer–Lemeshow test), measuring discrimination ability, and using likelihood ratio tests to compare nested models. Adequate sample size often at least 10 events per predictor is essential to avoid overfitting (38). Even a well-specified model should be validated to ensure it performs reliably beyond the original dataset.

The model was trained with default L2 regularization. Feature coefficients

were extracted and visualized to identify the most influential lncRNAs based on their magnitude and direction (positive or negative correlation with LN metastasis). Regarding explainability, SHAP values were computed using shap.LinearExplainer to quantify each feature's contribution to predictions.

2. Random Forest (RF)

Random Forest is a supervised ensemble learning technique which addresses classification and regression problems because it delivers fast and effective predictive performance and noise resistance in addition to performing well with big datasets that have many features (42). Breiman (2000) developed Random Forest as an enhancement of decision trees since individual trees have a tendency to overfit training data (43).

The Random Forest method operates by training multiple decision trees on distinct random subsets of data features which then generate ensemble predictions (44). The ensemble approach implements bagging (bootstrap aggregating) by using replacement sampling to draw bootstrap samples for training each tree (45). Through this sampling method the trees remain diverse thus reducing their correlation which enhances the overall model stability.

The training process of Random Forest includes two random components where each tree selection process chooses a random subset of features before determining the optimal split (46). Through this method, the additional step of tree decorrelation results in improved generalization capabilities and reduced overfitting risk.

During classification Random Forest uses majority voting to determine the final prediction based on the outputs of individual decision trees within the forest (44). The final regression prediction results from averaging all individual tree predictions (44). The combination process generates more reliable prediction limits which leads to better model performance on new data.

Generalization is improved by averaging multiple decision trees, because it decreases variance and makes the model more resistant to overfitting (44,47). It proves efficient when working with large datasets that contain numerous features compared to the number of samples (44,47). Random Forest also demonstrates strong tolerance to missing values and outliers because its ensemble structure reduces the negative effects of noisy or incomplete data. Furthermore, it generates meaningful insights about feature importance which proves essential for feature selection and improves model interpretability (44,47).

The model was trained with n_estimators=100 and a fixed random_state for reproducibility. Feature Importance: Extracted from the average impurity decrease across trees. Permutation To improve reliability, permutation_ importance from scikit-learn was applied to the top 50 features, retraining a

smaller Random Forest on this subset. This method evaluates the change in model performance (ROC AUC) when each feature is randomly permuted.

Model Evaluation

All models were evaluated using 5-fold stratified cross-validation, using ROC AUC and accuracy as the scoring metric. Hold-out test set, using classification report metrics and ROC AUC .An overlap analysis between SHAP-selected and permutation-selected features was performed.

Result:

Three machine learning models, Logistic Regression, Random Forest, and XGBoost were trained to classify lymph node metastasis (LN-negative vs LN-positive) in pancreatic cancer patients using lncRNA expression profiles. Their performance was evaluated using 5-fold stratified cross-validation and an independent test set. As shown in Table 1, Logistic Regression achieved the highest test ROC AUC of 1.000, followed closely by Random Forest (0.998) and XGBoost (0.946), indicating excellent discriminative performance across models. Figure 1 displays comparison of performances.

Table 1. Performance of classification models

Model	Accuracy	Test ROC AUC
Logistic Regression	0.9608	1.0000
Random Forest	0.9608	0.9985
XGBoost	0.8824	0.9462

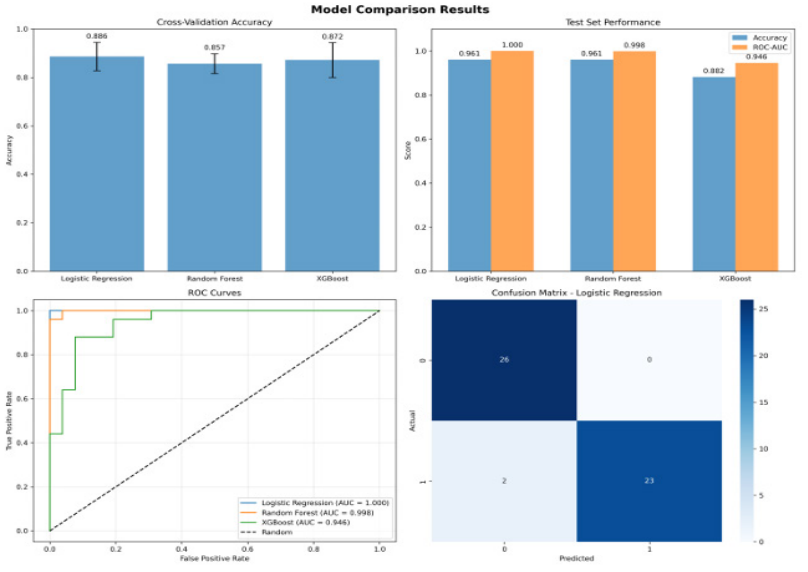


Figure 1: Comprehensive comparison of different machine learning models (Logistic Regression, Random Forest, and XGBoost) based on their performance metrics. The top left subplot displays the cross-validation accuracy with error bars, while the top-right subplot shows the accuracy and ROC-AUC scores on the test set. The bottom-left subplot illustrates the Receiver Operating Characteristic (ROC) curves for each model, and the bottom-right subplot provides the confusion matrix specifically for the Logistic Regression model.

To interpret model decisions and identify key predictive lncRNAs, feature importance was assessed using a combination of model-based and model-agnostic methods. Coefficients from Logistic Regression were visualized (Figure 2) to understand the directionality of each lncRNA's influence. Random Forest and XGBoost feature importances were also plotted (Figures 3 and 4, respectively). Additionally, to evaluate each feature's contribution to individual predictions, SHAP (SHapley Additive Explanations) values were computed for Logistic Regression and XGBoost. For the Random Forest model, permutation importance was utilized instead. The summary plots (Figures 5–7) revealed that several lncRNAs consistently ranked high across models.

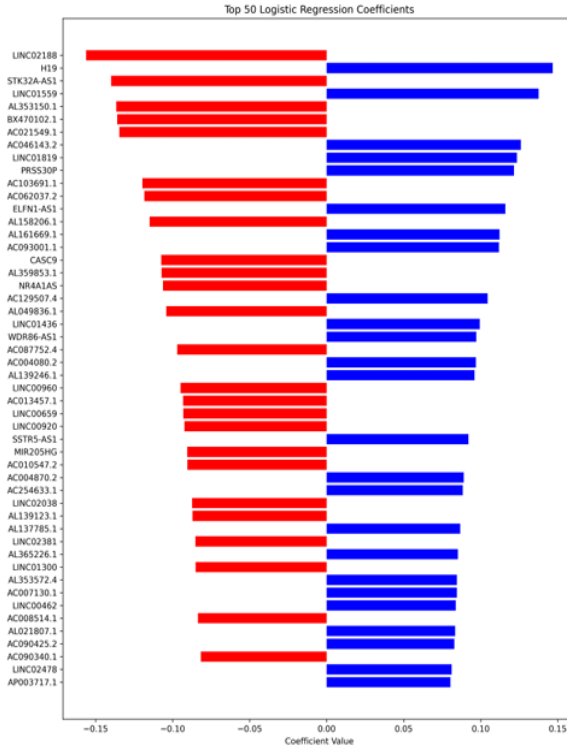
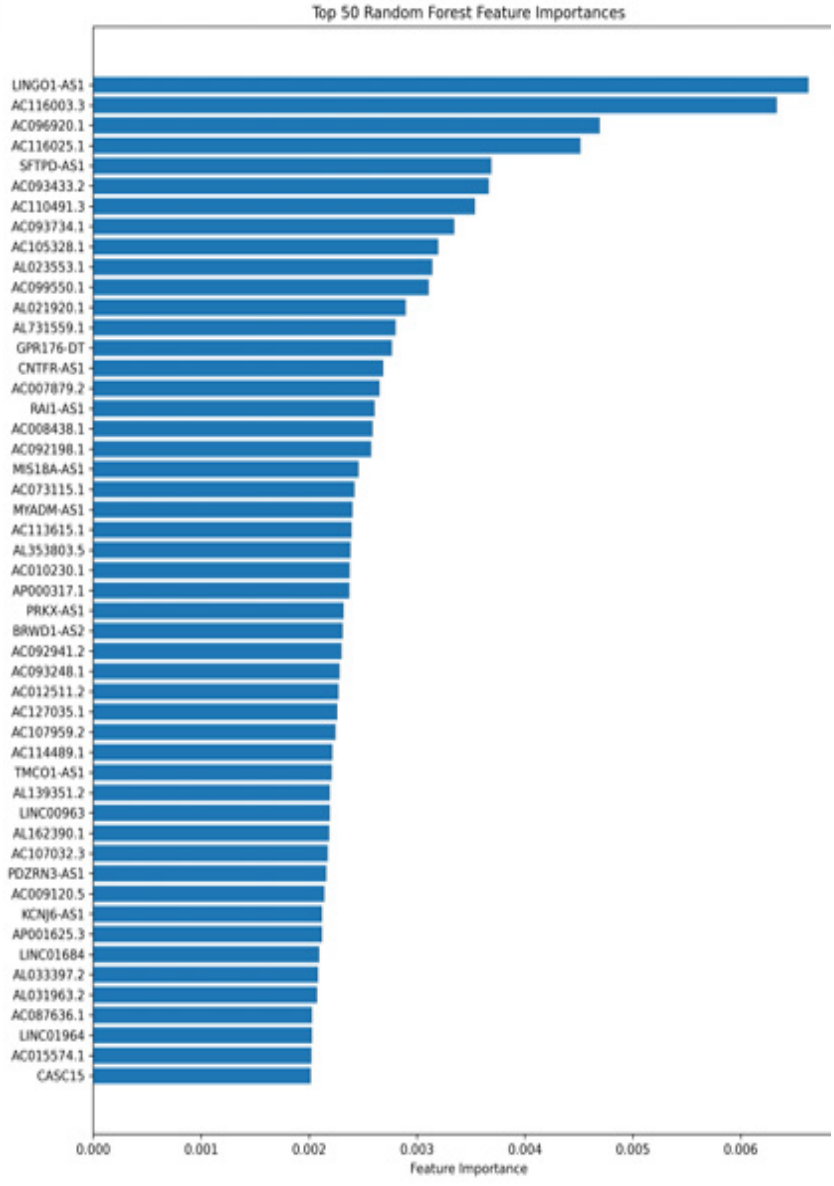


Figure 2. Top 50 coefficients from a Logistic Regression model. The x axis represents the coefficient value, with positive values (blue bars) indicating a positive correlation and negative values (red bars) indicating a negative correlation with the target variable. The y-axis lists the specific features or variables whose coefficients are being shown.



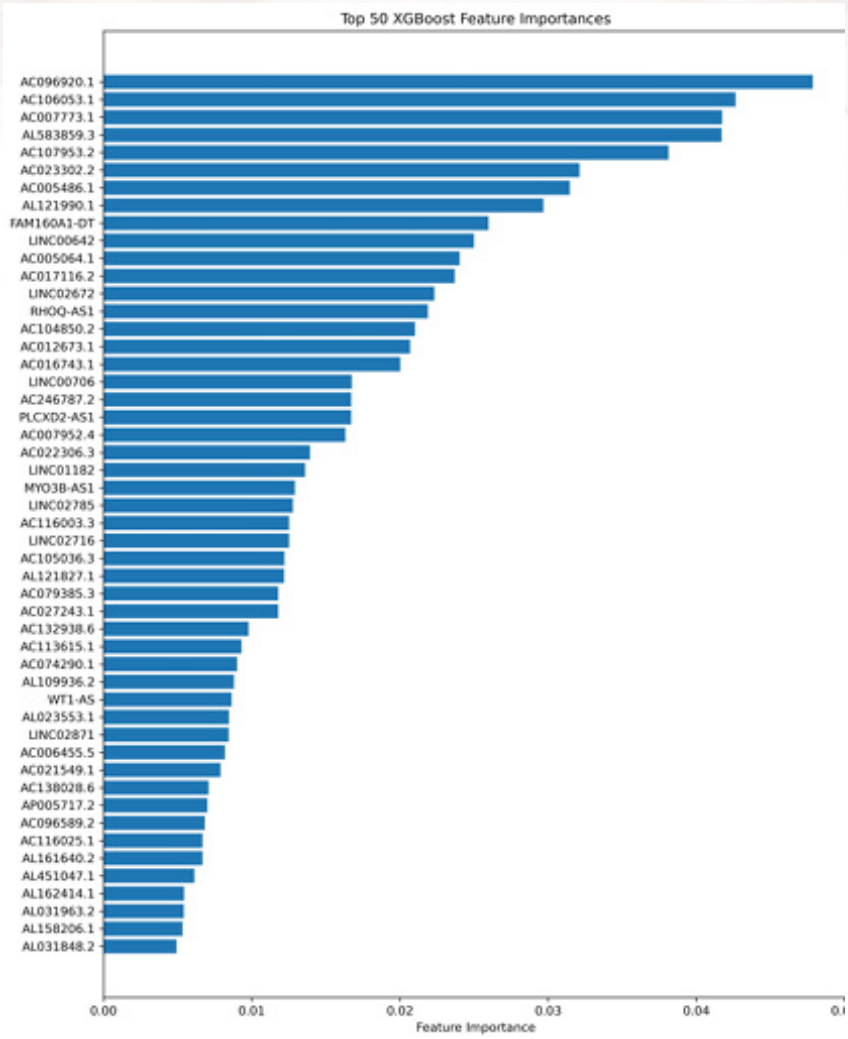


Figure 3. Top 50 feature importances derived from a Random Forest model. The x-axis represents the feature importance score, indicating the relative contribution of each feature to the model's predictions. The y-axis lists the specific features, ordered from most important (top) to least important (bottom) among the top 50.

Figure 4. Top 50 feature importances derived from a XGboost model. The x-axis represents the feature importance score, indicating the relative contribution of each feature to the model's predictions. The y-axis lists the specific features, ordered from most important (top) to least important (bottom) among the top 50.

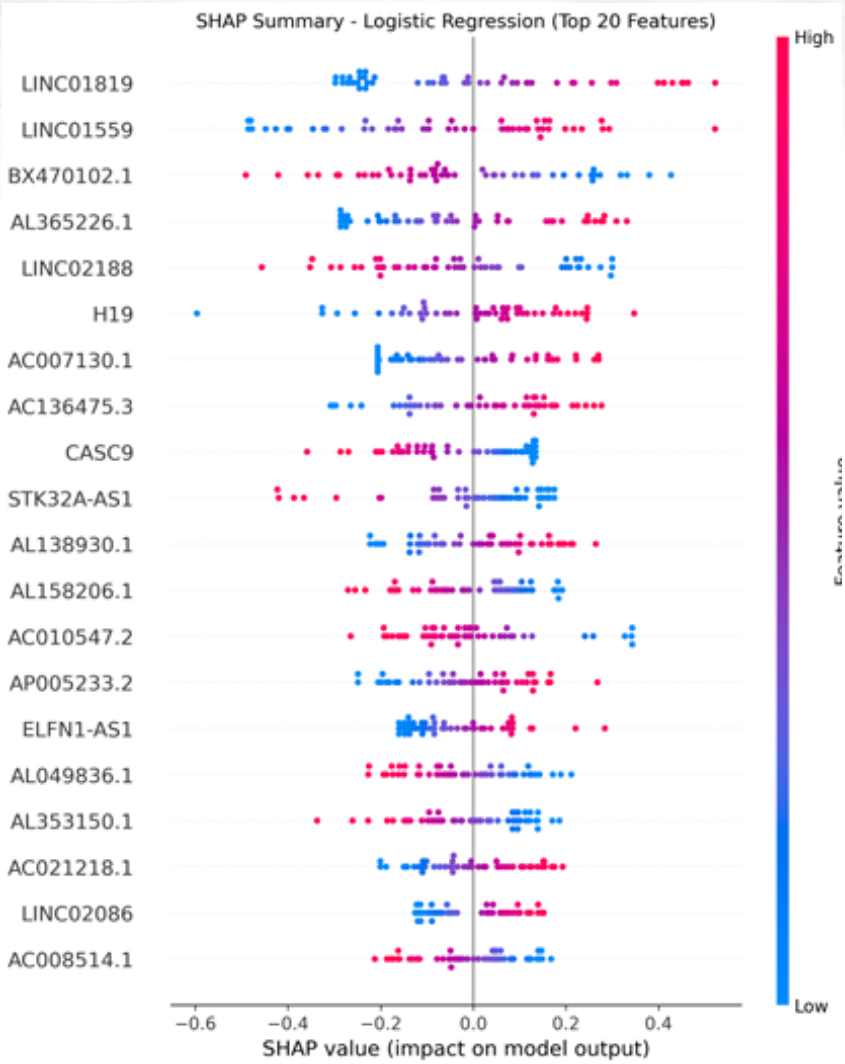


Figure 5. SHAP (SHapley Additive Explanations) summary plot visualizes the impact of the top 20 features on the Logistic Regression model's output. Each row represents a feature, and each point represents a single prediction. The horizontal position of a point indicates the SHAP value (impact on model output), where positive values increase the prediction and negative values decrease it. The color of each point indicates the feature value, with red indicating high values and blue indicating low values, as shown by the color bar on the right.

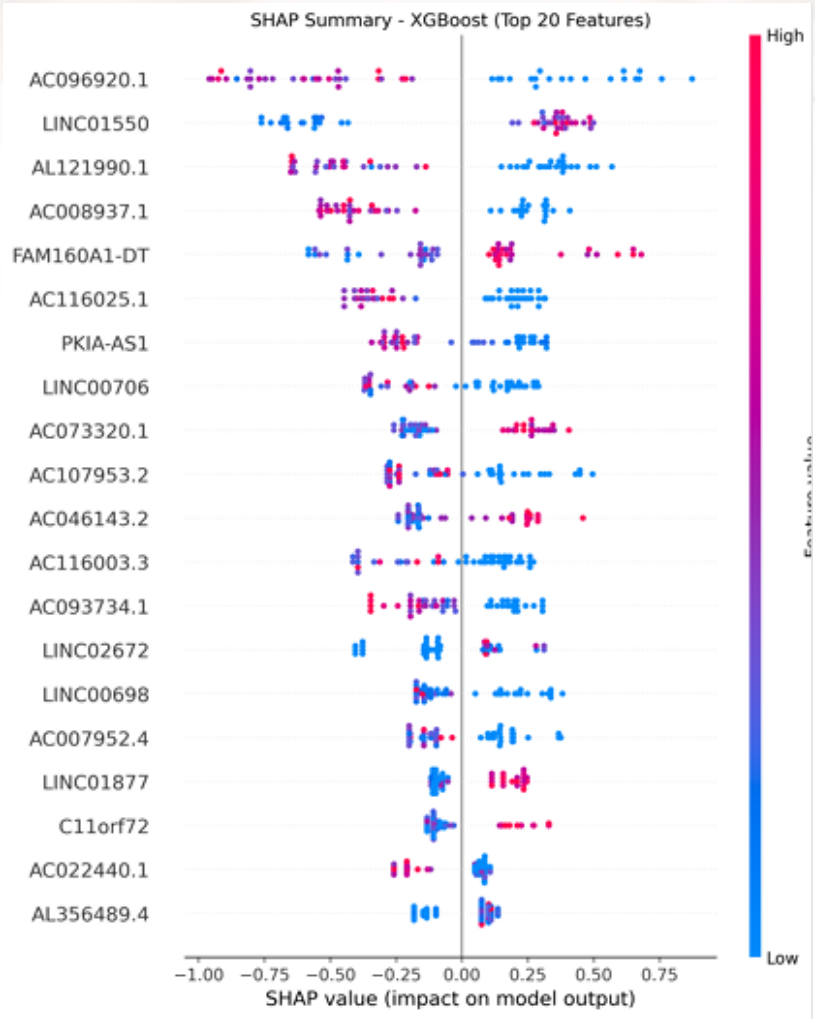


Figure 6. SHAP (SHapley Additive Explanations) summary plot visualizes the impact of the top 20 features on the XGboost model's output. Each row represents a feature, and each point represents a single prediction. The horizontal position of a point indicates the SHAP value (impact on model output), where positive values increase the prediction and negative values decrease it. The color of each point indicates the feature value, with red indicating high values and blue indicating low values, as shown by the color bar on the right.

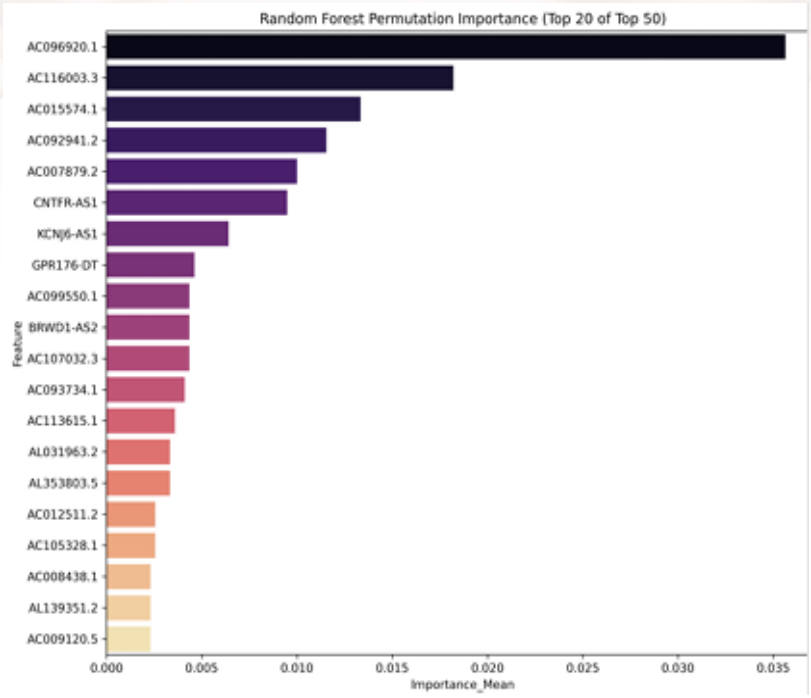


Figure 7. Top 20 features ranked by their permutation importance for the Random Forest model. The x-axis represents the mean importance score, indicating how much the model's performance decreases when the values of that feature are randomly shuffled. The y-axis lists the specific features, ordered from most important (top) to least important (bottom) among the top 20. The color gradient of the bars also visually represents the relative importance, transitioning from darker (more important) to lighter (less important).

A comparison between SHAP and permutation importance results yielded four overlapping lncRNAs: AC096920.1, AC116003.3, CNTFR-AS1, and AC093734.1. These lncRNAs were identified as the most robust and reliable predictors of lymph node status, appearing consistently across multiple importance-ranking frameworks. Expression patterns of these four overlapping lncRNAs were then visualized using boxplots, comparing their expression levels between LN-negative and LN-positive patient groups. The results, presented in Figures 8, showed clear differences in expression between the two classes. All four lncRNAs demonstrated statistically significant expression separation (Mann–Whitney U test, $p < 0.05$), further supporting their potential utility as biomarkers.

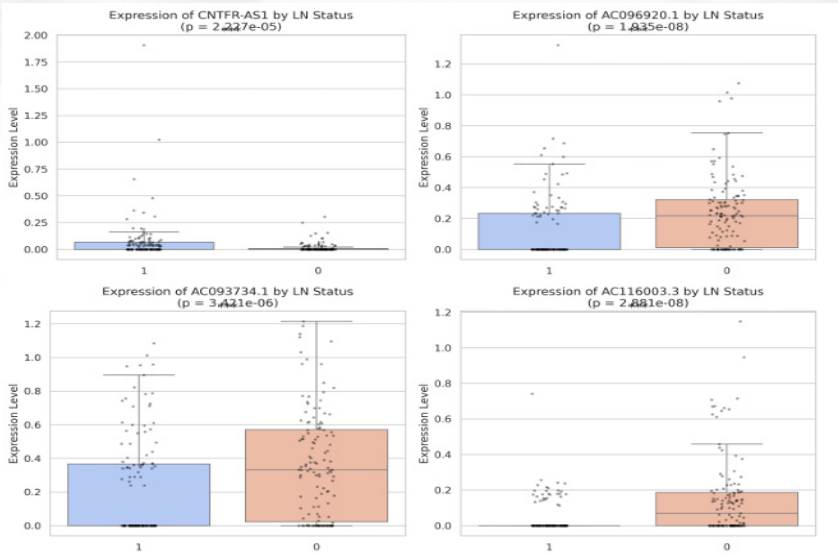


Figure 8. Box plot visualization of the expression levels of selected lncRNAs in relation to LN (lymph node) status. Each subplot displays the distribution of expression for a specific lncRNA, namely CNTFR-AS1 (top-left), AC096920.1 (top-right), AC093734.1 (bottom-left), and AC116003.3 (bottom-right). The x-axis differentiates between two LN status groups (labeled '1' and '0'), while the y-axis represents the expression level. Each plot also includes overlaid swarm plots showing individual data points and displays the p-value indicating the statistical significance of the differential expression between the groups.

Discussion:

This study highlights the remarkable potential of long non-coding RNA (lncRNA) expression profiles, when paired with machine learning (ML), to predict lymph node metastasis (LNM) in PAAD. The performance of the threemodels: Logistic Regression, Random Forest, and XGBoost were highly significant, with random forest having the highest score test ROC AUCs of 1.000, reflecting the better efficiency of the model for this specific task. These results could also indicate that the molecular "signature" carried by lncRNAs contains predictive information about the disease's behavior, even before the manifestation of symptoms. Similar results has been observed in recent studies: as Wen et al. (2024) demonstrated that ML applied to ultrasound image based features could predict LNM in PAAD with impressive accuracy, and Tang et al. (2024) showed that combining radiological data and genomic features, including information regarding RNAs, allowed reliable preoperative prediction of nodal involvement and hence an effective and non-invasive insights into the patient prognosis (48,49).

Another important characteristics of this study is the convergence of feature importance across different models. By leveraging both SHAP values and permutation importance, we consistently identified four lncRNAs: AC096920.1, AC116003.3, CNTFR-AS1, and AC093734.1 as strong predictors of lymph node status. Since these lncRNAs showed clear differences in expression between LN-positive and LN-negative patients , them they can act as red flags to indicate patients with high risk of disease progression. While these novel IncRNAs has not been publically reported before , other RNAs has been linked in previous studies to pancreatic cancer such as HOXA11-AS and LINC01559 which is up regulated in pancreatic adenocarcinoma (50,51). This reflects the complexity of the genetic landscape of pancreatic cancer and the inability to fully detect all the underling pathogenic mechanisms.

On the other hand, our study relied on TCGA-PAAD data, which may not fully capture the diversity of patient populations worldwide especially from developing countries. Moreover, High predictive performance as found in the current studies, while encouraging, does not equate to causal relationship and hence there is still need for functional studies to confirm the biological roles of these lncRNAs. Therefore more clinical studies are needed to confirm the prognosis significance of the four main IncRNAs detected in this study.

Conclusion:

This study demonstrates the significant potential of integrating long non-coding RNA(IncRNA) expression data with machine learning algorithms to predict lymph node involvement in pancreatic adenocarcinoma (PAAD). By analyzing TCGA-PAAD data we developed and evaluated three distinct models Logistic Regression, Random Forest, and XGBoost all of which exhibited excellent predictive performance with test ROC-AUC score reaching up to 1.0000. this indicating strong potential for clinical use. A key outcome was the identification of four IncRNA: AC096920.1, AC116003.3, CNTFR-AS1, and AC093734.1 as robust biomarker, showing significant differential expression between node-postive and node-negative.

These findings support the use of IncRNA-based classifiers for improving preoperative risk assessment and treatment planning. Future work should focus on external validation and functional studies to confirm biological mechanisms and clinical applicability.

Conflict of interest:

The authors have no conflict of interest to declare regarding this paper.

Ethical consideration:

No identifying information about the patients has been included in this study, ensuring that the patients privacy and confidentiality are fully protected.

References:

(1)Lippi G, Mattiuzzi C. The global burden of pancreatic cancer. Arch Med Sci. 2020;16(4):820–4.
(2)Simoes PK, Olson SH, Saldia A, Kurtz RC. Epidemiology of pancreatic adenocarcinoma. Chin Clin Oncol. 2017 June;6(3):24.
(3)Capasso M, Franceschi M, Rodriguez-Castro KI, Crafa P, Cambiè G, Miraglia C, et al. Epidemiology and risk factors of pancreatic cancer. Acta Bio Medica Atenei Parm. 2018 Dec 17;89(9-S):141–6.
(4)Yeo TP. Demographics, epidemiology, and inheritance of pancreatic ductal adenocarcinoma. Semin Oncol. 2015 Feb;42(1):8–18.
(5)The Complexity of the Pancreatic Lymphatic System and the Key Role of Para-Aortic Lymph Node Metastasis in Pancreatic Cancer Prognosis Prediction: A Comprehensive Review [Internet]. [cited 2025 July 28]. Available from: https://www.mdpi.com/2813-0545/3/2/10
(6)Kanda M, Fujii T, Nagai S, Kodera Y, Kanzaki A, Sahin TT, et al. Pattern of lymph node metastasis spread in pancreatic cancer. Pancreas. 2011 Aug;40(6):951–5.
(7)Analysis of paraaortic lymph node involvement in pancreatic carcinoma - Kayahara - 1999 - Cancer - Wiley Online Library [Internet]. [cited 2025 July 28]. Available from: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.1002/%28SICI%291097-0142%2819990201%2985%3A3%3C583%3A%3AAID-CNCR8%3E3.0.CO%3B2-J
(8)Yamamoto Y, Ikoma H, Morimura R, Konishi H, Murayama Y, Komatsu S, et al. The clinical impact of the lymph node ratio as a prognostic factor after resection of pancreatic cancer. Anticancer Res. 2014 May 1;34(5):2389–94.
(9)You MS, Lee SH, Choi YH, Shin BS, Paik WH, Ryu JK, et al. Lymph node ratio as valuable predictor in pancreatic cancer treated with R0 resection and adjuvant treatment. BMC Cancer. 2019 Oct 15;19(1):952.
(10)Li J, Zhang B, Cui G, Dai D. [Correlation between characteristics of lymph node metastases and prognosis in pancreatic cancer treated with pancreaticoduodenectomy]. Zhonghua Zhong Liu Za Zhi. 2014 Sept;36(9):688–92.
(11)Williams JL, Nguyen AH, Rochefort M, Muthusamy VR, Wainberg ZA, Dawson DW, et al. Pancreatic cancer patients with lymph node involvement by direct tumor extension have similar survival to those with node-negative disease. J Surg Oncol. 2015 Sept;112(4):396–402.
(12)Gong Z, Zhang S, Zhang W, Huang H, Li Q, Deng H, et al. Long non-coding RNAs in cancer. Sci China Life Sci. 2012 Dec 1;55(12):1120–4.
(13)Feng Y, Fan Y, Huiqing C, Zicai L, Quan D. [The emerging landscape of long non-coding RNAs]. Yi Chuan Hered. 2014 May;36(5):456–68.

(14)Rafiee A, Riazi-Rad F, Havaskary M, Nuri F. Long noncoding RNAs: regulation, function and cancer. Biotechnol Genet Eng Rev. 2018 Oct;34(2):153–80.
(15)Alsaedy H, Mirzaei A, Alhashimi RA. Investigating the Structure and Function of Long Non-Coding RNA (LncRNA) and Its Role in Cancer [Internet]. Rochester, NY: Social Science Research Network; 2023 [cited 2025 July 28]. Available from: https://papers.ssrn.com/abstract=4470175
(16)Alsaedy HK, Mirzaei AR, Alhashimi RAH. Investigating the structure and function of Long Non-Coding RNA (LncRNA) and its role in cancer. Cell Mol Biomed Rep. 2022 Dec 1;2(4):245–53.
(17)Yang G, Lu X, Yuan L. LncRNA: A link between RNA and cancer. Biochim Biophys Acta BBA - Gene Regul Mech. 2014 Nov 1;1839(11):1097–109.
(18)Bhan A, Soleimani M, Mandal SS. Long Noncoding RNA and Cancer: A New Paradigm. Cancer Res. 2017 Aug 1;77(15):3965–81.
(19)Applications of Machine Learning in Cancer Prediction and Prognosis - Joseph A. Cruz, David S. Wishart, 2006 [Internet]. [cited 2025 July 28]. Available from: https://journals.sagepub.com/doi/10.1177/117693510600200030
(20)Yue W, Wang Z, Chen H, Payne A, Liu X. Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. Designs. 2018 June;2(2):13.
(21)Jiang X. Exploring the Application of Machine Learning to Cancer Prediction. Sci Technol Eng Chem Environ Prot [Internet]. 2024 June 6 [cited 2025 July 28];1(1). Available from: https://www.deanfrancispress.com/index.php/te/article/view/923
(22)Mo Y, Adu-Amankwaah J, Qin W, Gao T, Hou X, Fan M, et al. Unlocking the predictive potential of long non-coding RNAs: a machine learning approach for precise cancer patient prognosis. Ann Med. 2023 Dec 12;55(2):2279748.
(23)Li S, Yang M, Ji L, Fan H. A multi-omics machine learning framework in predicting the recurrence and metastasis of patients with pancreatic adenocarcinoma. Front Microbiol. 2022 Nov 3;13:1032623.
(24)Alsharoh H. Machine learning predicts metastatic progression using novel differentially expressed lncRNAs as potential markers in pancreatic cancer [Internet]. Oncology; 2023 [cited 2025 July 28]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.11.01.23297724
(25)Chen AY, Zhang K, Liu GQ. LncRNA LINP1 promotes malignant progression of pancreatic cancer by adsorbing microRNA-491-3p. Eur Rev Med Pharmacol Sci. 2020 Sept;24(18):9315–24.
(26)26. Ghafouri-Fard S, Fathi M, Zhai T, Taheri M, Dong P. LncRNAs: Novel Biomarkers for Pancreatic Cancer. Biomolecules. 2021 Nov 10;11(11):1665.

(27)Zhang S, Zhang C, Du J, Zhang R, Yang S, Li B, et al. Prediction of Lymph-Node Metastasis in Cancers Using Differentially Expressed mRNA and Non-coding RNA Signatures. Front Cell Dev Biol. 2021 Feb 11;9:605977.

(28)28. Fang J, Wang M, Gao Y, Qi Y, Hong W, Xiao C. Prediction of overall survival in pancreatic cancer based on a twenty-four-gene risk model associated with lymph node metastasi. Medicine (Baltimore). 2025 May 16;104(20):e42448.

(29)Zhang X, Wang J, Li J, Chen W, Liu C. CRlncRC: a machine learning-based method for cancer-related long noncoding RNA identification using integrated features. BMC Med Genomics. 2018 Dec;11(S6):120.

(30)30. Ma D, Yang Y, Cai Q, Ye F, Deng X, Shen B. Identification of a lncRNA based signature for pancreatic cancer survival to predict immune landscape and potential therapeutic drugs. Front Genet. 2022 Sept 14;13:973444.

(31)Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, et al. Integrated Genomic Characterization of Pancreatic ductal adenocarcinoma. Cancer Cell [Internet]. 2017 Aug 1;32(2):185-203.e13. Available from: https://doi.org/10.1016/j.ccell.2017.07.007

(32)Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Research [Internet]. 2024 Nov 20;53(D1):D966–75. Available from: https://doi.org/10.1093/nar/gkae1078

(33)Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research [Internet]. 2002 Jun 1;16:321–57. Available from: https://doi.org/10.1613/jair.953

(34)Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2025 Aug 2]. p. 785–94. (KDD '16). Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

(35)Nasiri H, Hasani S. Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost. Radiography. 2022 Aug 1;28(3):732–8.

(36)Mohbey K, Khan M, Indian A. Credit-Card-Fraud-Prediction-Using-XGBoost -An-Ensemble-Learning-Approach. Int J Inf Retr Res. 2022 July 8;12.

(37)Boateng, E.Y. and Abaye, D.A. (2019) 'A Review of the Logistic Regression Model with Emphasis on Medical Research', *Journal of Data Analysis and Information Processing*, 7, pp. 190–207. doi:10.4236/jdaip.2019.74012.

(38). Diop, A., Diop, A. and Dupuy, J.F. (2011) 'Maximum likelihood estimation in the logistic regression model with a cure fraction', *Electronic Journal*Electronic Journal of Statistics*, 5, pp. 460–483. doi:10.1214/11-EJS616.

(39)Panda, N.R. et al. (2022) 'A Review on Logistic Regression in Medical Research', *National Journal of Community Medicine*, 13(4), pp. 265–270. doi:10.55489/njcm.134202222.

(40)Schober, P. and Vetter, T.R. (2021) 'Logistic Regression in Medical Research', *Anesthesia & Analgesia*, 132(2), pp. 365–366. doi:10.1213/ANE.0000000000005257.
(41)3Sperandei, S. (2014) 'Understanding Logistic Regression Analysis', *Biochemia Medica*, 24(1), pp. 12–18. doi:10.11613/BM.2014.003.
(42).Biau G, Scornet E. A random forest guided tour. Test [Internet]. 2016 Apr 19;25(2):197–227. Available from: https://doi.org/10.1007/s11749-016-0481-7
(43)Breiman L. Some infinite theory for predictor ensembles. Unpublished manuscript. 2000.Available from: https://www.stat.berkeley.edu/~breiman/infinite.pdf
(44).Salman HA, Kalakech A, Steiti A. Random Forest algorithm Overview. Deleted Journal [Internet]. 2024 Jun 8;2024:69–79. Available from: https://doi.org/10.58496/bjml/2024/007
(45)Novotny J, Bilokon PA, Galiotos A, Délèze F. Forests. In: Novotny J, Bilokon PA, Galiotos A, Délèze F, editors. Machine Learning and Big Data with kdb+/q. 1st ed. Chichester (UK): John Wiley & Sons; 2020. p. 495-508. doi:10.1002/9781119404729.ch25.
(46)Maudes J, Rodríguez JJ, García-Osorio C, García-Pedrajas N. Random feature weights for decision tree ensemble construction. Information Fusion [Internet]. 2010 Dec 3;13(1):20–30. Available from: https://doi.org/10.1016/j.inffus.2010.11.004
(47)Zhu T. Analysis on the applicability of the random forest. Journal of Physics Conference Series [Internet]. 2020 Aug 1;1607(1):012123. Available from: https://doi.org/10.1088/1742-6596/1607/1/012123
(48)Wen D yue, Chen J min, Tang Z ping, Pang J shu, Qin Q, Zhang L, et al. Noninvasive prediction of lymph node metastasis in pancreatic cancer using an ultrasound-based clinicoradiomics machine learning model. Biomed Eng OnLine. 2024 June 18;23(1):56.
(49)Tang Y, Su Y xi, Zheng J mei, Zhuo M ling, Qian Q fu, Shen Q ling, et al. Radiogenomic analysis for predicting lymph node metastasis and molecular annotation of radiomic features in pancreatic cancer. J Transl Med. 2024 July 29;22(1):690.
(50)Nishiyama H, Niinuma T, Kitajima H, Ishiguro K, Yamamoto E, Sudo G, et al. HOXA11-As Promotes Lymph Node Metastasis Through Regulation of IFNL and HMGB Family Genes in Pancreatic Cancer. Int J Mol Sci. 2024 Jan;25(23):12920.
(51)Lou C, Zhao J, Gu Y, Li Q, Tang S, Wu Y, et al. LINC01559 accelerates pancreatic cancer cell proliferation and migration through YAP-mediated pathway. J Cell Physiol. 2020 Apr;235(4):3928–38.