Early Pancreatic Cancer detection using Extracellular Vesicles and adaptive learning techniques

Konstadina Kourou, Angeliki Iosif, Ioanna Angelioudaki, Vassiliki Kigka, Alexandros Georgios Tzingounis, Vasiliki Pantazatou, Maria Vlachou, Nikolaos Memos, Agapi Kataki, Manousos M. Konstadoulakis, Dimitrios I. Fotiadis, *Fellow IEEE*

***Abstract*— Pancreatic cancer (PC) is a major global public health problem. Identifying potential biomarkers and individual patient characteristics that contribute to early PC detection is critical in clinical practice. The present study employed a comprehensive data-driven pipeline to address this important issue in a subset of patients diagnosed with Pancreatic Ductal Adenocarcinoma (PDAC) along non-oncologic samples. Putative extracellular vesicle (EV) characteristics in combination with clinical and laboratory features served as potential predictors of patient risk stratification and PDAC diagnosis. The machine learning (ML)-based pipeline entailed the appropriate steps for unbiased model training and validation. The two groups of samples were discriminating against a high degree of accuracy (= 0.96) with balanced sensitivity (=0.95) and specificity (=0.97) rates. Both EV-based variables and biochemical characteristics emerged as important predictors of PDAC diagnosis.**

***Clinical Relevance*— Minimally invasive technologies based on extracellular vesicles (EVs) along with adaptive learning methodologies could provide new directions and solutions with increased efficiency in PC clinical diagnosis.**

I. Introduction

Pancreatic cancer (PC) is a major global public health issue. It has been associated with a high incidence rate (approximately 510,992 incident cases) with 4.7 per 100,000 Age-Standardized Rate (ASR) in the world standard population [1]. Due to late symptom onset and the high proportion of patients having metastatic disease at diagnosis, PC has generally poor prognosis with an overall 5-year survival rate of around 10% [2]. According to the Global Cancer Observatory (GLOBOCAN) 2020 data, PC is the sixth leading cause of cancer-related deaths, owing to the high death rate of its most prevalent subtype (i.e. Pancreatic Ductal Adenocarcinoma (PDAC)) [1]. By 2030, it is expected to be the most lethal malignancy of the gastrointestinal tract and the second-leading cause of death from cancer [3].

PC diagnosis relies mostly on invasive, high cost techniques such as endoscopic ultrasonography-guided fine needle aspiration (EUS-FNA), magnetic resonance imaging (MRI), and computed tomography (CT) which additionally expose patients to radiation. To relieve patients from this significant burden, the use of liquid biopsy as an alternative has drawn attention in recent years. Liquid Biopsy via a minimally invasive and risk-free procedure, well tolerated by patients, can provide valuable information about the tumor. Its constituents include a variety of analytes such as circulating tumor cells (CTCs), cell free DNA (cfDNA) and extracellular vesicles (EVs), each one with unique properties and high potential to serve as a biomarker in cancer detection, prognosis, and monitoring.

EVs is one of the most promising element in this field although their isolation and characterization present certain challenges. EVs, initially characterized as microparticles, are small membrane-bound vesicles released from cells (cellular or subcellular origin) into the circulation. Research data indicates them as key mediators of intercellular communication, contributing to different stages of carcinogenesis as well as in therapeutic resistance and various studies have described their elevated levels in cancer patients' plasma [4]. In addition, Nanoparticle Tracking Analysis (NTA) is the most widely followed method for accurate size measurement of nanoparticles such as EVs. The mean squared displacement of particles moving with Brownian motion is measured in a sample chamber illuminated by a laser beam. An example of a thorough NTA analysis is given in [5], where samples derived from either cell cultures, human monocytes, bacteria or synthetic vesicles were analyzed to delineate possible intra- and inter-assay variations.

While ML-based approaches have been widely developed and applied in the last decade for cancer diagnosis, prognosis, and treatment, there is still a lack of research into the role of EV-based characteristics as novel biomarkers for early detection and patient risk stratification in PC. Our study focuses on this specific challenging clinical issue, by exploring putative EV-based variables in combination with clinical and laboratory features, to distinguish PDAC patients from non-

* Research funding from the European Union's HORIZON-CL4-2022-DIGITAL-EMERGING-01 project under GA nos.1010931712020 Institutional Review Board Statement (https://horizon-de-phorever.eu/).

K. K., A.I., V.K. and D.I.F are with the Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, GR-45110 GREECE and with the Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, GR-45110, GREECE (e-mail: k.kourou@uoi.gr, e-mail: agg.iosif@uoi.gr, email: kigkavaso@gmail.com , e-mail: fotiadis@uoi.gr , tel: +30 265100 9006).

I. A., A. G. T., M. V. and M. M. K. are with the 2nd Department of Surgery, Aretaieion Hospital, Medical School of Athens, National and Kapodistrian University of Athens, Athens, Greece (e-mail: ioangel95@gmail.com, e-mail: konstadoulakismm@yahoo.com).

A. K. and N. M. are with the 1st Department of Propaedeutic Surgery, "Hippokratio" General Hospital, Medical School of Athens National and Kapodistrian University of Athens, Greece (e-mail: akataki@med.uoa.gr, email: nikolaosmemos@gmail.com).

V. P. is with the Laboratory of Reliability and Quality Control in Laboratory Hematology (HemQcR), Department of Biomedical Sciences, School of Health & Welfare Sciences, University of West Attica, Egaleo, Greece (email: pantazatouvasiliki@gmail.com).

oncological samples. We assessed the overall accuracy of a flexible and comprehensive ML-based framework, guided by a combination of EVs and clinical profiles, as potential predictors for patients, to identify those having PDAC among non-oncologic samples. The proposed methodology achieved promising results based on adaptive learning techniques for early disease detection.

II. MATERIALS AND METHODS

A. Study population

The study enrolled 122 patients treated in two Surgical Clinics of National and Kapodistrian University of Athens (NKUA) between March 2024 and December 2024 from patients with PDAC n=87, and age-matched non-oncologic samples n=35. All patients followed a fasting diet before blood drawing and surgical operation. For the purpose of the present study, 20ml of blood was collected in EDTA coated tubes (BD vacutainers) from each patient prior to surgical intervention after 20hr fasting. The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics and Deontology Committee of Aretaieion Hospital (488/20-02-2023). All patients or their legal representatives provided written informed consent for the assessment of their data for research purposes. Part of the plasma samples analyzed was from the "Pancreas" biobank (blood and tissue samples) approved by the Ethics and Deontology Committees of the Aretaieion Hospital (237/10-7-2020) and the Hippokrateio General Hospital of Athens (34/14-7-2020). Inclusion criteria involved histopathologically confirmed PDAC diagnosis, and absence of malignancy for non-oncologic age-matched samples. Clinical and laboratory features were retrieved from patients' medical files.

B. Nanoparticle Tracking Analysis (NTA)

After a series of centrifugation processes aiming to remove cells and cellular debris from the plasma, samples were aliquoted and processed using NTA (NanoSight S300, NTA Version 3.4) for EVs quantification and size measurement [6].

C. Predictor and outcome variables

The current analysis was based on demographic data and clinical features obtained from patients' prior surgery. EV-based characteristics were extracted following NTA analysis. The histological category considered for patient risk stratification was patients with PDAC. This histological group served as the outcome variable in the current analyses. Non-oncologic samples were also included in the current analysis. Samples with PDAC were annotated as *Class 1* (n=87) and non-oncological samples as *Class 0* (n=35). Thirteen patients out of a total of 122 samples were not included in the model's training process and were used for the external validation process.

D Data analysis

The analysis pipeline entailed preprocessing steps, feature selection, model training and validation. During the training phase a 5-fold cross-validation was applied. The Python programming language was used along with the scikit-learn ML library for the design and development of a flexible ML-based pipeline for PDAC diagnosis. [7]. The ML-based model was designed to optimize prediction performance for PC diagnosis considering PDAC tumors (C*lass 1*). To choose only the important features for training and testing the final model, the ML-based pipeline included a meta-estimator with the Random Forest (RF) algorithm [7], with reference to the classifier selected. The cross-validated model's performance on the test set was evaluated using the area under the Receiver Operating Characteristic curve (ROC AUC), which involved assessing the following metrics: specificity, sensitivity, accuracy, F-measure, and AUC.

Data preprocessing and handling of missing data

Initially, raw data was rescaled to zero mean and unit variance, and ordinal variables were converted to dummy binary variables. All cases and variables have less than 30% missingness. The remaining numbers that were missing were imputed using the *KNeighbors* classifier implementing the k-nearest neighbors vote based on the default parameters [7].

Feature Selection

Feature selection was conducted using a meta-transformer built on a Random Forest (RF) algorithm [7, 8] which assigns weights to the features and ranks them according to their relative importance. The maximum number of features to be selected by the estimator was set to 30 to identify the most significant predictors that contribute to the risk prediction of PDAC. The feature selection scheme was incorporated into the ML-based pipeline alongside the classification algorithm to select only the relevant features for training the model.

Model training and validation

A cross-validation technique was implemented, with holdout data retained for the final model evaluation to deal with the relatively prominent issue of model overfitting in ML applications in clinical research. This technique reduces both generalization errors and misclassifications during the training phase. To evaluate the model's performance and estimate the class of the new samples (namely, controls and PDAC) not utilized during the training phase, the trained models were also evaluated against an external validation dataset that included new, unseen samples (n=13). Class imbalance handling was addressed using the SMOTE - Synthetic Minority Over-sampling Technique [7]. SMOTE was implemented to over-sample the minority class and further deal with the classification task of the dataset. Class imbalance handling was considered only during the training phase in order to avoid any bias during the validation of the prediction model. Although the present study deals with mixed data types (namely, continuous and categorical characteristics) the vast majority of are of numeric type. Hence, the SMOTE technique was selected for class imbalance and applied considering also the number of samples in *Class 0* (namely the non-oncological samples) which is under-represented. Under-sampling techniques for class imbalance handling was applied for comparison reasons, but the performance was mediocre compared to SMOTE.

To build the composite estimator for model training and testing the XGBoost classifier was applied [7]. For comparison reasons we also incorporated the RF algorithm into the data-driven analytical framework [7, 8]. A 5-fold cross validation was selected for model's evaluation during the training procedure. The default parameters were applied to each estimator to compare model's performance and ensure model's optimization and reproducibility throughout the

TABLE I. PERFORMANCE OF THE XGBOOST MODEL IN PREDICTING PDAC AS OUTCOME (MEAN ± SD).

Performance metrics	XGBoost model
Specificity	0.97 +/- 0.07
Sensitivity	0.95 +/- 0.08
Accuracy	0.96 +/- 0.06
F1	0.98 +/- 0.03
AUC	0.98 +/- 0.03

whole process. The following metrics were used to assess the performance of the classification model: specificity (true negative rate); sensitivity (true positive rate); AUC, and F-measure. The Receiver Operating Characteristic (ROC) curve was also used to depict the trade-off between false negative and false positive rates at each possible cut-off. Explainability analysis was performed using SHAP values to identify the major factors driving model predictions [9, 10]. A positive SHAP number indicates that the specific trait improves the prediction, whilst a negative value suggests the opposite. This analysis provides a more detailed knowledge of each feature's impact on the model's behavior for a specific prediction case.

III. RESULTS

A. *Optimized prediction of PDAC outcomes*

As shown in Table I, the XGBoost model correctly predicted the occurrence of PDAC in 95% of patients. Moreover, the model identified patients who were non-oncological samples with approximately 96% certainty. The shape of the ROC curve with AUC=0.79, entails a fair balance between sensitivity and specificity. For comparison reasons we also calculated the model's performance with a smaller set of variables (i.e. the 10 most important predictors) and the results obtained exhibited similar metrics with the XGBoost model after retaining the 30 features (accuracy: 0.91+/-0.08, external validation score: 0.84). Performance assessment of the RF model was comparable to XGBoost model but with slight differences in the evaluation metrics. The XGBoost model was deployed to an external validation dataset including new unseen samples to quantify model's performance and better predict the class of the new samples

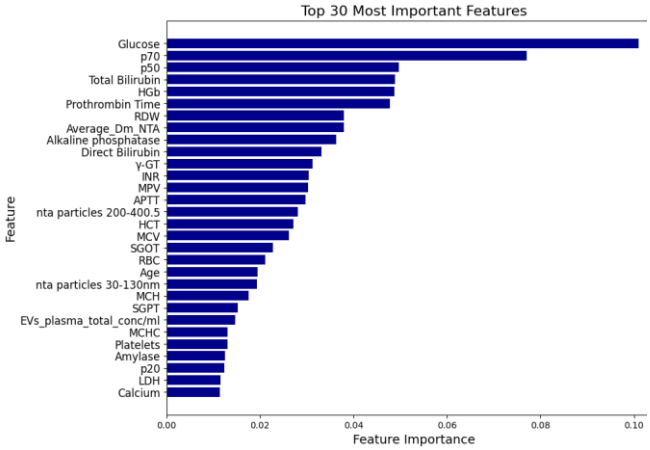

Figure 1. The selected features for XGBoost model ranked according to their relative importance for prediction of PDAC.

(i.e. non-oncological, and exocrine neoplasms). The samples were organized accordingly and followed the same preprocessing analysis as the training dataset, except for the balancing step. This outcome arises from the necessity of preserving the original data distribution during the validation process. The external validation score based on the trained model was 0.76.

B. *Identification of potential predictors of PDAC diagnosis*

Fig. 1 presents the 30 top-ranked features used for the classification of the 2 groups of patients according to the best performing model XGBoost. Highly ranked features are mainly EV-based features including average EVs' diameter, particle size distributions p70 and p50, meaning the particle sizes under which 70% and 50% of the overall particles are detected respectively, total EVs concentration, as well as concentrations of both smaller (30nm-130nm) and larger EVs (200nm-400.5nm), along with some biochemical parameters including indices associated with red-blood-cells, liver function, or blood-clotting. It should be noted that P50 and P70 reflect different points on a cumulative distribution curve, and their values differ. P50 is the median, reflecting the central tendency of vesicle size, the most "typical" EV size in the sample, while P70 gives more weight to the presence of outliers, probably indicating subpopulations like large microvesicles (in all measurements P70 size is larger than P50 size). Both can offer complementary biological insight:

The results of the SHAP analysis are illustrated in Fig. 2. Features with significant impact on the model include smaller NTA particles of sizes from 30nm-130nm and importantly, particle size distribution p70. Moreover, among the top predictors of the model are biochemical parameters such as glucose, amylase, calcium ion, red blood cell indices such as Red blood cell count (RBC), Red Cell Distribution Width (RDW), hemoglobin (HGb), hematocrit (HCT), Mean Corpuscular Volume (MCV), indices associated with liver or bile duct problems such as direct and total bilirubin, gamma-glutamyl transferase (γ-GT), Alkaline phosphatase, as well as indices associated with blood clotting such as prothrombin time, Activated Partial Thromboplastin Clotting Time (APTT), Mean Platelet Volume (MPV), and international normalized ratio for blood clotting (INR).

IV. DISCUSSION

Early diagnosis of PC at a curable stage and specifically of its most common and aggressive subtype (PDAC) constitutes a clinical unmet need, and a large amount of research is dedicated to this aim. Many studies have directed their interest into the clinical application of non-invasive and effective biomarkers like EVs whose unique physical and biological properties can provide diagnostic input [11]. The present study used adaptive learning techniques to address the clinical problem of early PC detection among PDAC patients in terms of potential EV-based characteristics and relevant clinical variables. A rigorous analytical framework was developed based on supervised learning analysis to avoid

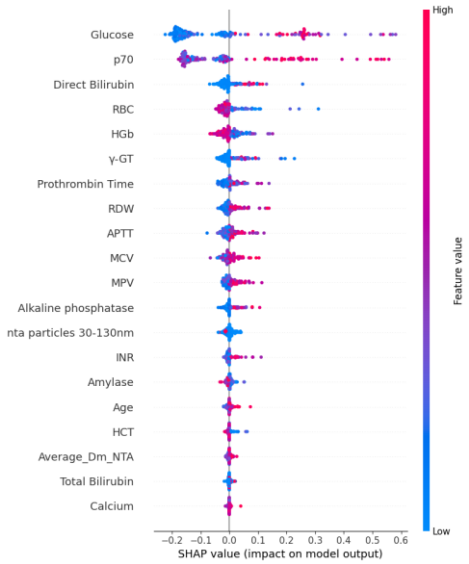

Figure 2. SHAP plot presenting the contribution of each feature to the prediction of the XGBoost model.

overfitting and model's generalizability based on external validation. To the best of our knowledge, this is the first reported effort to apply NTA analysis in plasma samples of PC patients. In the past, Nesteruk et al. performed NTA analysis and showed that the composition of isolated EVs from pancreatic juice was altered in patients with PDAC compared to controls especially of larger size EVs cases [12]. In the present study, plasma samples of PDAC cases were analyzed and particle size distribution (p70) has a significantly high impact on the ability of the classification model to discriminate between patients and non-oncologic samples. Hinestrosa et al. [13] developed ExoVita Pancreas classifier that differentiate early-stage PDAC based on multiplex immunoassays. Moreover, Yu et al. supported the high diagnostic potential of long-RNAs based signature found inside EVs derived from PDAC patients' plasma [14].The present study focuses on small EVs and examines solely the contribution of physical properties of EVs such as size along with hematological parameters.

Furthermore, the generalizability of the prediction model was assessed on cases that were not considered during the training phase, to avoid overfitting and reduce classification errors. The prediction model correctly classified 95% of the cases that had been diagnosed with PDAC. Moreover, the model identified the non-oncological samples with approximately 86% certainty. Hence, our results inspire optimism on transferring the risk stratification model for early disease detection on relevant clinical settings. The findings of our study should be considered with certain limitations. The most important include the expansion of the patient group and the collection of measurements during patient follow-up that would further substantiate the results and the accuracy of the prediction model.

V. Conclusion

The clinical potential of the physical properties of EVs facilitated by ML algorithm is highlighted in the present study. Understanding and applying such novel innovative technologies in the clinic should bring advances in the medical field, guiding critical decisions in a rapid, and efficient way. Increasing efforts to substantiate the results and the accuracy of the adaptive learning algorithms by monitoring the disease progression should provide new directions and solutions in healthcare.

Acknowledgment

Retrospective samples were from the Biobank funded during TRANSCAN-2 program ERA-NET JTC 2017, Translational research on rare cancers" General Secretariat for Research and Technology, Greece/ T9EPA3-00012.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,* vol. 71, pp. 209-249, 2021.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: a cancer journal for clinicians,* vol. 73, pp. 17-48, 2023.

[3] H. An, H. Dai, and X. Liu, "Changing Trends in the Global Disease Burden of Pancreatic Cancer from 1990 to 2030," *Digestive Diseases and Sciences,* pp. 1-12, 2024.

[4] R. Kalluri and K. M. McAndrews, "The role of extracellular vesicles in cancer," *Cell,* vol. 186, pp. 1610-1626, 2023.

[5] B. Vestad, A. Llorente, A. Neurauter, S. Phuyal, B. Kierulf, P. Kierulf, *et al.*, "Size and concentration analyses of extracellular vesicles by nanoparticle tracking analysis: a variation study," *Journal of extracellular vesicles,* vol. 6, p. 1344087, 2017.

[6] D. Bachurski, M. Schuldner, P.-H. Nguyen, A. Malz, K. S. Reiners, P. C. Grenzi, *et al.*, "Extracellular vesicle measurements with nanoparticle tracking analysis–An accuracy and repeatability comparison between NanoSight NS300 and ZetaView," *Journal of extracellular vesicles,* vol. 8, p. 1596016, 2019.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[8] W. I. D. Mining, *Introduction to data mining*: Springer, 2006.

[9] M. Scott and L. Su-In, "A unified approach to interpreting model predictions," *Advances in neural information processing systems,* vol. 30, pp. 4765-4774, 2017.

[10] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature machine intelligence,* vol. 2, pp. 56-67, 2020.

[11] L. Ma, H. Guo, Y. Zhao, Z. Liu, C. Wang, J. Bu, *et al.*, "Liquid biopsy in cancer current: status, challenges and future prospects," *Signal Transduction and Targeted Therapy,* vol. 9, p. 336, 2024.

[12] K. Nesteruk, I. J. Levink, N. F. Dits, D. L. Cahen, M. P. Peppelenbosch, M. J. Bruno, *et al.*, "Size and concentration of extracellular vesicles in pancreatic juice from patients with pancreatic ductal adenocarcinoma," *Clinical and Translational Gastroenterology,* vol. 13, p. e00465, 2022.

[13] J. P. Hinestrosa, R. C. Sears, H. Dhani, J. M. Lewis, G. Schroeder, H. I. Balcer, *et al.*, "Development of a blood-based extracellular vesicle classifier for detection of early-stage pancreatic ductal adenocarcinoma," *Communications Medicine,* vol. 3, p. 146, 2023.

[14] S. Yu, Y. Li, Z. Liao, Z. Wang, Z. Wang, Y. Li, *et al.*, "Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma," *Gut,* vol. 69, pp. 540-550, 2020.