A large language model for predicting pancreatic ductal adenocarcinoma patients from blood-derived exosomal transcriptomics data

Shubham Choudhury[1], Naman Kumar Mehta[1], Gajendra P. S. Raghava[1, *]

[1] Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi-110020, India.

Mailing Address of Authors

Shubham Choudhury (SC): shubhamc@iiitd.ac.in ORCID ID: https://orcid.org/0000-0002-4509-4683
Naman Kumar Mehta (NKM): namanm@iiitd.ac.in ORCID ID: https://orcid.org/0009-0009-0244-2826
Gajendra P. S. Raghava (GPSR): raghava@iiitd.ac.in ORCID ID: https://orcid.org/0000-0002-8902-2876

***Corresponding Author**

Prof. Gajendra P. S. Raghava

Head and Professor

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

Okhla Industrial Estate, Phase III, (Near Govind Puri Metro Station)

New Delhi, India – 110020

Office: A-302 (R&D Block)

Phone: 011-26907444

Email: raghava@iiitd.ac.in

Website: http://webs.iiitd.edu.in/raghava/

ABSTRACT

Conventionally, machine learning models for text or sequence classification are developed by converting textual data into numerical features. In this study, we explore an inverse approach wherein numerical values/features are transformed into sequence representations and classified using natural language processing (NLP) methods, including large language models (LLMs). We applied this methodology to predict pancreatic ductal adenocarcinoma (PDAC) using the expression profiles of 50 genes from 284 PDAC and 217 non-PDAC patients. Gene expression values were converted into sequence data, with each gene represented as a residue in a 50-residue protein sequence. Major LLMs like PeptideBERT, ProtBERT, and ESM2 were fine-tuned on a protein training dataset and evaluated on an independent dataset. The best-performing model, esm2-t6_8M_UR50D, achieved an AUC of 0.95 on independent dataset. Additionally, an alignment-based approach employing BLAST and MERCI motifs was explored, and an ensemble model combining the LLM-based and alignment-based methods was developed. Our LLM-based model outperformed traditional machine learning models developed using exosomal transcriptomics data, offering a novel strategy for prediction of cancer biomarker prediction from transcriptomics data.

KEYWORDS: RNA, Large Language Model, Pancreatic ductal adenocarcinoma, Exosome, Transcriptomics

HIGHLIGHTS

- Prediction of PDAC patients based on transcriptomics data
- Convert numeric expression data to peptide sequence
- Leverage ability of LLMs to classify peptides
- Achieved performance better than conventional techniques

AUTHOR'S BIOGRAPHY

1. Shubham Choudhury is currently working as Ph.D. Student in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
2. Naman Kumar Mehta is currently working as Ph.D. Student in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
3. Gajendra P. S. Raghava is currently working as Professor and Head of Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC), which accounts for over 90% of all pancreatic tumors, is characterized by a robust inflammatory desmoplastic reaction [1]. It is the fourth leading cause of cancer-related deaths globally, with a five-year survival rate of less than 8% [2]. This dismal prognosis arises largely because PDAC is often diagnosed at advanced stages, leaving limited treatment options. Moreover, the incidence of PDAC is projected to more than double within the next decade due to factors such as population ageing, obesity, and type 2 diabetes, highlighting an urgent need for innovative and effective diagnostic strategies [3–5].

Current diagnostic approaches for PDAC rely on a combination of surgery, chemotherapy, radiotherapy, and imaging techniques, such as computed tomography (CT), magnetic resonance imaging (MRI), and endoscopic ultrasound (EUS). However, these methods face substantial limitations, including difficulties in detecting small or early-stage tumors, dependence on operator expertise, and high costs, making them less accessible and effective. Furthermore, with only 15-20% of patients eligible for surgery and most presenting with distant metastasis, the window for meaningful intervention is narrow [6,7].

Recognizing these challenges, researchers have sought innovative solutions, with transcriptomic markers in blood emerging as a particularly promising avenue. Despite significant progress, existing diagnostic strategies leveraging transcriptomic biomarkers—such as CA19-9 and extracellular vesicle protein markers like GPC1—often rely on traditional machine learning models like Random Forest (RF) and Lasso regression. While these methods have shown promise, they are limited in their ability to capture complex, high-dimensional patterns in gene expression data [8–11]. The emergence of the role of exosomal proteins has also been highlighted by computational methods that can accurately predict exosomal proteins [12,13]. Recent studies have demonstrated that even simple machine learning models, such as Logistic Regression and Support Vector Machines (SVM), can achieve high diagnostic accuracy when trained on curated biomarker panels [11,14–16]. Transcriptomics, which examines RNA molecules expressed by cells, provides a window into tumor activity and function. Tumor-derived RNA, including microRNAs, circulates in the bloodstream, offering a non-invasive means to identify unique molecular signatures of PDAC, even at early stages. Such

transcriptomic approaches hold the potential for developing highly sensitive and specific blood tests that could transform PDAC diagnosis and prognosis.

In 2019, Wang et al. developed a diagnostic biomarker based on the expression of eight genes [11]. The major highlight of this study was the source of the samples – the gene expression was calculated from blood-based extracellular vesicles. A SVM model was trained on gene expression values calculated in transcripts per million (TPM). The eight gene biomarker signature developed in this study achieved an AUC of 0.939 on the external validation dataset, highlighting the potential of non-invasive blood-based biomarkers compared to conventional.

Building on this foundation, we introduce a novel approach that employs large language models (LLMs) to analyze transcriptomic data for PDAC prediction. While LLMs have revolutionized natural language processing (NLP) by capturing intricate patterns in textual data, their application to numerical datasets has been largely unexplored. In this study, we adapted LLMs to process numerical transcriptomic features by converting them into text-based representations, enabling the models to leverage their robust pattern-recognition capabilities. We have utilized the dataset used by Wang et al. for training and validation. Our results demonstrate that the performance of LLM-based models is of a similar order to established methods, while providing a unique perspective on data interpretation and feature representation. By harnessing the power of LLMs, this study explores a novel and complementary tool for PDAC biomarker discovery and early detection, paving the way for further advancements in transcriptomic analysis.

MATERIAL AND METHODS

In order to provide a clear overview of our methodological approach, we present Figure 1, which illustrates the key steps involved in our analysis. This figure outlines the process of utilizing a transcriptome profiling microarray (TPM) expression dataset obtained from the Gene Expression Omnibus (GEO), followed by the feature selection phase aimed at identifying genes predictive of pancreatic ductal adenocarcinoma (PDAC) samples. Post feature selection, these values were used to generate peptide sequences that was used to train LLMs. This visual representation serves to summarize the workflow and key components of our study, facilitating a better understanding of our analytical strategy.

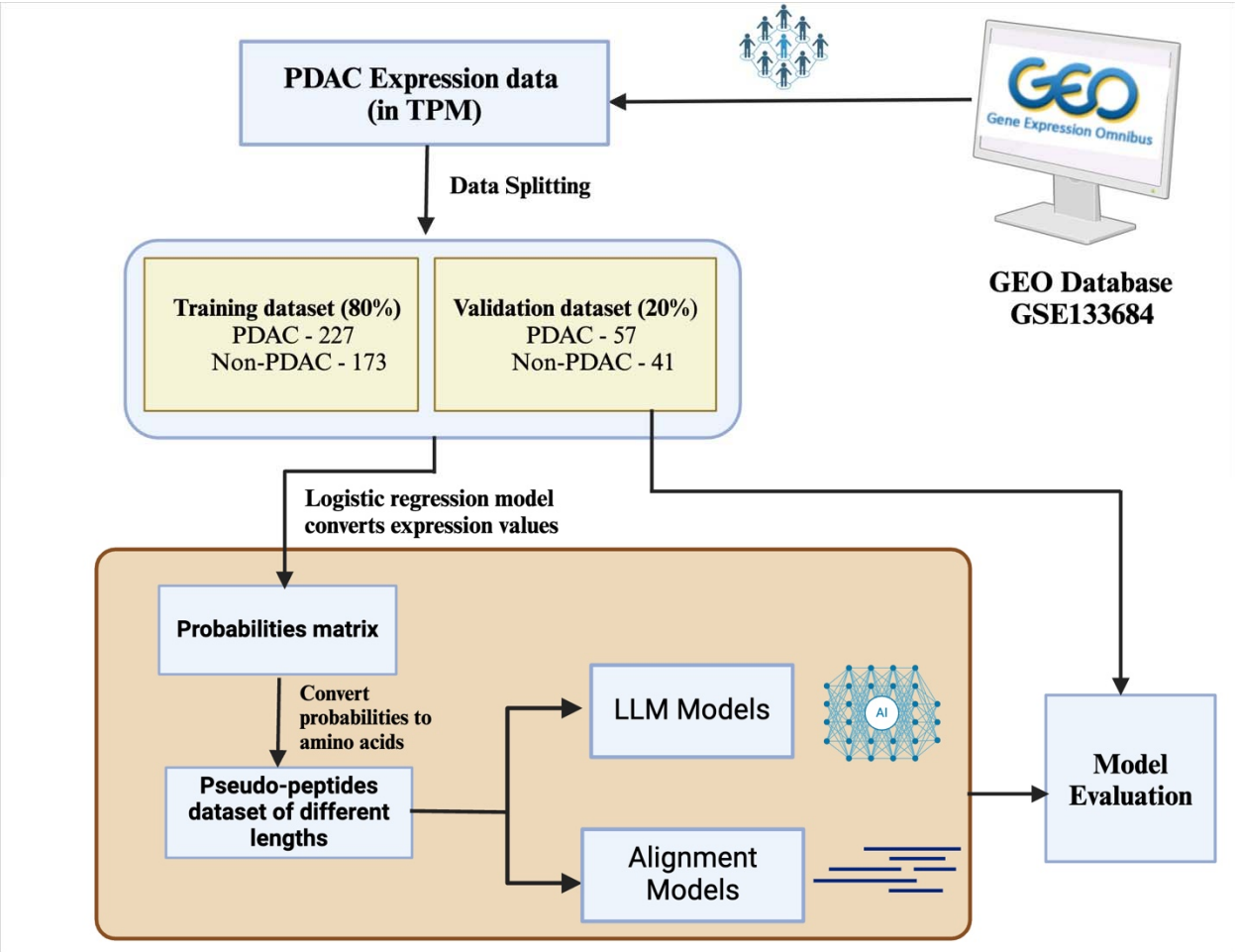

Figure 1- *Methodology overview*

Dataset creation

Publicly available sequencing data was obtained from GEO (accession number: GSE133684) which was used in the study by Wang et al. [11]. Read counts were converted to TPM values by the depositors and the same was used for our study. The dataset consisted of 501 patient samples, which included - 284 PDAC patients, 100 Chronic Pancreatitis (CP) patients and 117 Healthy Controls (HC). The number of genes reported in this dataset was 54,148. For our analysis, labels were generated for each sample – with the positive label being PDAC samples (284 samples) and the negative label for CP and HC samples (217 samples). Subsequently the entire dataset was split into 2 parts - 80% in the training dataset (401 samples) and 20% in the validation dataset (100 samples).

Expression to Probability

Following the preparation of the dataset, the TPM values were transformed into a probability matrix of the same dimensions (501 samples × 54,148 genes) to serve as features for subsequent analysis. This conversion was achieved by training a logistic regression (LR) model on each gene individually using the training dataset (401 samples). For each gene, the LR model was fitted to predict the binary outcome (PDAC vs. non-PDAC) based on its TPM values, and then applied to both the training and validation datasets to generate probability scores. This process was iteratively performed for all 54,148 genes, resulting in a probability matrix where each entry represents the likelihood of a sample being classified as PDAC based on the expression of a single gene. These probability values were then utilized as input features for the next stage of our analysis.

Feature Selection

To reduce the feature space from 54,148 to a manageable range, three feature selection methods were employed: Minimum Redundancy Maximum Relevance (mRMR) [17], Random Forest Out-of-Bag Error (RF-OOB) [18], and Least Absolute Shrinkage and Selection Operator (LASSO) [19]. Each method generated varying sets of features, ranging from 5 to 25 in increments of 5, which were subsequently evaluated using different machine learning models to identify the optimal feature set.

The mRMR method aims to select features that enhance relevance to the target variable while minimizing redundancy, making it particularly useful in high-dimensional datasets. It was implemented in Python using the *pymrmr* library. Random Forest is an ensemble learning technique that builds multiple decision trees and utilizes the Out-of-Bag error for internal validation. This method assesses feature importance by measuring the increase in prediction error when feature values are permuted. LASSO, on the other hand, performs variable selection and regularization by adding a penalty based on the absolute values of coefficients, promoting sparsity in the model. This allows for some coefficients to become zero, effectively simplifying the model. LASSO was implemented using the sklearn library in Python, demonstrating its effectiveness in feature selection alongside RF-OOB and mRMR.

Textual representation

A probability matrix was generated from patient gene expression values, where rows represent samples and columns correspond to genes. To generate sample-specific peptides, each probability value for all the genes selected was converted into an amino acid based on predefined probability ranges. Specifically, probabilities between 0 and 0.05 were mapped to alanine (A), 0.05 to 0.10 to cysteine (C), and so forth, following a systematic mapping scheme. A sample conversion of probabilities to peptides is shown in Table 1. For each sample, the amino acids derived from the probability values across all genes were concatenated to produce a unique peptide sequence for each sample. This process was repeated for all samples, resulting in a set of sample-specific peptide sequences reflective of their gene expression profiles.

Table 1 - Assignment of probabilities to amino-acids and a toy data indicating how the probabilities were converted to pseudo-amino acids

			G 1		G 2		G 3		G 4		G 5		Pseudo-peptide
Prob	**AA**	**Patient**	**Prob**	**Symbol**	**Prob**	**Symbol**	**Prob**	**Symbol**	**Prob**	**Symbol**	**Prob**	**Symbol**	
0.00 - 0.05	A	P 1	0.905	W	0.897	V	0.948	W	0.947	W	0.591	N	WVWWN
0.05 - 0.10	C	P 2	0.968	Y	0.904	W	0.963	Y	0.924	W	0.910	W	YWYWW
0.10 - 0.15	D	P 3	0.958	Y	0.897	V	0.955	Y	0.858	V	0.762	S	YVYVS
0.15 - 0.20	E	P 4	0.958	Y	0.904	W	0.953	Y	0.917	W	0.891	V	YWYWV
0.20 - 0.25	F	P 5	0.933	W	0.812	T	0.940	W	0.937	W	0.580	N	WTWWN
0.25 - 0.30	G	P 6	0.942	W	0.886	V	0.946	W	0.944	W	0.620	P	WVWWP
0.30 - 0.35	H	P 7	0.936	W	0.874	V	0.942	W	0.845	T	1.000	Y	WVWTY
0.35 - 0.40	I	P 8	0.922	W	0.878	V	0.933	W	0.941	W	0.532	M	WVWWM
0.40 - 0.45	K	P 9	0.899	V	0.833	T	0.919	W	0.913	W	0.596	N	VTWWN
0.45 - 0.50	L	P 10	0.947	W	0.895	V	0.948	W	0.839	T	0.807	T	WVWTT
0.50 - 0.55	M	P 11	0.088	C	0.415	K	0.250	G	0.215	F	0.523	M	CKGFM
0.55 - 0.60	N	P 12	0.190	E	0.225	F	0.106	D	0.004	A	0.484	L	EFDAL
0.60 - 0.65	P	P 13	0.218	F	0.017	A	0.104	D	0.278	G	0.479	L	FADGL

0.65 - 0.70	Q	P 14	0.512	M	0.605	P	0.513	M	0.006	A	0.493	L	MPMAI
0.70 - 0.75	R	P 15	0.304	H	0.446	K	0.498	L	0.201	F	0.563	N	HKLFN
0.75 - 0.80	S	P 16	0.096	C	0.031	A	0.266	G	0.187	E	0.485	L	CAGEI
0.80 - 0.85	T	P 17	0.537	M	0.010	A	0.574	N	0.753	S	0.619	P	MANSI
0.85 - 0.90	V	P 18	0.414	K	0.801	T	0.504	M	0.032	A	0.451	L	KTMAI
0.90 - 0.95	W	P 19	0.261	G	0.394	I	0.240	F	0.202	F	0.516	M	GIFFM
0.95 - 1.00	Y	P 20	0.138	D	0.681	Q	0.746	R	0.419	K	0.555	N	DQRKN

Large Language Models

To classify PDAC patients based on the generated sample-specific peptide sequences, we employed several large language models (LLMs), including PeptideBERT [20], ProtBERT [21], t6_8M_UR50D [22], t12_35M_UR50D [22], and t33_650M_UR50D [22]. Each model was fine-tuned on the peptide sequences derived from our probability matrix, enabling the extraction of relevant features that capture the biological significance of the peptides. The training process involved using the training dataset (401 samples) to optimize the models for distinguishing PDAC patients from non-PDAC samples. Following training, we evaluated the performance of each model on the validation dataset (100 samples) using metrics such as accuracy, precision, recall, and F1-score. This comprehensive approach aimed to identify the most effective LLM for classifying PDAC patients, thereby enhancing our understanding of the molecular signatures associated with pancreatic cancer.

Alignment-based methods

In this study, alignment-based methods were employed to classify pseudo-peptides of length-50 using PSI-BLAST and the MERCI motif search tool [23,24]. For the PSI-BLAST analysis, a database was created using 80% of the training dataset, and sequences from the validation dataset were queried against this database. The label of the best-matching sequence in the database was assigned to the query pseudo-peptide. Concurrently, the MERCI tool was used to identify motifs specific to both the positive and negative classes in the training data. These class-specific motifs were then applied to classify the validation sequences: if a positive-class motif was detected in a validation sequence, it was assigned a positive label, whereas the presence of a negative-class motif resulted in a negative label.

Evaluation metrics

The binary classification performance of our fine-tuned model was evaluated using the following metrics: Sensitivity (SENS), Specificity (SPEC), Precision (PREC), Accuracy (ACC), Matthew's Correlation Coefficient (MCC), F1-Score (F1) and Area Under the Receiver Operator Characteristic curve (AUC). The aforementioned metrics were calculated using the four different types of prediction outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN):

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

$$F1 - score = \frac{TP}{TP + (0.5 \times (FN + FP))}$$

The evaluation of our binary classification model using various metrics provides critical insights into its performance. Sensitivity (SENS) measures the model's ability to identify positive instances, while Specificity (SPEC) assesses its accuracy in recognizing negative instances. Precision (PREC) reflects the accuracy of positive predictions, and Accuracy (ACC) offers an overall measure of correctness, though it may be misleading in imbalanced datasets. Matthew's Correlation Coefficient (MCC) provides a balanced view by considering all prediction outcomes, with values close to 1 indicating strong predictive capability. The F1-Score (F1) combines Precision and Sensitivity into a single metric, ideal for balancing the trade-off between false positives and negatives. Finally, the Area Under the Curve (AUC) evaluates the model's ability to distinguish between classes across different thresholds, with higher values indicating better performance. Together, these metrics enable a comprehensive evaluation of the model, guiding necessary improvements and refinements.

Ensemble methods

In this ensemble approach, the prediction probabilities from LLMs were combined with motif information from MERCI and sequence similarity data obtained through PSI-BLAST to provide a more accurate and biologically meaningful prediction of PDAC class peptide. Similar to how motifs from MERCI were incorporated, the query peptide was compared against the test data using PSI-BLAST, with varying e-values. If any motif or similarity hits were detected in the peptide, the probability score for the peptide was adjusted: for positive class peptides, the score was increased by 0.5, while for negative class peptides, the score was decreased by 0.5. This adjustment allowed the model to output a probability score, reflecting the likelihood of a peptide belonging to a specific class rather than a simple binary outcome.

RESULTS

In this research, we aimed to develop a predictive model using a large language model (PeptideBERT) that is capable of identifying phenotypes of pancreatic ductal adenocarcinoma (PDAC) by analyzing gene expression data derived from blood-based exosomes. To achieve this objective, we utilized expression data obtained from a publicly accessible Gene Expression Omnibus (GEO) dataset. This dataset was generated using a non-invasive method for extracting RNA from blood-based exosomes, which were subsequently sequenced to gather the necessary genetic information.

Selection of relevant genes

In this study, we focused exclusively on identifying discriminating genes that could serve as potential biomarkers for pancreatic ductal adenocarcinoma (PDAC). By employing multiple feature selection methods, including minimum Redundancy Maximum Relevance (mRMR), Least Absolute Shrinkage and Selection Operator (LASSO), and random forest (RF) out-of-bag methods, we successfully isolated a relevant set of genes that exhibited significant differences in expression levels. Five sets of genes were identified using each algorithm and starting from 5 to 25 in increments of 5. This targeted approach enhances our ability to pinpoint genetic features that may contribute to the diagnosis and understanding of PDAC. The 5-gene set identified by RF-OOB algorithm was selected for further analysis as they showed significant diagnostic

potential while keeping the number of genes small. Figure 2 highlights the difference in average gene expression for the 5 selected genes between PDAC and non-PDAC samples.

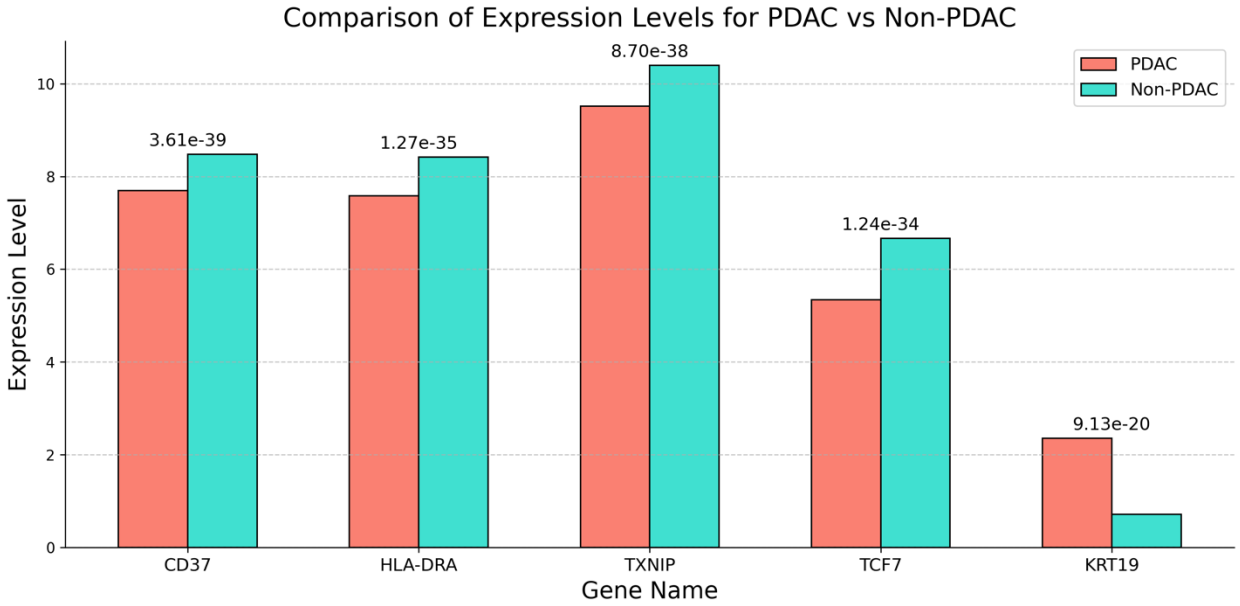

Figure 2 - Bar plot highlighting the difference in the average expression of the genes selected by the RF-OOB algorithm between PDAC and non-PDAC samples. The values above the bar indicate the p-values.

Performance of LLMs

In this study, we employed a set of generated peptides to train several large language models (LLMs)—PeptideBERT, ProtBert, t6_8M_UR50D, t12_35M_UR50D, and t33_650M_UR50D - for the classification of pancreatic ductal adenocarcinoma (PDAC) patients. The models were trained on peptide sequences derived from PDAC patient samples to differentiate between diseased and healthy states, with performance evaluated using the area under the receiver operating characteristic curve (AUC). Among the models, t6_8M_UR50D achieved the highest AUC of 0.9460, followed closely by t12_35M_UR50D (AUC = 0.9381) and t33_650M_UR50D (AUC = 0.9345), suggesting robust discriminatory power influenced by their pre-training on the UR50D dataset. PeptideBERT and ProtBert also exhibited strong performance with AUC values of 0.9266 and 0.9202, respectively, likely benefiting from their specialized architectures for

peptide and protein sequence analysis. The detailed comparison between the LLM models on the 50-length dataset is provided in Table 2. Table 3 shows the AUC values for all the LLM models corresponding to each length of peptide used in the study. These results underscore the effectiveness of peptide-trained LLMs for PDAC classification and support further exploration of model optimization for clinical applications.

Table 2 - Performance of LLMs on the validation dataset for peptide length 50

Model	Sensitivity	Specificity	Precision	Accuracy	MCC	F1-score	AUC
Peptidebert	0.8750	0.8667	0.8909	0.8713	0.7402	0.8829	0.9266
Protbert	0.7500	0.9333	0.9333	0.8317	0.6833	0.8317	0.9202
t6_8M_UR50D	0.9286	0.8444	0.8814	0.8911	0.7795	0.9043	0.9460
t12_35M_UR50D	0.9107	0.8667	0.8947	0.8911	0.7792	0.9027	0.9381
t33_650M_UR50D	0.9286	0.8444	0.8814	0.8911	0.7795	0.9043	0.9345

Table 3 - AUC of LLMs on the validation dataset for peptide lengths - 5, 10, 15, 20, 25, 30, 40, 50

Model	k=5	k=10	k=15	k=20	k=25	k=30	k=40	k=50
Peptidebert	0.9393	0.9433	0.9262	0.9464	0.9262	0.919	0.9357	0.9266
Protbert	0.8937	0.9099	0.9175	0.8893	0.9337	0.8956	0.9246	0.9202
t6_8M_UR50D	0.9417	0.9333	0.948	0.9313	0.9139	0.9194	0.931	0.946
t12_35M_UR50D	0.9405	0.9313	0.9437	0.9306	0.9214	0.9091	0.9317	0.9381
t33_650M_UR50D	0.9369	0.9286	0.927	0.9377	0.9262	0.9028	0.9381	0.9345

Performance of Alignment based method

PSI-BLAST was used to perform a similarity search against a training dataset consisting of pseudo-peptides of length 50. The search was conducted with PSI-BLAST using an e-value range from 10^-7 to 10. Optimal performance was achieved at an e-value of 10^-3, correctly

identifying 32 pseudo-peptides of the target class and 16 pseudo-peptides from the non-target class, with 10 incorrect hits. E-values lower than 10^-3 resulted in inadequate sequence coverage, while higher e-values led to a higher error rate. For peptide sequences of length 5, no hits were obtained while BLAST query. The performance of this approach for peptide length 50 is shown in Table 4.

Table 4 - Results for the BLAST-based alignment approach on the validation dataset

E-value	T_hits	C_hits	I_hits	P_hits	N_hits	CP_hits	CN_hits
1.00E+03	101	57	44	48	29	9	15
1.00E+02	101	57	44	48	29	9	15
1.00E+01	101	57	44	48	29	9	15
1.00E+00	97	56	41	46	27	10	14
1.00E-01	82	48	34	41	23	7	11
1.00E-02	58	37	21	32	16	5	5
1.00E-03	41	30	11	27	9	3	2
1.00E-04	31	27	4	24	2	3	2
1.00E-05	25	23	2	23	0	0	2

Similarly, we used the MERCI tool for motif search, testing various fp and k values. Among them, fp 15 was chosen as the optimal parameter due to its high accuracy, even though it offered limited coverage. The result of the motif search is provided in Table 5.

Table 5 - Results for the MERCI-based alignment approach on the validation dataset for peptides of length 50

K	T_hits	C_hits	I_hits	P_hits	N_hits	CP_hits	CN_hits
10	19	18	1	18	0	18	0
100	28	27	1	27	0	27	0
1000	28	27	1	27	0	27	0

Performance of Ensemble models

To further enhance the predictive performance of our models, we implemented ensemble methods that combined the predicted probabilities from the various large language models (LLMs) trained on peptide sequences with alignment-based methods. Specifically, we utilized both the MERCI algorithm and BLAST to integrate the predictions from PeptideBERT, ProtBert, t33_650M_UR50D, and t12_35M_UR50D. By combining the predicted probabilities with the sequence similarity information provided by these methods, we aimed to refine our classification of PDAC patients and leverage the strengths of both the LLMs and the alignment-based approaches. The integration of predictions with MERCI resulted in a significant improvement in classification accuracy as compared to BLAST, demonstrating the effectiveness of combining predictive modeling with sequence alignment techniques. The performance of LLMs combined with BLAST and MERCI on the validation dataset are reported in Table 6 and Table 7, respectively.

Table 6 – Performance of LLM + BLAST ensemble model on the validation dataset

Model	**k=10**	**k=15**	**k=20**	**k=25**	**k=30**	**k=40**	**k=50**
Peptidebert	0.9433	0.9278	0.9472	0.9266	0.9198	0.9079	0.9214
Protbert	0.9099	0.9175	0.8901	0.9337	0.9063	0.9079	0.9131
t6_8M_UR50D	0.9333	0.9480	0.9329	0.9131	0.9242	0.9341	0.9468
t12_35M_UR50D	0.9317	0.9437	0.9317	0.9206	0.9119	0.9349	0.9417
t33_650M_UR50D	0.9290	0.9270	0.9389	0.9290	0.9143	0.9369	0.9385

Table 7 - Performance of LLM + MERCI ensemble model on the validation dataset

Model	**k=5**	**k=10**	**k=15**	**k=20**	**k=25**	**k=30**	**k=40**	**k=50**
Peptidebert	0.9468	0.9484	0.9278	0.9466	0.9286	0.9208	0.9365	0.9302
Protbert	0.9089	0.9204	0.9183	0.8994	0.9353	0.9091	0.9296	0.9296
t6_8M_UR50D	0.9464	0.9405	0.9492	0.9395	0.9218	0.924	0.9363	0.9518
t12_35M_UR50D	0.946	0.9375	0.9464	0.9381	0.9272	0.9145	0.9409	0.9464

t33_650M_UR50D	0.944	0.9357	0.9341	0.9411	0.93	0.9179	0.9423	0.9425

Comparison with the existing methods

We compared the performance of our ensemble models with two existing methods that have previously attempted to classify pancreatic ductal adenocarcinoma (PDAC) based on blood exosome derived transcriptomics. Wang et al., which utilizes a SVM model to classify PDAC patients based on the expression of an eight gene biomarker signature, managed to achieve an AUC of 0.936 on their external validation dataset [11]. A method by Wang et al., identified a four lncRNA signature that could classify PDAC patients with an AUC of 0.848 [25]. Our ensemble models, which integrated predictions from multiple large language models with alignment-based methods, consistently outperformed both the tools. The detailed comparison of the existing methods with our approach is reported in Table 8. The performance of the existing tools represents the values that were reported in their respective studies. Notably, the AUC values for our models were significantly higher, indicating superior discriminative ability in classifying PDAC patients. This comparison highlights the advantages of our approach, which leverages the strengths of both LLMs and alignment techniques, ultimately providing a more robust and accurate tool for clinical applications in PDAC detection and management.

Table 8 - Comparison of performance reported by existing methods with our method for peptide lengths 5, 10 and 15

	Number of gene in biomarker panel	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**	**MCC**	**F1-score**	**AUC**
Wang et al. 2020	7	0.9368	0.9157	-	0.9269	-	-	0.936
Wang et al. 2024	4	0.72	0.89	-	-	-	-	0.8476
ESM2-t6-8M + MERCI	5	0.9643	0.8222	0.871	0.901	0.8029	0.9153	0.9464
PeptideBERT + MERCI	10	0.9107	0.8667	0.8947	0.8911	0.7792	0.9027	0.9484

ESM2-t6-8M + MERCI	15	0.8929	0.8444	0.8772	0.8713	0.7391	0.885	0.9492

DISCUSSION

This study marks a significant milestone in the application of artificial intelligence (AI) and machine learning (ML) for pancreatic ductal adenocarcinoma (PDAC) diagnosis by introducing large language models (LLMs) as a novel methodology for leveraging transcriptomic data. By transforming gene expression profiles into pseudo-peptides and training a PeptideBERT model, we achieved a robust classification performance with an AUC of 0.949. This outcome underscores the transformative potential of LLMs in bioinformatics, particularly in translating complex transcriptomic data into interpretable and actionable insights for disease prediction.

The unique ability of LLMs to process and model sequential data positions them as a powerful tool for analyzing biological sequences. In this study, the transformation of gene expression data into a textual format proved pivotal, enabling the PeptideBERT model to identify meaningful patterns within the data. This approach not only demonstrated comparable or superior performance to traditional methodologies but also highlighted the scalability of LLMs in analyzing high-dimensional biological datasets.

Moreover, this study broadens the horizon of LLM applications in precision oncology, providing a robust framework for integrating diverse types of omics data for diagnostic purposes. The versatility of LLMs in handling complex sequential data, as demonstrated here, represents a paradigm shift in how transcriptomic data can be used for early and accurate disease detection. The results of this study establish a strong foundation for future exploration of LLMs in precision diagnostics, including their potential for analyzing data beyond transcriptomics, such as proteomics or metabolomics.

However, there are still certain limitations of the study, that can be addressed in the future. In the process of generating the peptide sequences, the position of the amino acids is determined by the probabilities and the peptides generated from them do not represent real world peptides. This

could mislead pattern recognition in LLMs which are trained on real peptides. The length of the generated peptides in our study are limited to 50, whereas it may be possible that longer peptide may report better accuracies in identifying PDAC samples.

CONCLUSION

While this study highlights the potential of large language models (LLMs) in advancing precision diagnostics for pancreatic ductal adenocarcinoma (PDAC), challenges such as computational complexity and resource demands must be addressed for clinical implementation. Optimizing LLMs for efficiency and scalability is crucial to their adoption in routine diagnostics. Additionally, as LLMs evolve, their integration with multi-omics data and predictive modelling for treatment response offers significant promise. These advancements could transform PDAC diagnosis and management, paving the way for more accurate, comprehensive, and personalized healthcare solutions.

FUNDING SOURCE

The current work has been supported by the Department of Biotechnology (DBT) grant BT/PR40158/BTIS/137/24/2021.

CONFLICT OF INTEREST

The authors declare no competing financial or non-financial interests.

ACKNOWLEDGEMENTS

Authors are thankful to the Department of Science and Technology (DST-INSPIRE) and Indraprastha Institute of Information Technology, New Delhi, for fellowships and financial support. Authors are also thankful to Department of Computational Biology, IIITD New Delhi for infrastructure and facilities.

AUTHORS' CONTRIBUTIONS

SC collected and processed the datasets. SC and NKM implemented the algorithms and developed the prediction models. SC, NKM and GPSR analyzed the results. SC, NKM and

GPSR penned the manuscript. GPSR conceived and coordinated the project. All authors have read and approved the final manuscript.

REFERENCES

1. Kleeff J, Korc M, Apte M, et al. Pancreatic cancer. Nat. Rev. Dis. Primers 2016; 2:16022

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J. Clin. 2018; 68:7–30

3. Calle EE, Rodriguez C, Walker-Thurmond K, et al. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N. Engl. J. Med. 2003; 348:1625–1638

4. Font-Burgada J, Sun B, Karin M. Obesity and Cancer: The Oil that Feeds the Flame. Cell Metab. 2016; 23:48–62

5. Rahn S, Zimmermann V, Viol F, et al. Diabetes as risk factor for pancreatic cancer: Hyperglycemia promotes epithelial-mesenchymal-transition and stem cell properties in pancreatic ductal epithelial cells. Cancer Lett. 2018; 415:129–150

6. Gillen S, Schuster T, Meyer Zum Büschenfelde C, et al. Preoperative/neoadjuvant therapy in pancreatic cancer: a systematic review and meta-analysis of response and resection percentages. PLoS Med. 2010; 7:e1000267

7. Werner J, Combs SE, Springfeld C, et al. Advanced-stage pancreatic cancer: therapy options. Nat. Rev. Clin. Oncol. 2013; 10:323–333

8. Poruk KE, Gay DZ, Brown K, et al. The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. Curr. Mol. Med. 2013; 13:340–351

9. Melo SA, Luecke LB, Kahlert C, et al. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. Nature 2015; 523:177–182

10. Yang KS, Im H, Hong S, et al. Multiparametric plasma EV profiling facilitates diagnosis of pancreatic malignancy. Sci. Transl. Med. 2017; 9:

11. Yu S, Li Y, Liao Z, et al. Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. Gut 2020; 69:540–550

12. Arora A, Patiyal S, Sharma N, et al. A random forest model for predicting exosomal proteins using evolutionary information and motifs. Proteomics 2024; 24:e2300231

13. Choudhury S, Bajiya N, Patiyal S, et al. MRSLpred-a hybrid approach for predicting multi-label subcellular localization of mRNA at the genome scale. Front. Bioinform. 2024; 4:1341479
14. Nakamura K, Zhu Z, Roy S, et al. An exosome-based transcriptomic signature for noninvasive, early detection of patients with pancreatic ductal adenocarcinoma: A multicenter cohort study. Gastroenterology 2022; 163:1252-1266.e2
15. Bhalla S, Kaur H, Kaur R, et al. Expression based biomarkers and models to classify early and late-stage samples of Papillary Thyroid Carcinoma. PLoS One 2020; 15:e0231629
16. Arora C, Kaur D, Raghava GPS. Universal and cross-cancer prognostic biomarkers for predicting survival risk of cancer patients from expression profile of apoptotic pathway genes. Proteomics 2022; 22:e2000311
17. Zhao Z, Anand R, Wang M. Maximum Relevance and Minimum Redundancy feature selection methods for a marketing machine learning platform. arXiv [stat.ML] 2019;
18. Mitchell MW. Bias of the random forest out-of-bag (OOB) error for certain input parameters. Open J. Stat. 2011; 01:205–211
19. Kukreja SL, Löfberg J, Brenner MJ. A least absolute shrinkage and selection operator (lasso) for nonlinear system identification. IFAC Proc. Vol. 2006; 39:814–819
20. Guntuboina C, Das A, Mollaei P, et al. PeptideBERT: A language model based on transformers for peptide property prediction. J. Phys. Chem. Lett. 2023; 14:10427–10434
21. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 2022; 44:7112–7127
22. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic level protein structure with a language model. bioRxiv 2022;
23. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009; 10:421
24. Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics 2011; 27:1231–1238
25. He X, Chen L, Di Y, et al. Plasma-derived exosomal long noncoding RNAs of pancreatic cancer patients as novel blood-based biomarkers of disease. BMC Cancer 2024; 24:961