H2020-SC1-FA-DTS-2020-1
AI for Genomics and Personalized Medicine

PANCAIM

Pancreatic cancer AI for genomics and personalized Medicine

Starting date of the project: 01/01/2021
Duration: 48 months

= Deliverable D4.2 =

Stability of the established and new PDAC subtypes quantified

Due date of deliverable: 30/06/2024
Actual submission date: 30/06/2024

Responsible WP: Tero Aittokallio, WP4, Oslo University Hospital
Responsible TL: Tero Aittokallio, Oslo University Hospital
Revision: V1.0

Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016851.

AUTHOR

Author	Institution	Contact (e-mail, phone)
John Zobolas	OUS	ioannisz@uio.no
Tero Aittokallio	OUS	t.a.aittokallio@medisin.uio.no
Alberto López Sánchez	OUS	a.l.sanchez@medisin.uio.no

DOCUMENT CONTROL

Document version	Date	Change
V0.1	20/06/2024	First draft
V0.2	25/06/2024	Review by project manager
V1.0	01/07/2024	Final version for submission

VALIDATION

Reviewers		Validation date
Work Package Leader	Tero Aittokallio	20/06/2024
Project Manager	Kristin Aldag	25/06/2024
Coordinator	Henkjan Huisman	01/07/2024

DOCUMENT DATA

Keywords	Pancreatic cancer; explainable AI; deep learning; clustering; unsupervised learning; multi-omics
Point of Contact	Name: Alberto López Sánchez Partner: OUS Address: Ullernchausseen 70, OUS, Radiumhospitalet, 0379, Oslo E-mail: a.l.sanchez@medisin.uio.no
Delivery date	01/07/2024

DISTRIBUTION LIST

Date	Version	Recipients
24/06/2024	V0.1	Coordinator, project manager
25/06/2024	V0.2	WP and task leaders
01/07/2024	V1.0	European Commission via portal, Partners via OwnCloud, public via project website (pancaim.eu) and Zenodo

DISCLAIMER

Any dissemination of results reflects only the authors' view and the European Commission Horizon 2020 is not responsible for any use that may be made of the information Deliverable D4.2 contains.

Executive Summary

This document details the review of established pancreatic ductal adenocarcinoma (PDAC) subtypes and the development, implementation and beta-testing of an artificial intelligence (AI) multi-modal deep learning (DL)-based clustering approach for stratifying PDAC patients. A review of state-of-the-art PDAC subtyping methods is presented, evaluating and comparing their conclusions and methodologies. New patient stratification findings are presented based on analyses on the publicly-available TCGA-PDAC dataset [1], assessing cluster stability and validating the methodology with an external PDAC dataset. Key innovations include an unsupervised multi-modal DL algorithm for patient stratification and a new DL methodology to analyze the importance of various data modalities. The report highlights the clinical relevance of the identified patient clusters, alongside the application of explainable AI (XAI) techniques for a better understanding of the disease biology. The new methodologies have been implemented as open-source code and are freely accessible on GitHub. In summary, the report illustrates the progress made in developing clustering algorithms to stratify patients and assist translational researchers and clinicians in the effective management of pancreatic cancer patients.

Table of Contents

List of figures

List of tables

Abbreviations

PDAC	Pancreatic Ductal Adenocarcinoma
XAI	Explainable Artificial Intelligence
NMF	Non-negative matrix factorization
SNF	Similarity Network Fusion
CNA	Copy Number Aberrations
CPTAC	Clinical Proteomic Tumor Analysis Consortium
HC	Hierarchical Clustering
MAD	Mean Absolute Deviation
DL	Deep Learning
DAE	Deep AutoEncoder
MAE	Mean Absolute Error
AI	Artificial Intelligence
CNV	Copy Number Variation
AMI	Adjusted Mutual Information

1. Introduction

The Deliverable 4.2 directly relates to WP4 Task 4.2. The Task 4.2 aims to develop a data-driven PDAC patient classification by integrating data from multiple experiments (data modalities), and the study of the information provided by the different data sources. It also includes the evaluation of established PDAC subgroups and assessing the reproducibility of the new PDAC subtypes using various stability metrics, such as those derived from data resampling and cross-validation. This will ensure that the clustering structure is not unique to the specific datasets used. The methodology is validated on an external PDAC dataset. The identification of novel multi-factorial subtypes revealed distinct biological characteristics and potential subtype-specific therapeutic vulnerabilities, which will be crucial for selecting the most effective treatments for the PDAC patients.

2. Results and Discussion

2.1. Review of clustering models for PDAC patient stratification

Pancreatic cancer patient stratification is critical due to the heterogeneous nature of the disease, which impacts treatment efficacy and patient outcomes. Stratifying patients based on molecular and clinical characteristics allows for the identification of distinct subgroups with unique biological behaviors and treatment responses. This precision approach facilitates the development of tailored therapies, improves prognostic predictions, and enhances the ability to identify patients who may benefit from specific interventions. Therefore, by understanding the diverse pathways and mechanisms driving pancreatic cancer, stratification enables more effective and personalized treatment strategies, ultimately aiming to improve survival rates and quality of life for patients.

Table 1 summarizes the state-of-the-art in pancreatic cancer patient stratification at the time of writing this report (June 2024). The table includes various studies that have employed different datasets and methodologies to find pancreatic cancer subtypes. It highlights the diversity in data sources, clustering techniques, number of identified subtypes and clinical relevance of the clusters. From this table, we can conclude that **gene expression is the most common modality used for patient stratification**, both in single-omics and multi-omics studies; in recent years, there has been a higher interest in multi-omics analysis, although that is usually associated to a lower sample size; NMF and Consensus Clustering, both applied individually or together, are the most common techniques to cluster patient samples; two is the most common number of subtypes, in particular, when using the TCGA pancreatic cohort dataset; clusters are not always associated to the survival outcome, especially in unsupervised analyses where the survival information is not used to form the groups.

Table 1: Overview of previous pancreatic cancer patient stratification studies, sorted by publication year

Dataset	Patients	Biological sources	Multi-modal	Method	Number of subtypes	Log-rank test*	Year/ Publication
GSE17891	27	Gene expression (microarray)	No	NMF + Consensus Clustering	3	0.038	2011 / [2]
GSE71729	147	Gene expression (microarray)	No	NMF + Consensus Clustering	2	0.007	2015 / [3]
EGAS00001000154	93	Gene expression (RNA-seq)	No	NMF	4	0.0302	2016 / [4]
TCGA	184	DNA methylation	No	NMF	3	Not provided	2017 / [5]
TCGA	76	Gene expression (RNA-seq + miRNA + lncRNA); DNA methylation	Yes	SNF	2	Not provided	2017 / [6]
Not provided	288	Gene expression (microarray)	No	Consensus clustering	5	4×10^{-9}	2018 / [7]
TCGA	164	Gene expression (RNA-seq)	No	HC	2	0.05	2020 / [8]
TCGA	45	Protein (RPPA);	Yes	SNF	2	0.18	2020 / [9]

		gene expression (RNA-seq + miRNA); DNA methylation					
TCGA	146	Gene expression (RNA-seq + miRNA); DNA methylation; clinical	Yes	Deep learning (autoencoder) + K-Means	2	$1x10^{-6}$ (supervised) / 0.005 (unsupervised)	2021 / [10]
CPTAC	105	DNA (CNA); gene expression (RNA-seq); protein (expression + phosphorylation + glycosylation sites abundances)	Yes	NMF	2	$1.2x10^{-5}$	2021 / [11]
IPX0002796002	217	Protein (expression)	No	Consensus clustering	3	0.01	2022 / [12]
TCGA	160	Gene expression (RNA-seq + miRNA + lncRNA); DNA methylation + DNA (mutation); clinical	Yes	Ensemble of 10 clustering algorithms	2	< 0.001 (supervised)	2023 / [13]

**The log-rank test is a non-parametric statistical test used to determine whether there are significant differences in the survival distributions of two or more groups of patients (here, the clusters of PDAC patients).*
SNF: Similarity Network Fusion; HC: Hierarchical Clustering; NMF: non-negative matrix factorization; CNA: Copy Number Aberrations; CPTAC: Clinical Proteomic Tumor Analysis Consortium.

As shown in Table 1, NMF is the most commonly-applied approach for stratifying patients. However, in recent years, DL-based clustering has been shown to outperform traditional clustering algorithms in various bioinformatics applications, including bioimaging, cancer genomics and biomedical text mining [14]. Previous works have predominantly used architectures like autoencoders or variational autoencoders, typically as two-step models (training the DL component separately from the clustering step) or applied only to single-omics data [15–17]. Yet, optimizing a joint loss function that integrates both the autoencoder and the clustering objectives has outperformed previous methods in tasks such as cancer category recognition, survival analysis, and clinical parameter enrichment in the pan-cancer TCGA dataset [18]. Its efficacy has also been demonstrated in single-cell multi-omics data [19]. Consequently, DL-based clustering approaches hold significant promise for future clinical applications in biomedical data analysis.

Lautizi *et al.* systematically evaluated the established PDAC subtypes by Collisson et al. [2], Moffit et al. [3], Bailey et al. [4] and Puleo et al. [7], across nine publicly available gene expression datasets [20]. The clustering analysis showed inconsistencies in subtype identification across different datasets, and in some instances, it revealed a different number of PDAC subgroups than initially reported, questioning the true number of the PDAC subtypes. Next, they developed sixteen classification models to assess the predictive capability of these signatures for tumor subtypes. The classification accuracy varied significantly, ranging from approximately 35% to 90%, indicating **instability of the signatures**. Furthermore, permuted subtypes and random gene sets yielded similar performance levels. This study revealed significant limitations and inconsistencies stemming from technical biases in sample preparation and tumor purity, indicating that PDAC molecular signatures lack generalizability across datasets. This emphasizes the difficulties of using transcriptome data for PDAC subtyping and the need for more robust biomarker signatures.

Other PANCAIM partners also examined the stability of the established PDAC subtypes in the Deliverable 3.4, confirming the limits of the current established subtypes. To overcome this situation, a new transcriptomics-based Consensus Classification System (**PDAConsensus**) was developed to integrate these subtypes, providing a more comprehensive understanding of PDAC molecular subtypes. Additionally, a deep learning tool (**CLAM-Kalimuthu**) was developed to automatically classify haematoxylin-eosin (H&E) staining whole-slide images into their corresponding Kalimuthu classification [21,22].

Yet, a multi-omics approach is crucial for deriving more robust and clinically valuable insights. By integrating multiple omics data within the same cohort, we can obtain more detailed information. This integrative analysis reduces the impact of experimental and biological noise, identifies various aspects within the same molecular layer (such as mutations and copy number variations in DNA), and enhances our understanding of different levels of biological organization.

Therefore, this report focuses on finding novel PDAC subtypes using multi-omics data. We developed a deep clustering model using multi-omics data to identify subgroups of PDAC patients by optimizing both the autoencoder and the clustering metrics. RNA-seq and DNA methylation omic profiles were selected as these were the most commonly used across the published methods (Table 1). Cluster stability is assessed through data resampling techniques and stability metrics. Our approach also leverages the capabilities of NMF for feature selection. We compared our method with several others listed in Table 1, including two early integration techniques (K-Means and HC), two NMF variants (jNMF and intNMF), and SNF. The model development and results are shared in https://github.com/albertolzs/edc_mo_pdac. To further validate our methodology, we used an external dataset [23], whose results are publicly available in https://github.com/albertolzs/edc_mo_pdac_val.

2.2. A DL-clustering model applied to the TCGA PDAC cohort

Dataset: We used PAAD-TCGA data from the Firehose Broad GDAC using the R packages curatedTCGAData and TCGAutils [24]. Patients with PDAC as primary tumor and having both RNA-seq data (Illumina HiSeq, upper quartile normalized RSEM TPM gene expression values) and DNA methylation data (Illumina Human Methylation 450) were selected. In total, we had 147 patients with both omics profiles.

Preprocessing: High-dimensional data, especially when combined with small patient cohorts, which is the typical case in omics profiling, limits the learning capacity of machine learning models. Therefore, it is usually recommended to reduce the data dimensionality before the clustering. The omics data were preprocessed using a Scikit-learn pipeline [25]. Similar to previous studies [10,26,27], the steps included:

- RNA-seq data:
  - Initially consisted of 20,501 features (genes).
  - Removing features with zero values in more than 20% of patients.
  - Retaining 50% of the most variable features using mean absolute deviation (MAD).
  - Removing features with a Pearson correlation higher than 0.85.
  - Applying log2 transformation.
  - Performing a NMF-based feature selection.
  - Normalizing the data using z-score normalization.
- DNA methylation data:
  - Initially consisted of 485,577 features (CpGs).
  - Excluding features from sexual chromosomes.
  - Selecting only CpGs with gene symbols in the array.
  - Filtering out features with missing values in more than 20% of patients.
  - Retaining 10% of the most variable features using MAD.
  - Imputing missing values using average values.
  - Performing a NMF-based feature selection.
  - Normalizing the data using z-score normalization.

Feature selection: For the NMF-based feature selection, we adopted a strategy similar to Carmona-Saez *et al. [28]*, which is reported as an unsupervised method to identify important features [29,30]. NMF aims to find two non-negative matrices whose product approximates the original matrix. The resultant weight matrix, with dimensions equal to the number of components by the number of final features, is composed of vectors called basis components. We ranked the features for each NMF basis component in descending order of significance and selected the most influential features. The number of features from each basis component was determined through a hyperparameter optimization. To determine the optimal number of NMF components, we experimented with various numbers (8, 16, 32, 64, 128, 256, and 512) and computed the reconstruction error using beta-divergence.

While selecting more components reduces the error, it also increases overfitting and computational demands. Based on some preliminary hyperparameter optimization results, which aimed to identify the best set of features, we ultimately used 512 components in our analysis.

The two data matrices were used as input for a deep clustering algorithm that combines a multi-modal deep autoencoder (DAE) and K-Means. A DAE is an artificial neural network used for feature extraction, reducing the dimensionality of the input data in a non-linear manner by mapping it into a hidden representation. The encoder layers create this hidden representation (or embedding), while the decoder layers attempt to reconstruct the original input. The encoder branches were concatenated before the latent representation, following a middle integration approach that maps multi-omics data to a joint latent representation. Each block consisted of a linear layer, PReLU activation, and a batch normalization layer. We used a joint loss function, which is the weighted sum of the reconstruction error (mean absolute error, MAE) of the DAE and the sum of squared distances of the samples to their closest cluster center (inertia):

$$L_{ae} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \hat{x}_i|$$

$$L_{cl} = \frac{1}{n} \sum_{\substack{i=1 \\ x_i \in cluster_j}}^{n} (x_i - \mu_j)^2$$

$$L = L_{ae} + \lambda L_{cl}$$

where:
n is the number of samples,
x is the input sample,
x̂ is the reconstructed sample,
μ is the cluster center that contains the sample,
λ is the coefficient that controls the trade-off between the autoencoder and clustering loss function.

Both the multi-modal deep autoencoder and the cluster centers were initialized with prior independent pretraining. This pretraining step is a common practice with small datasets [19] due to K-Means sensitivity to initialization. The algorithm workflow is illustrated in Figure 1.

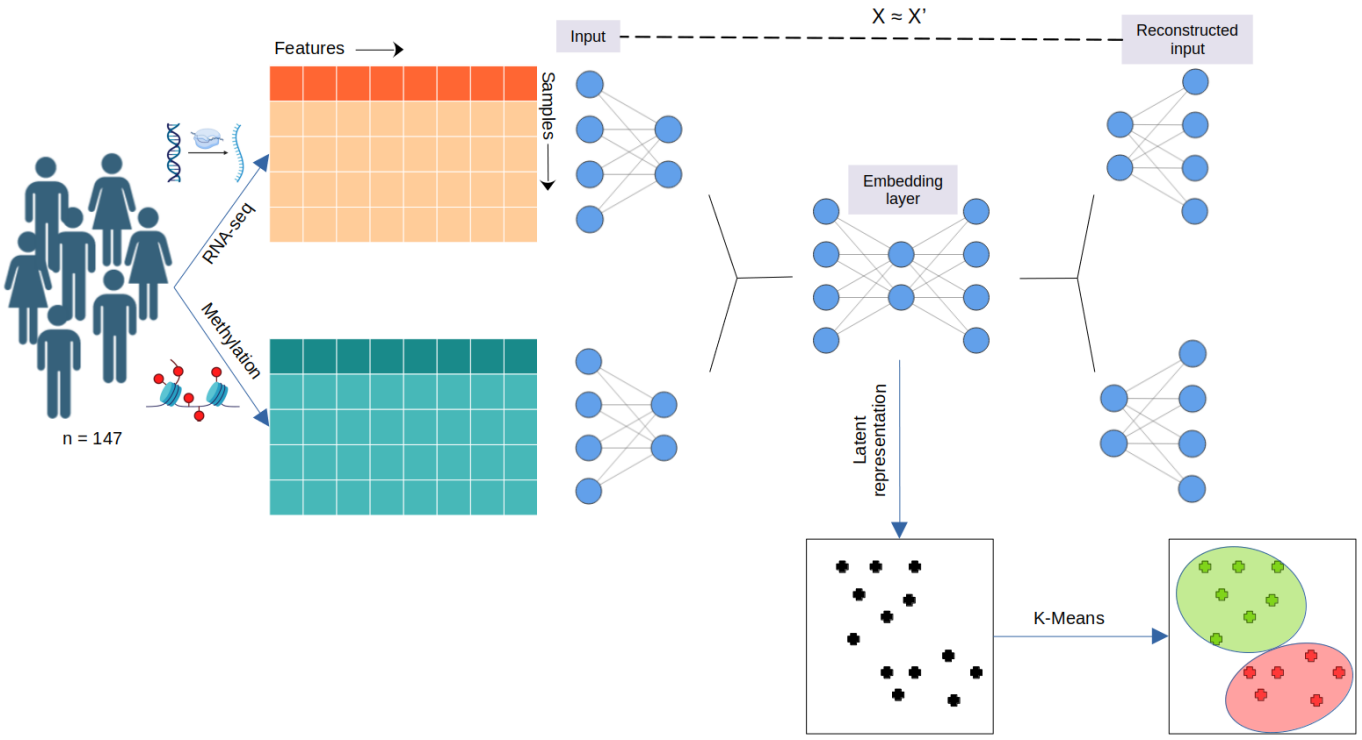

Figure 1: Deep clustering algorithm architecture

Hyperparameter optimization: To optimize the model performance and ensure it accurately captures underlying data patterns, we performed hyperparameter tuning using the Silhouette score as the objective function. The Silhouette score, a widely used clustering metric, is calculated as the average of the mean intra-cluster distance and the mean nearest-cluster distance for each sample [31]. We utilized the Tree-structured Parzen Estimator (a Bayesian optimization method) algorithm to find the best hyperparameters [32]. We employed a 5-nested 5-fold cross-validation strategy to evaluate and ensure the robustness of the clusters. During this process, the training data was employed for model training, the validation data for hyperparameter selection and model evaluation, and the testing dataset only for assessing the model performance and stability on an independent dataset. To prevent data leakage, both the preprocessing pipeline and the clustering model were applied in each iteration. The optimized hyperparameters included: the number of hidden layers, the number of neurons, latent space dimensionality, number of input features, epochs for both pretraining and training, lambda coefficient for the joint loss function, and the number of clusters. The learning rate was automatically set using a learning rate finder strategy [33].

Figure 2 illustrates the model loss function and the Silhouette score used as an independent evaluation metric for the clustering solution. As there is no baseline for clustering due to the absence of ground truth, for each metric, we utilized the results from the hyperparameter optimization process to establish comparison baselines (dotted red lines): the largest loss function value when the lambda coefficient was within {best lambda coefficient} ± 0.02; and the average Silhouette score obtained during optimization. These results demonstrated that the clustering model outperformed a random clustering solution and was not overfitted to the training data. Despite significant differences (Wilcoxon test) in the model loss function between training and validation data, likely due to the small dataset size and the metric's sensitivity to outliers; the Silhouette score remained consistent across training, validation, and test datasets.

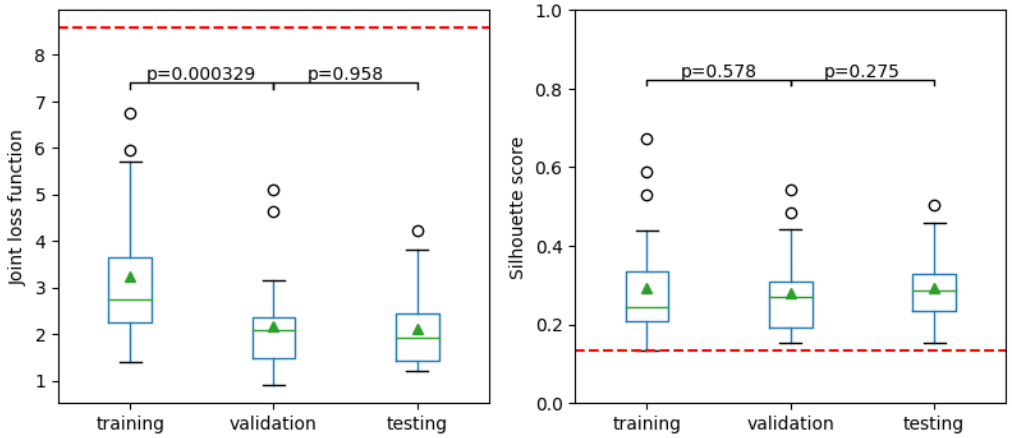

Figure 2: Model loss function (left) and Silhouette score (right) during hyperparameter optimization.
The results were obtained using a nested cross-validation strategy, ensuring the testing set remains completely blind to the training process, thus demonstrating the algorithm generalizability. Baseline scores are shown as red lines

The number of clusters was shown to be the most significant hyperparameter during optimization, as determined by the fANOVA hyperparameter importance evaluation algorithm [34]. Most models in the nested cross-validation achieved the best metrics with two clusters (Figure 3). Table 1 further supports this, showing that two clusters is the most common solution, indicating a robust number of subtypes in the current dataset. When the model was trained on the whole dataset, it identified a cluster with 97 samples, and another one with 50.

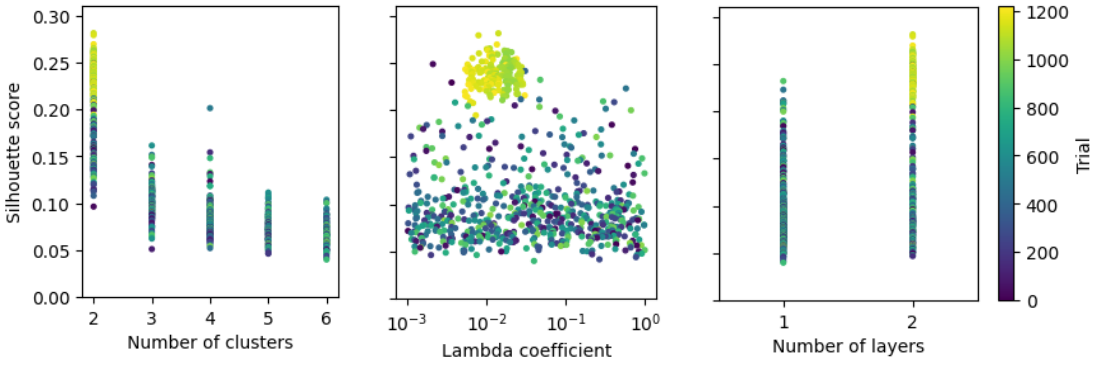

Figure 3: Slice plot scoring of every possible option of the hyperparameters
The x-axis represents the options, and the y-axis the silhouette score; the color corresponds to the number of the trial during the optimization process.

Cluster stability: During the 5-nested 5-fold cross validation, 25 clustering solutions are computed and validated. For each of the 25 clustering solutions, we performed pairwise comparisons among the generated clusters, keeping only the samples present in each pair. This yields a total of 300 comparisons. The metric used was the **adjusted mutual information** (AMI). The mutual information score measures the similarity between two clustering solutions of the same data, while the AMI modifies the mutual information score to account for randomness [35]. We selected AMI for its practical properties. It is unaffected by the specific label values: changing the class or cluster label values does not alter the score. Additionally, AMI is symmetric, meaning that swapping the folds will yield the same score. This is useful for evaluating the consistency of two independent clustering solutions on the same dataset when the true labels are unknown.

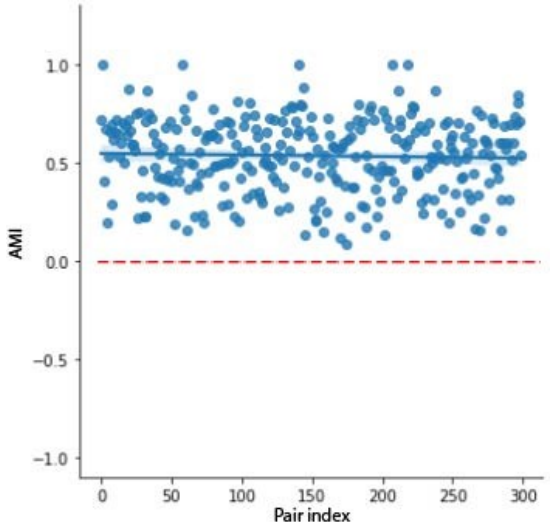

Figure 4: Cluster stability plot representing the AMI scores for pairwise comparisons of clusters generated during nested cross-validation.

Clusters across splits are relatively consistent with respect to the average value of 0.53, with most comparisons resulting in moderate to high AMI scores. Random baseline score is shown as a red line.

The AMI returns a score of 1 when the two partitions are identical, meaning they match perfectly. For random partitions, which are independent labelings, the expected AMI score is around 0 on average. In our analysis, the pairwise AMI had an average value of 0.53 (Figure 4). The cluster stability, as measured by AMI, is relatively consistent, as the scores do not deviate drastically. Overall, the figure indicates that the clusters generated across different splits of the nested cross-validation are generally stable, with most comparisons yielding moderate to high AMI scores. Stable clusters indicate that the clustering results are consistent and robust to variations in the data, a crucial step for ensuring that the clustering method produces meaningful and consistent results.

2.3. Clinical relevance of the identified PDAC clusters

Understanding the clinical relevance of clusters in pancreatic cancer is vital for advancing patient care and treatment strategies. Clustering patients based on molecular characteristics can reveal subgroups with distinct biological behaviors and therapeutic responses, providing insights into disease mechanisms and progression. However, since clusters were identified using unsupervised learning, we cannot directly determine their clinical relevance. Thus, we tested their associations to clinical parameters.

The log-rank test revealed that the patients in these two clusters had significantly different survival probabilities (Figure 5). The Cox proportional hazards model returned a hazard ratio of 1.75 (95% confidence interval: 1.24-2.49), indicating that patients in cluster 0 have a 75% higher risk of adverse outcomes compared to those in cluster 1. This suggests that cluster 0 corresponds to a more aggressive disease subtype.

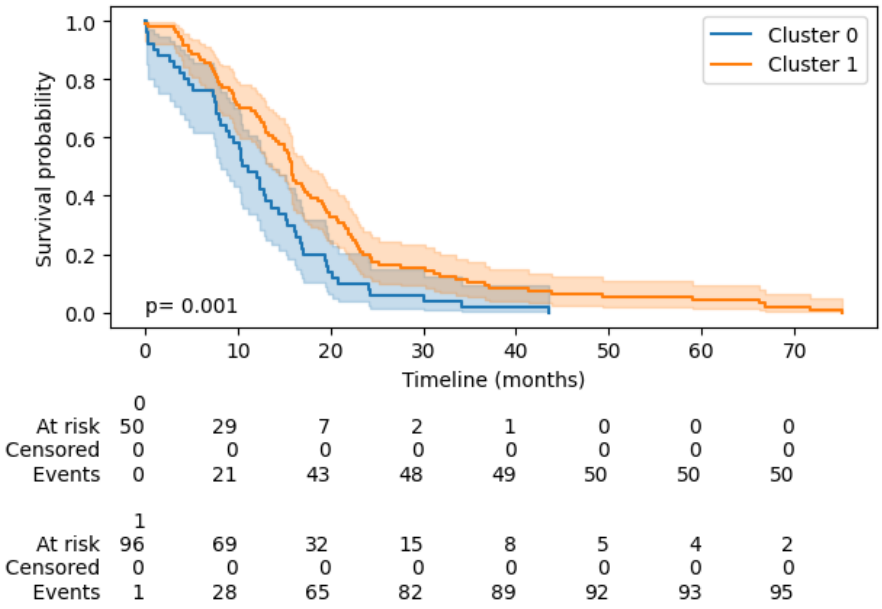

Figure 5: Kaplan-Meier analysis of the survival differences between clusters identified by our DL-clustering method.

Table 2 shows the p-values of testing the association of clusters with clinical parameters. We used the Kruskal-Wallis test on the age at initial diagnosis, as well as the chi-square test on sex and three discrete clinical pathological parameters quantifying the progression of the tumor (tumor stage), cancer in lymph nodes (neoplasm histologic grade) and metastases (metastasis stage). No other significant associations were identified. For reference comparisons, we applied HC, K-Means, and advanced multi-omics clustering algorithms, including SNF [36], intNMF [37], and jNMF [38]. Interestingly, only the clusters identified by the multi-omics clustering algorithms showed an association with sex, whereas those identified by HC and K-Means did not yield any significant associations with other clinical parameters, including overall survival. Furthermore, compared with the other previous methods in Table 1, our algorithm identifies the clusters with the highest association to survival data in TCGA in an unsupervised way. This underscores the impact of our method on patient stratification, demonstrating its capability to identify clusters with significant prognostic differences.

Table 2: p-values of testing the association of clusters with clinical parameters

	Deep clustering	K-Means	HC	SNF	intNMF	jNMF
Overall Survival	**0.001**	0.212	0.469	0.138	0.155	0.165
Diagnosis age	0.920	0.074	0.210	0.250	0.555	0.689
AJCC tumor stage	0.970	0.810	0.597	0.810	0.805	0.906
AJCC metastasis stage	0.457	1.000	1.000	1.000	1.000	0.865
AJCC neoplasm histologic grade	0.876	0.192	0.909	0.159	0.523	0.809
Sex	0.359	0.076	0.432	0.035	0.047	0.023

The lack of significant associations to clinical parameters highlights the challenge of identifying clusters associated with multiple external variables, as most multi-omics models in similar studies in other cancer types found only 1-2 significant associations with clinical parameters across various cancer types [26,39]. This difficulty arises because clustering algorithms aim to uncover novel biological insights rather than being biased by known clinical associations. Overall survival provides a reliable measure, unlike more subjective clinical parameters. Thus, being the overall survival an objective clinical endpoint, it would imply that these clusters are based on novel biological insights and could be valuable for patient stratification in future studies.

2.4. Explainable AI sheds light on PDAC biology

To better understand the biology of pancreatic cancer, we used XAI techniques to unbox how decisions of the black-box deep learning model are made. In medicine, explainability is crucial as it can reveal new insights and shared characteristics among the most important biomarkers or features. By identifying molecular signatures and biological markers associated with specific patient subgroups, biomarker discovery enables precise characterization and classification of individuals based on their unique disease profiles. Therefore, it has the potential to contribute to the recommendation of more personalized treatments.

We firstly employed post-hoc interpretation methods to assign importance scores to each input feature. Attribution methods were utilized to determine the contribution of each input feature to the output, elucidating their impact on predictions for specific neurons. We used various algorithms available in Captum [40]. Specifically, we employed two gradient-based methods and one perturbation-based method: integrated gradients [41], gradient SHAP [42] and feature ablation. After computing attributions for every feature in all samples using these methods, we computed the mean (in absolute values) across the methods and sorted the features accordingly.

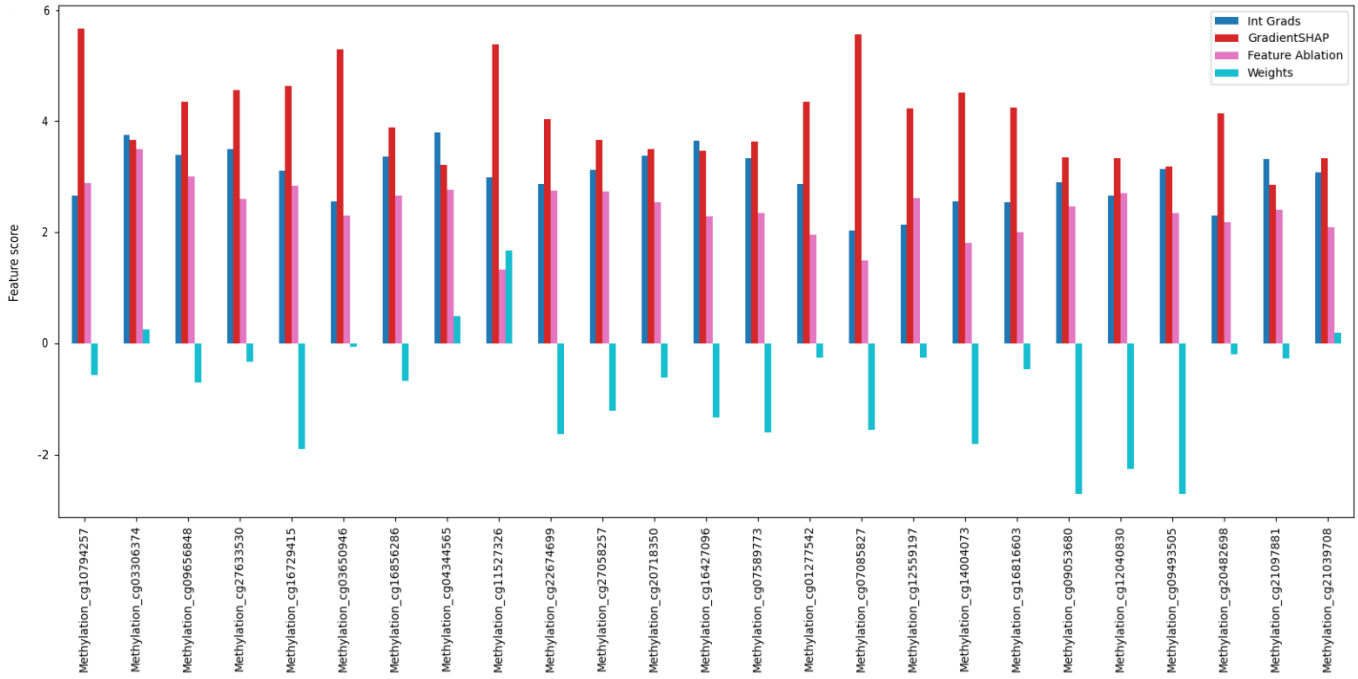

Figure 6: Top-25 input features across feature importance methods (the score is represented as the values of attributions x 10–4 on the y-axis).

The feature importance analysis reveals the crucial role of DNA methylation in unsupervised PDAC patient stratification.

Figure 6 presents a visualization of the feature scores for the top 25 features. The analysis highlights the feature cg10794257, a CpG site associated with Hox genes, which play a crucial role in organogenesis and animal development [43,44]. This gene has been implicated in promoting tumors in pancreatic cancer [5,45]. The second methylation feature, cg03306374, was previously included in a DNA methylation signature for identifying PDAC [46]. A positive correlation was found between sites located in the CpG island at the 5'-end and PDAC-hypermethylated cg03306374. Interestingly, there was no significant difference in the values of the top 25 features between the clusters.

An important observation is that all of the most important features were **methylation** features. To go deeper into this finding, we assessed the contribution of each omics to the final clustering solutions using MM-SHAP. MM-SHAP is a performance-agnostic multi-modal score based on Shapley values that quantifies the extent to which a multi-modal model uses individual modalities [47]. Surprisingly, the relative contribution of each modality across the samples was found to be quite homogeneous, as depicted in Figure 7a. On average, DNA methylation patterns contributed 71% to the final predictions, while the remaining 29% originated from RNA-seq profiles. This

underscores the **crucial role of DNA methylation in PDAC patient stratification**. We further investigated whether there was a difference in the omics relative contributions between the two clusters (Figure 7b). However, the Mann-Whitney U test returned a non-significant p-value. Additionally, no differences were found in the DNA methylation or gene expression value distributions between the two clusters (Figure 7c and 7d).This suggests that it is not the individual features alone that determine the prognosis, but rather the **interaction** between different entities (such as gene expression and methylation features) that is more closely linked to the phenotype. Therefore, the relationships between two or more variables are not merely additive; instead, their combined effects significantly contribute to cancer aggressiveness.

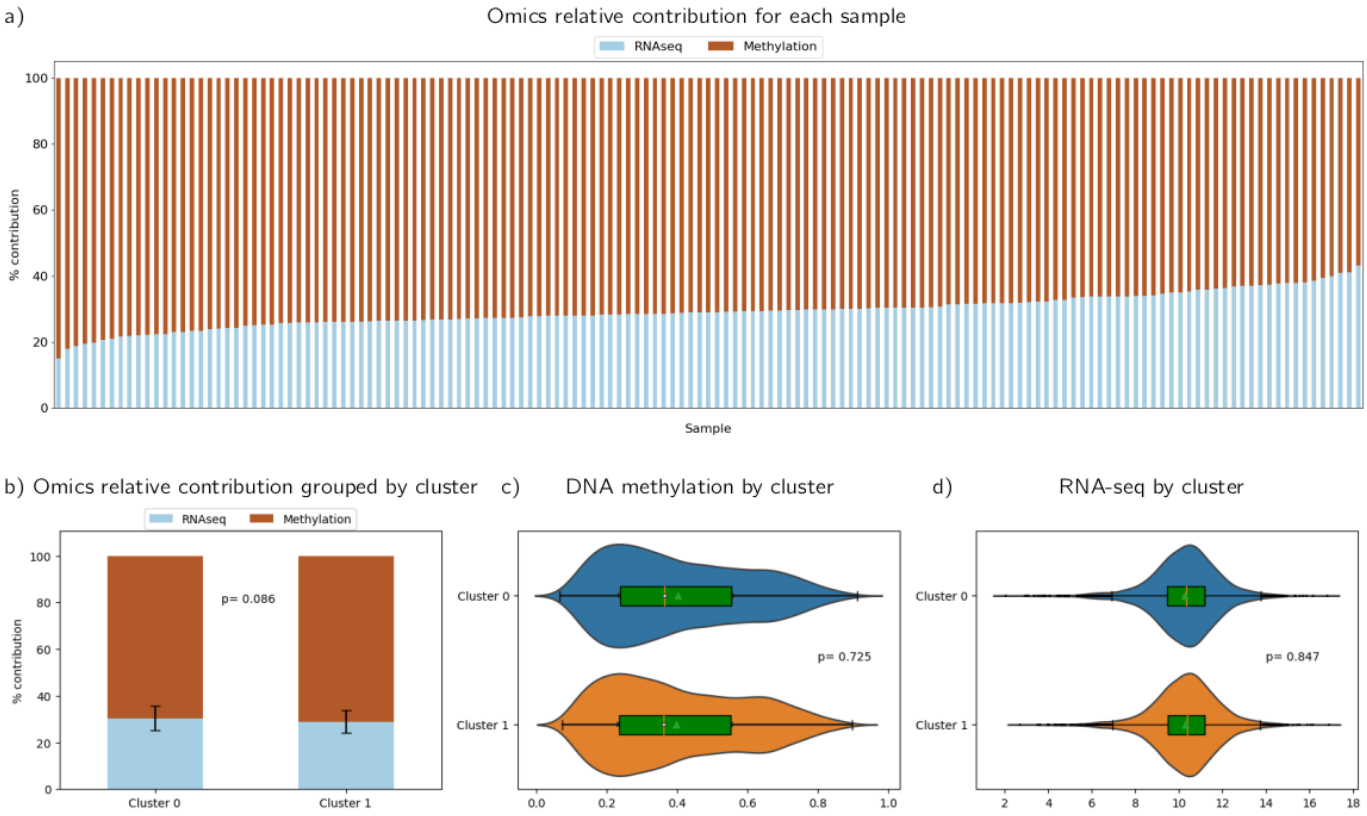

Figure 7: Omics contribution analysis and distribution by cluster

As the neural network functions through the combined activity of various neurons across each layer to generate predictions, we also examined the roles of individual neurons. We first identified the neuron importance in the embedding layer using the layer conductance [48], and then we analyzed the influence of the input features and the relative contribution of each omic using the neuron conductance [49]. The contribution of each data modality on every neuron in the embedding layer is illustrated in Figure 8. On average, methylation data contributes 54%, while RNA-seq accounts for the remaining 46%. However, when considering neuron importance, the disparity increases; for example, methylation influences 63% of the first eight neurons. Generally, the most important neurons were predominantly activated by methylation patterns, whereas less important neurons were more influenced by gene expression. We found that 8 out of the 50 neurons were activated by only a single modality: 6 by methylation and 2 by RNA-seq data.

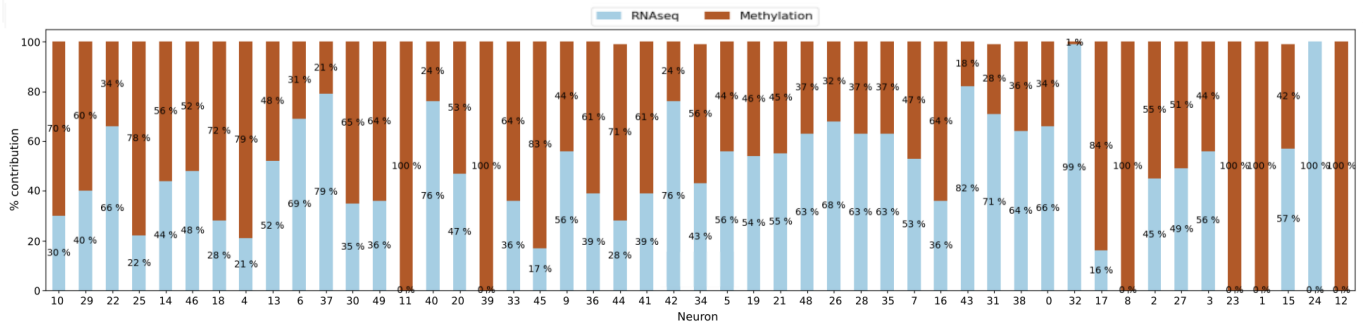

Figure 8: Omics relative contribution to neurons in the embedding layer sorted by importance

We integrated transcriptomics and epigenomics data, the most common combination in multi-omics cancer studies [50]. Overall, transcriptomics is the most frequent omic modality used in multi-omics studies for both cancer and non-cancer diseases. Although gene expression has been considered to provide the highest predictive power, predictive models can be based on different omics profiles and cancer types, as there is no universal best omics data type, and results depend on the cancer type [51]. Interestingly, our findings indicated that DNA methylation patterns were more crucial for our clustering solution. Both attribution methods, for the entire model, and neuron conductance, for specific units in the network, confirmed the significant role of methylation data. Our findings are supported by research studies, where it has been demonstrated that changes in DNA methylation profiles drive tumor progression in PDAC [52,53] and are strongly associated with patient survival [54]. Also a recent study identified DNA methylation as the most predictive omic data type for accurately distinguishing PDAC from chronic pancreatitis among DNA methylation, mRNA, and miRNA data [46].

2.5. Validating the results using an external dataset

To further validate our approach, we used an external dataset. We obtained the dataset from Osipov *et al [23]*. This dataset includes several data modalities: DNA (SNVs, INDELs and CNVs), gene expression (RNA-seq), proteomics (tissue and plasma), lipidomics and pathology features. We integrated four modalities - gene expression (2000 features), pathology (820 features), CNVs (648 features), and SNVs (611 features) - which resulted in a cohort of 57 patients.

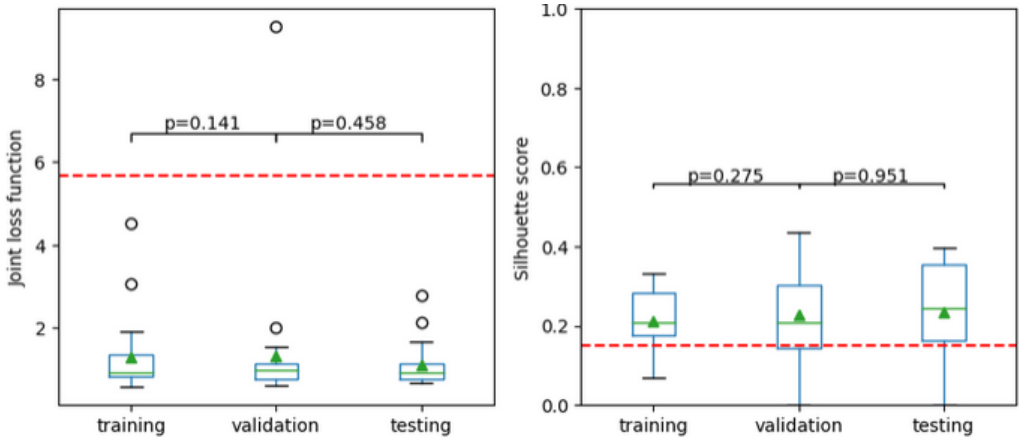

Figure 9: Model loss function (left) and Silhouette score (right) during hyperparameter optimization in an external dataset.

We first validated our algorithm. As this dataset has been already preprocessed with a feature selection, the only preprocessing we applied was a z-score normalization. Same as in the DL-clustering model applied to the TCGA PDAC cohort section, the result of the optimization is shown in Figure 9. Compared to the previous Silhouette score of 0.28, the current score is 0.23, indicating a very similar performance. While the number of clusters remains the most influential hyperparameter, determining the precise number of subtypes is less clear in this dataset, with three clusters showing a slightly higher score (Figure 10).

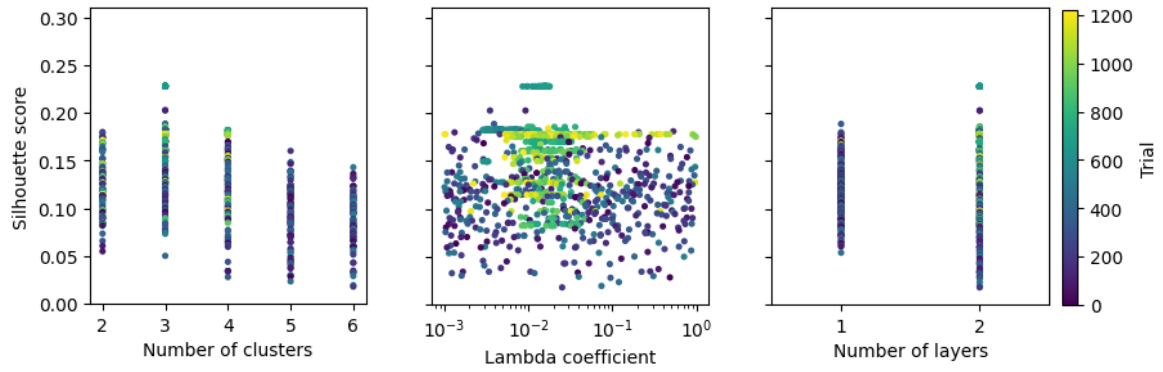

Figure 10: Slice plot scoring of every possible option of the hyperparameters during optimization in an external datase
The x-axis represents the options, and the y-axis the silhouette score; the color corresponds to the number of the trial during the optimization process.

Clusters were significantly associated with survival data, with the log-rank test achieving a similar p-value as in the other dataset. One of the clusters had a much poorer prognosis, as shown in the survival curves in Figure 11. The clusters 0 and 1 had very similar survival curves, however, we decided to keep them separately for a better statistical analysis, as the model optimization suggested.

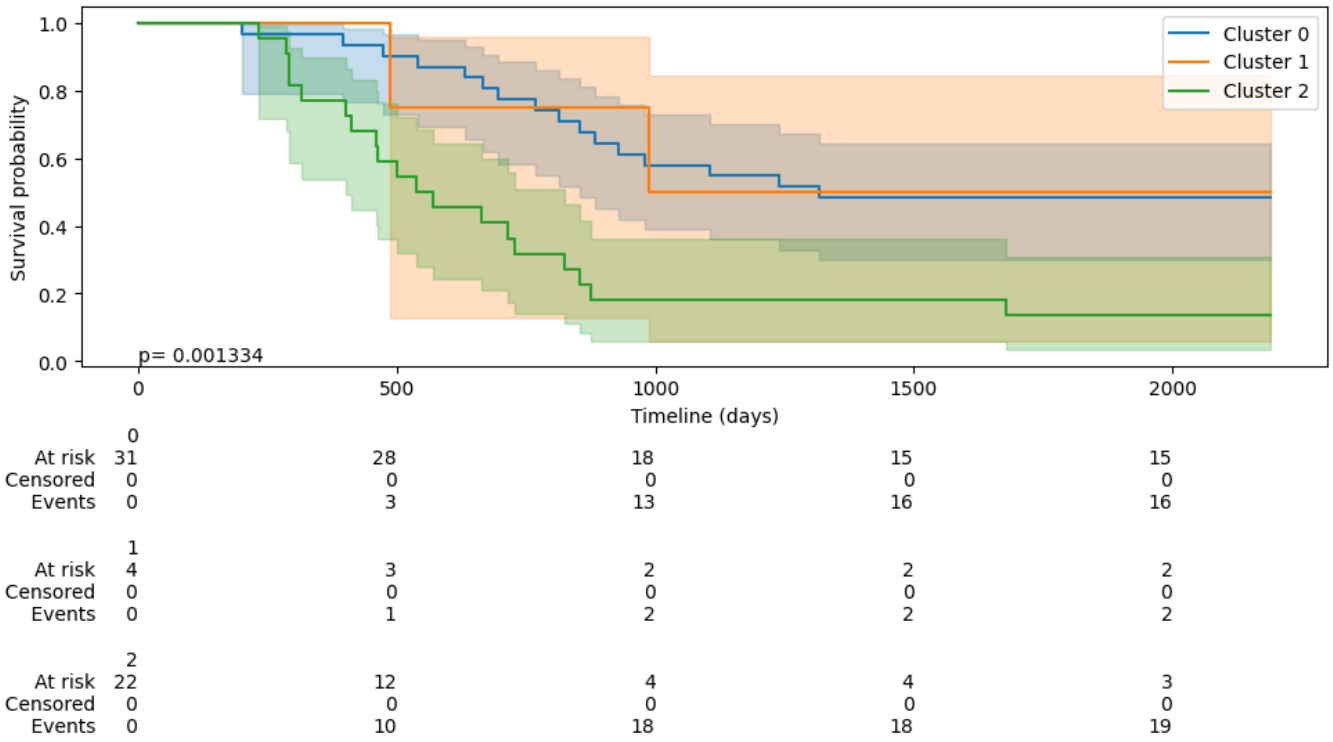

Figure 11: Kaplan-Meier analysis of the survival differences between clusters in the external dataset

As with previous results, the clusters demonstrated a significant association only with overall survival (Table 3). Standard methods like K-Means and HC did not show any significant associations.

Table 3: p-values of testing the association of clusters with clinical parameters.

	Deep clustering	K-Means	HC
Overall Survival	0.001	0.522	0.466
Diagnosis age	0.153	0.842	0.653
TNM Stage	0.804	0.551	0.497
Sex	0.529	0.348	0.356

The multi-omics relative contribution analysis using MM-SHAP showed that the pathology was the most important data modality (Figure 12). RNA-seq obtained a very similar value to our previous result (29% in the TCGA dataset; 33% in the external dataset).

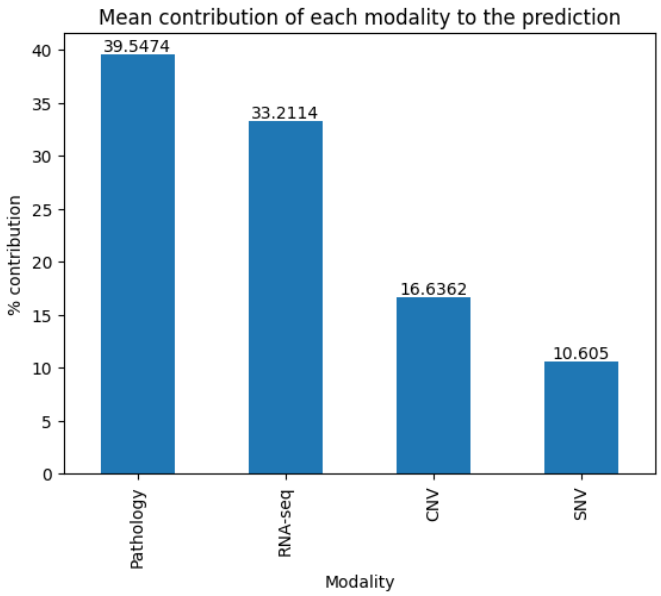

Figure 12: Omics relative contribution. CNV: Copy Number Variation

As a final step, we assessed the importance of the identified potential biomarkers (Figure 8). We selected the top 25 features for the three most important neurons, using only RNA-seq input features since this dataset lacks methylation data. Out of them, only six were present in the external dataset. We then evaluated the significance of these genes in the survival data, both as a combined biomarker panel and individually, as shown in Figure 13. Based on their p-values, three genes seem to be associated with survival when considered as a combined panel, and two when analyzed individually. This could support our methodology and findings as a strategy for discovering biomarkers.

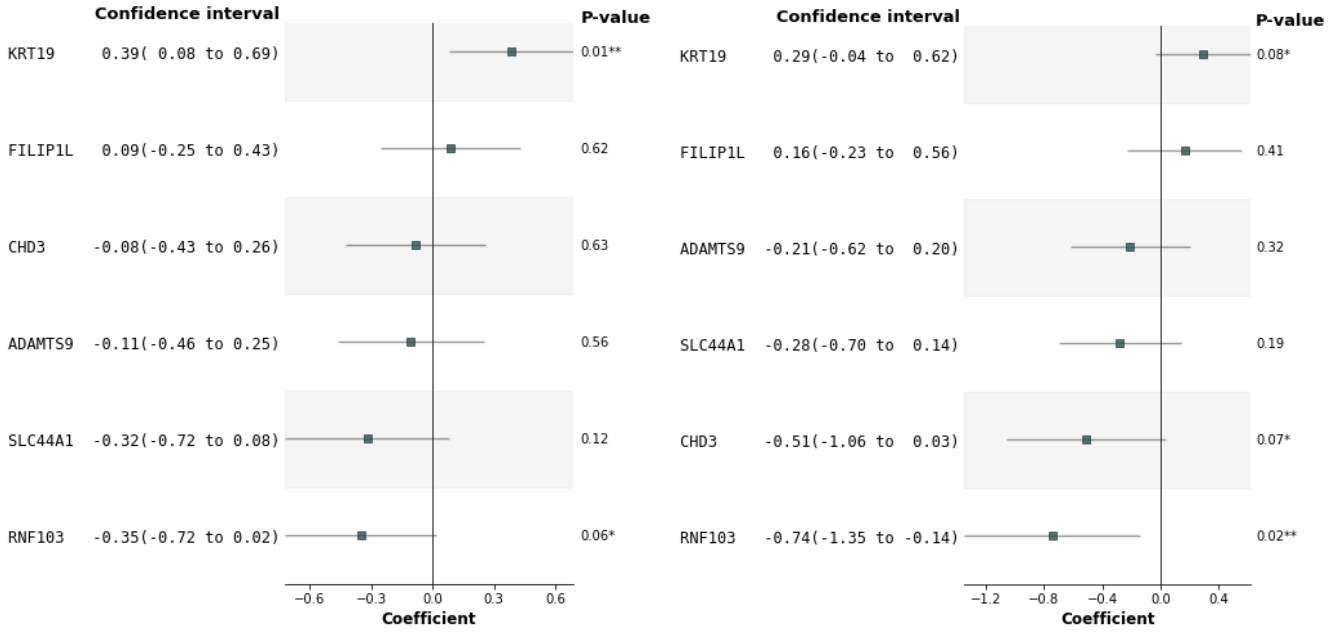

Figure 13: Forest plot of identified biomarkers and their association to patient survival in the external dataset
a) Hazard ratio and p-values using each biomarker individually; b) Hazard ratio and p-values using a combined biomarker panel.

Identifying a robust molecular signature in PDAC is a challenging problem. Previous molecular signatures have produced controversial and inconsistent results across different datasets [20], as discussed in the previous "Review of clustering models for PDAC stratification" section. Despite numerous reports of potential biomarkers, none have yet been translated into clinical practice [55]. Consequently, further research is essential to discover a reliable biomarker panel for stratifying PDAC patients, which would facilitate personalized medicine and potentially lead to more effective treatments for this highly lethal cancer.

3. Conclusions

The report describes the development of a multi-modal deep clustering algorithm for patient stratification and the use of XAI techniques to elucidate disease biology and model decisions [1]. Algorithms, performance metrics, XAI techniques, methodologies, source code, and results are available in online GitHub repositories. Due to the unavailability of PANCAIM project multi-modal data at the time of writing this report, the PAAD TCGA dataset was utilized for the development, application, and testing of the aforementioned methodologies, and an external multi-modal PDAC dataset was used for further validation [48].

We expanded the previous work of other PANCAIM partners (Deliverable 3.4) in analyzing the stability of established PDAC subtypes, and successfully stratified the PAAD pancreatic cohort into two new groups through unsupervised learning. These two subtypes showed a significant association to survival, proving that they were clinically relevant. The subsequent application of XAI techniques allowed us to assess the relative contributions of various omics and identify potential biomarkers. The multi-omics profile analysis revealed an important role of DNA methylation, partially supported by previous experimental studies. This approach was validated using an external dataset, yielding results that support our patient stratification strategy. We hope this study will help to promote more explainable AI in real-world clinical applications, where the knowledge of the decision factors is crucial.

4. Degree of Progress

100%

5. Dissemination Level

The Deliverable 4.2 is public.

6. References

[1] López A, Zobolas J, Nebdal D, Lingjærde OC, Fleischer T, Aittokallio T. Explainable multi-omics deep clustering reveals an important role of DNA methylation in PDAC. Zenodo. 2024. doi:10.5281/zenodo.10635657

[2] Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med. 2011;17: 500–503.

[3] Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet. 2015;47: 1168–1178.

[4] Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;531: 47–52.

[5] Mishra NK, Guda C. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. Oncotarget. 2017;8: 28990–29012.

[6] Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell. 2017;32: 185–203.e13.

[7] Puleo F, Nicolle R, Blum Y, Cros J, Marisa L, Demetter P, et al. Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. Basic and Translational—Pancreas. doi:10.1053/j.gastro.2018.08.033

[8] Jonckheere N, Auwercx J, Hadj Bachir E, Coppin L, Boukrout N, Vincent A, et al. Unsupervised Hierarchical Clustering of Pancreatic Adenocarcinoma Dataset from TCGA Defines a Mucin Expression Profile that Impacts Overall Survival. Cancers . 2020;12: 3309.

[9] Sinkala M, Mulder N, Martin D. Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics. Sci Rep. 2020;10: 1–14.

[10] Ju J, Wismans LV, Mustafa DAM, Reinders MJT, van Eijck CHJ, Stubbs AP, et al. Robust deep learning model for prognostic stratification of pancreatic ductal adenocarcinoma patients. iScience. 2021;24. doi:10.1016/j.isci.2021.103415

[11] Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell. 2021;184: 5031–5052.e26.

[12] Tong Y, Sun M, Chen L, Wang Y, Li Y, Li L, et al. Proteogenomic insights into the biology and treatment of pancreatic ductal adenocarcinoma. J Hematol Oncol. 2022;15: 1–47.

[13] Chen Y, Meng J, Lu X, Li X, Wang C. Clustering analysis revealed the autophagy classification and potential autophagy regulators' sensitivity of pancreatic cancer based on multi-omics data. Cancer Med. 2023;12: 733–746.

[14] Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, et al. Deep learning-based clustering approaches for bioinformatics. Brief Bioinform. 2020;22: 393–415.

[15] Min E, Guo X, Liu Q, Zhang G, Cui J, Long J. A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. [cited 5 Jun 2024]. Available: https://doi.org/10.1109/ACCESS.2018.2855437

[16] Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. doi:10.1158/1078-0432.CCR-17-0853

[17] Hira MT, Razzaque MA, Angione C, Scrivens J, Sawan S, Sarker M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Sci Rep. 2021;11: 6265.

[18] MCluster-VAEs: An end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data. Comput Biol Med. 2022;150: 106085.

[19] Lin X, Tian T, Wei Z, Hakonarson H. Clustering of single-cell multi-omics data with a multimodal deep learning method. Nat Commun. 2022;13: 1–18.

[20] Lautizi M, Baumbach J, Weichert W, Steiger K, List M, Pfarr N, et al. The limits of molecular signatures for pancreatic ductal adenocarcinoma subtyping. NAR Cancer. 2022;4: zcac030.

[21] Kalimuthu SN, Wilson GW, Grant RC, Seto M, O'Kane G, Vajpeyi R, et al. Morphological classification of pancreatic ductal adenocarcinoma that predicts molecular subtypes and correlates with clinical outcome. Gut. 2020;69: 317–328.

[22] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering. 2021;5: 555–570.

[23] Osipov A, Nikolic O, Gertych A, Parker S, Hendifar A, Singh P, et al. The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. Nature Cancer. 2024;5: 299–314.

[24] Ramos M, Geistlinger L, Oh S, Schiffer L, Azhar R, Kodali H, et al. Multiomic Integration of Public Oncology Databases in Bioconductor. JCO Clinical Cancer Informatics. 2020 [cited 3 Jun 2024]. doi:10.1200/CCI.19.00119

[25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

[26] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46: 10546–10562.

[27] Duan R, Gao L, Gao Y, Hu Y, Xu H, Huang M, et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. PLoS Comput Biol. 2021;17: 1–33.

[28] Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A. Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC Bioinformatics. 2006;7: 1–18.

[29] Zhao L, Zhao H, Yan H. Gene expression profiling of 1200 pancreatic ductal adenocarcinoma reveals novel subtypes. BMC Cancer. 2018;18: 1–13.

[30] Liu D, Zhou B, Liu R. An RNA-sequencing-based transcriptome for a significantly prognostic novel driver signature identification in bladder urothelial carcinoma. PeerJ. 2020;8: e9422.

[31] Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20: 53–65.

[32] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. Adv Neural Inf Process Syst. 2011;24. Available: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

[33] Schaul T, Zhang S, LeCun Y. No more pesky learning rates. International Conference on Machine Learning. PMLR; 2013. pp. 343–351.

[34] Hutter F, Hoos H, Leyton-Brown K. An Efficient Approach for Assessing Hyperparameter Importance. International Conference on Machine Learning. PMLR; 2014. pp. 754–762.

[35] Vinh NX, Epps J, Bailey J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. J Mach Learn Res. 2010;11: 2837–2854.

[36] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11: 333–337.

[37] Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS One. 2017;12: e0176278.

[38] Wang H-Q, Zheng C-H, Zhao X-M. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinformatics. 2014;31: 572–580.

[39] Yang B, Yang Y, Wang M, Su X. MRGCN: cancer subtyping with multi-reconstruction graph convolutional network using full and partial multi-omics dataset. Bioinformatics. 2023;39: btad353.

[40] Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: A unified and generic model interpretability library for PyTorch. 2020. Available: http://arxiv.org/abs/2009.07896

[41] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. International Conference on Machine Learning. PMLR; 2017. pp. 3319–3328.

[42] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst. 2017;30. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[43] Zhang L, Silva TC, Young JI, Gomez L, Schmidt MA, Hamilton-Nelson KL, et al. Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex implicates the immune processes in Alzheimer's disease. Nat Commun. 2020;11: 1–13.

[44] Smeester L. A Critical Role for Imprinted Genes in the Placenta in the Developmental Origins of Health and Disease. 2019 [cited 5 Jun 2024]. Available: http://cdr.lib.unc.edu/downloads/pn89db83z

[45] Kuo T-L, Cheng K-H, Chen L-T, Hung W-C. Deciphering The Potential Role of Hox Genes in Pancreatic Cancer. Cancers . 2019;11: 734.

[46] Wu Y, Seufert I, Al-Shaheri FN, Kurilov R, Bauer AS, Manoochehri M, et al. DNA-methylation signature accurately differentiates pancreatic cancer from chronic pancreatitis in tissue and plasma. Gut. 2023;72: 2344–2353.

[47] Parcalabescu L, Frank A. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023. pp. 4032–4059.

[48] Dhamdhere K, Sundararajan M, Yan Q. How Important is a Neuron. International Conference on Learning Representations. 2018. Available: https://openreview.net/pdf?id=SylKoo0cKm

[49] Shrikumar A, Su J, Kundaje A. Computationally Efficient Measures of Internal Neuron Importance. 2018. vailable: http://arxiv.org/abs/1807.09946

[50] A guide to multi-omics data collection and integration for translational medicine. Comput Struct Biotechnol J. 2023;21: 134–149.

[51] Naulaerts S, Menden MP, Ballester PJ. Concise Polygenic Models for Cancer-Specific Identification of Drug-Sensitive Tumors from Their Multi-Omics Profiles. Biomolecules. 2020;10: 963.

[52] Roalsø MTT, Hald ØH, Alexeeva M, Søreide K. Emerging Role of Epigenetic Alterations as Biomarkers and Novel Targets for Treatments in Pancreatic Ductal Adenocarcinoma. Cancers . 2022;14: 546.

[53] Differential methylation landscape of pancreatic ductal adenocarcinoma and its precancerous lesions. Hepatobiliary Pancreat Dis Int. 2020;19: 205–217.

[54] Thompson MJ, Rubbi L, Dawson DW, Donahue TR, Pellegrini M. Pancreatic Cancer Patient Survival Correlates with DNA Methylation of Pancreas Development Genes. PLoS One. 2015;10: e0128814.

[55] Kenner B, Chari ST, Kelsen D, Klimstra DS, Pandol SJ, Rosenthal M, et al. Artificial Intelligence and Early Detection of Pancreatic Cancer: 2020 Summative Review. Pancreas. 2021;50: 251.