Am J Cancer Res 2024;14(2):643-654
www.ajcr.us /ISSN:2156-6976/ajcr0154069

Original Article

Development and validation of machine learning models for diagnosis and prognosis of cancer by urinary proteomics, based on the FLEMENGHO cohort

Shuncong Wang[1,2], Dongmei Wei[1], Yanling Zhao[1], Xin Pang[3], Zhenyu Zhang[1]

[1]*Studies Coordinating Centre, Research Unit Hypertension and Cardiovascular Epidemiology, Department of Cardiovascular Sciences, KU Leuven, University of Leuven, Campus Sint Rafaël, Kapucijnenvoer 7, Block H, Box 7001, BE-3000 Leuven, Belgium;* [2]*Biomedical Group, Campus Gasthuisberg, KU Leuven, 3000 Leuven, Belgium;* [3]*Faculty of Economics and Business, KU Leuven, 3000 Leuven, Belgium*

Received October 24, 2023; Accepted December 25, 2023; Epub February 15, 2024; Published February 28, 2024

Abstract: The current study aims to develop and validate machine learning (ML) models for the prediction of cancer status by the non-invasive urinary proteomic in a population-based cohort. In this retrospective study, urinary proteome profiles in 804 cases from the FLEMENGHO cohort were measured by mass spectrometry. After feature selection by LASSO on both clinical variables and urinary proteome profile, benchmark models by clinical variables were built with six different ML algorithms. Proteome-based models and combined models were built and compared with the benchmark models. The models' performance, i.e. area under the curve (AUC) was compared by Delong method. The 95% confidence interval was estimated by the bootstrapping method. The best-performing model was explained by Shapley Additive Explanations (SHAP) method. The predictive role of proteome biomarkers in longitudinal cancer diagnosis was also explored. A clinical model, based on age, blood sugar and blood lipid profile, yielded the best AUC of 0.75 (0.68-0.82), with 0.80 (0.72-0.91) for the proteome model based on 13 selected biomarkers and 0.83 (0.77-0.90) for the combined model (P=0.01 for comparison with clinical model). SHAP on the support vector machine in the combined setting showed that except for age, proteome biomarkers contribute to the final prediction of the model. After adjusting with clinical factors, three proteome biomarkers are independent risk factors for longitudinal cancer development. Urinary proteome profiling, together with fine-tuned machine learning algorithms, demonstrates the predictive potential for cancer diagnosis transparently.

Keywords: Cancer, machine learning, spectrophotometry, proteins, urine

Introduction

Cancer remained a huge burden for the healthcare system globally, with an estimated two million new cases and six hundred thousand cancer-related deaths in 2023 [1, 2]. Currently, clinically adopted diagnostic methods include radiology, pathological examination, and body fluid (blood, cerebral spinal fluid, etc.) tests dedicated to well-established cancer biomarkers such as alpha-fetoprotein, CA125, circulating tumor DNA and so on. A great majority of cancer types were diagnosed in the metastatic stage, which is associated with a dismal prognosis [1].

Development of novel, non-invasive and easy-to-implement diagnosis/screening techniques may boost cancer early diagnosis and ultimately improve patients' survival. Urinary samples, which are of large volume and easy to collect, represent a potential candidate. Previously, the detection of cancer cells in urine samples facilitates diagnosis, staging, and treatment monitoring of urinary tract malignancy [3]. Furthermore, given the development of biotechnology, cancer DNA debris, even in low abundance in urine, can provide a highly sensitive and specific diagnosis of urinary tract malignancies [4] and such methods were adopted in clinical practices. Protein debris from malignant, apoptotic, and necrotic cells into the blood and milieu are filtered and reabsorbed in the kidneys and accumulated in the urine. Urine protein excretion is typically less than 150 mg per day for healthy individuals and is utilized as a

https://doi.org/10.62347/COTC6463

biomarker for diagnosis and monitoring kidney-involved diseases like SLE and kidney failure [5]. Stable peptides and a few infrequent but crucial proteins may be tested with robustness in the urine [6].

Based on mass spectrometry coupled with different separation techniques (like liquid chromatography, and electrophoresis), the urinary proteome represents a high-throughput, sensitive, and cheap method, which provides the quantity for up to thousands of proteins [7]. The application of proteome in urinary tract cancer has been explored with high accuracy [8, 9]. Given the tumor heterogeneity and inter-individual variation, a single biomarker strategy may not lead to sufficient discrimination between cancer patients and healthy individuals and thus a combinatorial method would be preferred. Furthermore, despite information richness (high dimensionality) in the urinary proteome, most urinary proteins are not cancer-specific and are confounded by other co-variates (sex, ageing) [10, 11], thus a dedicated selected proteome-based signature may provide a more accurate cancer diagnosis.

Feature engineering was performed to preselect biomarkers that are highly relevant to cancer diagnosis to avoid overfitting. To handle the high-dimensional data, machine learning (ML) algorithms may learn the complex association between various protein biomarkers and cancer diagnosis. ML has been widely applied in medical research, especially to high-dimensional data like proteome, genetic sequencing, radiomics, transcriptomics and so on. A random forest classifier based on five protein markers was constructed to preoperationally differentiate benign and malignant ovarian tumours based on data from a single cancer center, with an AUC of 0.952 in the test dataset [12]. This study included patients admitted to a centralized cancer center, with a diagnosis of either benign or malignant ovarian cancer, thus posing the risk of selection bias. Furthermore, the classifier is only eligible for classification between benign and malignant ovarian cancer (assumingly excluding other cancer types).

Therefore, a study elaborating on the diagnostic role of the urinary proteome in a general population is necessary, which may be helpful for community screening. Additionally, the prediction for the risk of cancer development in a longitudinal setting via urinary proteomics was also elaborated. To this end, we designed this cross-sectional study based on a population-based Flemish Study on Environment, Genes and Health Outcomes (FLEMENGHO) cohort, whose enrollment can date back to 1985 with an initial participation rate of 78% [13].

Given the complexity of proteome data, we adopted an explainable machine-learning technique for the prediction of cancer diagnosis [14]. Additionally, the future risk of developing cancer in a cancer-free population was assessed in longitudinal settings by Cox regression analyses, after adjusting for clinical covariates.

Methods

The retrospective study was approved by the ethics committee of KU Leuven. The predictive role of the proteome in cancer diagnosis was elaborated in both cross-sectional (cancer status at the time of urine sampling) and longitudinal (cancer status after urine sampling) settings.

Study population

This research was based on the population-based Flemish Study on Environment, Genes, and Health Outcomes cohort (FLEMENGHO). Starting from 1985, the initial participation rate was 78%. Written consent was acquired from every participant. Participants were periodically followed up until 2016-12-31. Using a 95% confidence level with a 5% margin of error, a sample of 385 is sufficient to answer our research question. The inclusion criterion is having proteome measurement documented and having the validated outcome data available. Finally, 804 participants were included in the current study (**Figure 1**).

Clinical data

The following clinical data were routinely collected: sex, age, body mass index (BMI), the status of a current smoker, the status of current alcohol intake, history of a cardiovascular event, usage of antihypertensive drugs. Blood pressure was repeatedly measured for 5 times after sitting for 10 minutes. Fasting blood sample was collected for measurement of serum creatine, blood sugar level, total cholesterol

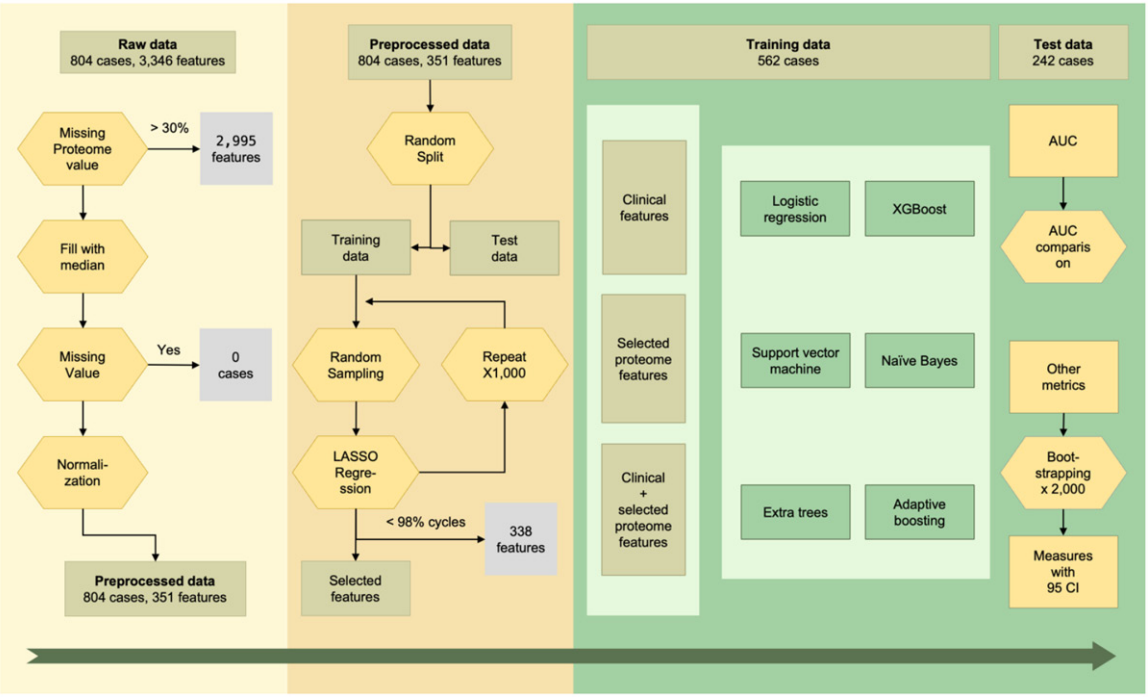


Figure 1. Flowchart of the study, consisting of data cleaning, feature selection and model training and validation. Data are shown as rectangles in olive, with operations on data as hexagons in yellow, excluded data as squares in grey and machine learning algorithms in rectangles in green. Abbreviations: LASSO: least absolute shrinkage and selection operator, AUC: area under the curve, CI: confidence interval.

and low-density lipiprotein (LDL) cholesterol. BMI is defined as body mass (kilogram) divided by the square of the body height (meter). The serum creatine and blood sugar are expressed in μmol/L and mmol/L. History of a cardiovascular event includes stroke, myocardial infarction, acute coronary syndrome, coronary artery bypass graft, percutaneous transluminal coronary angioplasty, congestive heart failure, pacemaker implantation, aortic aneurysm, pulmonary heart disease, pulmonary embolism or infarction, artheriosclerosis, other peripheral vascular disease, arterial embolism or thrombosis, gangrene, other diseases of arteries and arterioles. The mean blood pressure is defined as diastolic blood pressure + 1/3 [systolic blood pressure - diastolic blood pressure]. Cancer diagnosis status was collected at the time of urine sampling from medical history and updated during follow-up.

Urinary proteomics

Proteome data were collected with the following pipeline: sample preparation, proteome analyses by capillary electrophoresis (Beckman Coulter, Fullerton, CA) followed by mass spectrometry (micrOTOF MS, Bruker Daltonics, Bremen, Germany) and sequencing of the peptide (MosaiquesVisu software), as described in previous publications [15, 16]. Mass spectrum data were processed to produce a raw list of peptides with their molecular mass, migration time, and signal intensity. To ensure the comparability of various data sets, these raw lists were calibrated using internal urinary reference peptides [17, 18]. By fragmenting peptides and comparing the fragmentation spectra to the previously sequenced peptides from the Human Urinary Proteome Database, identified peptides were annotated [19]. Post-translational modifications and particular mass spectra were taken into account while annotating proteins. When peptides from separate samples had differences in their molecular weight and migration time of 100 parts per million and one minute, respectively, those peptides were considered to be the same. Peptides were excluded from further analyses when they were undetectable in more than 30% of cases. As a result, 351 of the 3346 urine peptides were included for further analyses.

Outcomes

The primary outcomes for cross-sectional analyses and longitudinal analyses were cancer status and incidence of cancer after baseline urinary proteome measurement, respectively. Physicians ascertained the diagnosis against the medical records of general practitioners or hospitals.

Statistical analyses

Continuous variables were shown as either mean ± standard deviation or median ± interquartile range as appropriate and were compared by t-test or Mann-Whitney U test as appropriate. Categorical variables were described as frequency and percentage and either Chi-square analyses or the Fisher's exact test was applied as appropriate. A two-sided *p*-value of less than 0.05 was considered statistically significant.

Machine learning: feature selection

After randomly splitting the entire dataset as training data (fine-tuning and validation) and test data in a ratio of 7:3, we performed feature selection in the training dataset (**Figure 1**; Supplementary Table 1). The distribution normality of each peptide was checked and log transformation at a base of 10 was applied in case of violation of normality. Proteome features missing in more than 30% of participants were excluded from the current study, and for those missing values, interpolation with median value was performed. After normalization for each biomarker, to select proteome features that are highly relevant to cancer diagnosis, the least absolute shrinkage and selection operator (LASSO) method was adopted, during which only the important features were retained in the model due to its regularization constraint. To ensure generalizability, 80% of the training data were randomly sampled to tune the LASSO, and this process was repeated 1000 times. To achieve a balance between model simplicity and model robustness, we empirically define that only the biomarkers that were retained in more than 98% of 1000 cycles were included. Finally, a correlation matrix based on Person's coefficient was built to exclude any potential collinearity. Similarly, the feature selection of the abovementioned clinical variables was also executed. For the modelling of the clinical and clinical-proteome model, only these preselected features were enrolled.

Given the low incidence of cancer in FLEMENGHO (< 5%), which may lead to deteriorated performance for ML algorithms, we adopted random upsampling of the minor class (positive cancer diagnosis) to the number of the major class (cancer-free individual). Upsampling was applied to the training data for model training, but not to the test data.

Machine learning: modelling

To prove the predictive value of proteome biomarkers, three categories of models were developed: models based on only clinical variables, models based on only proteome variables and models based on both clinical and proteome variables (combined model). The models based on clinical variables represent the benchmark models. For each of the three categories, ML modelling by different algorithms was conducted, including logistics regression, naïve Bayes, XGBoost, support vector machine, extra tree, and adaptive boosting.

To get the best performance of these algorithms, hyperparameter tuning was performed by the grid search method to achieve higher AUC, with respective tunable hyperparameters for each algorithm. Among these, untuned logistic regression served as a baseline algorithm. ML models were built based on the scikit-learn package in Python 3.10 [20].

For each model built, given the binary classification nature, model performance in the test data was quantitatively evaluated by AUC, specificity, sensitivity, and weighted f1 score, as defined below.

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

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

$$f1\ score = \frac{2 * (precision * recall)}{precision + recall}$$

$$Weighted\ f1\ score = \sum_{i=1}^{N} w_i \times f1\ score_i$$

Where w_i *is the average number of true instances for each label*.

Comparison between models was conducted by the Delong method [21]. 95% confidence

interval (CI) for each of the measures above was estimated by bootstrapping 80% of cases with a repetition of 2000 times.

Machine learning: explainable ML

To give consistent and locally correct attribution values for each feature in the model built by the support vector machine, we computed SHapley Additive exPlanation (SHAP) values for each variable [14]. This unifying strategy is used to describe the results of any machine learning model. The SHAP values assess the significance of the output that would arise from the inclusion of each feature and demonstrate this as feature importance. The intensity and direction of the impact (either positive or negative association) of each variable on the output probability were shown in a summary plot. The decision plot of each case shows how the prediction of each case was made from the same starting point at the bottom to the various probabilities. The initial probability of being cancer for each case is around 5% (if no ML model prediction was executed, based on the ratio of cancer over non-cancer) and the inclusion of more input variables will dynamically change this probability and finally reach the output probability.

Exploration of the predictive role of the proteome in a longitudinal setting

Patients with a previous cancer diagnosis at the time of urine sampling were excluded from this setting. In a univariate Cox regression setting, each proteome biomarker was input to calculate the corresponding hazard ratio (HR) and *p*-value. For biomarkers associated with a significant *p*-value in a univariate setting, each of these biomarkers was input into a multivariate Cox model, with adjustment for clinical variables mentioned above, to calculate the corresponding HR and *p*-value.

Data and code availability

The data used here can be shared under reasonable request with the corresponding author.

Model availability

The model used here can be shared in the following Github link: https://github.com/VAIOJuvenia/Proteome_prediction.git.

Result

Clinicopathological characteristics of the study population

Based on the study flowchart, 804 participants were included in this study (**Figure 1**; **Table 1**). Within the cohort, participants with cancer diagnoses are generally older ($P < 0.01$), with higher blood sugar levels ($P < 0.01$), higher mean blood pressure ($P=0.02$), higher prevalence of heart disease ($P < 0.01$) and a higher administration rate of anti-hypertensive drugs (**Table 1**). Baseline characteristics between training and test data show no significant difference between the two datasets ($P > 0.05$), except for the total-to-low cholesterol ratio (Supplementary Table 1).

Feature selection

Three clinical variables (age, blood sugar and high-to-low cholesterol ratio) and 13 peptide biomarkers (e11452, e11855, e09989, e16811, e07093, e10266, e12488, e07622, e19885, e01132, e15237, e08463 and e06068) were retained in 98% of feature selection cycles by LASSO regression. The correlation matrix reveals no strong correlation between any two biomarkers (**Figure 2**). Additionally, a comparison of the normalized value of selected biomarkers revealed statistical differences between cancer and non-cancer patients (Supplementary Figure 1). Based on database annotation, these biomarkers mostly corresponded to the collagen family. Specifically, these are collagen alpha-1(I) chain, collagen alpha-1(II) chain, collagen alpha-1(III) chain, collagen alpha-1(III) chain, uromodulin, collagen alpha-1(I) chain, collagen alpha-1(I) chain, gelsolin, collagen alpha-1(I) chain, matrix Gla protein, collagen alpha-1(III) chain, fibrinogen alpha chain and collagen alpha-1(III) chain.

Explainable ML prediction in different models

After hyperparameter tuning, clinical models show equivalent predicting power with proteome models in all ML algorithms ($P > 0.05$ for algorithm-wise comparison) (Supplementary Table 2). The best-performing clinical model was built by logistic regression, with an AUC of 0.75 (0.68-0.82), with extra trees in the proteome-based model, with an AUC of 0.80 (0.72-0.91). Interestingly, the combined models con-

Table 1. Baseline demographic and clinical characteristics of the participants

Categories	All (n=804)	Non-cancer (n=763)	Cancer (n=41)	*P* value
SEX - no. (%)				0.59
Male	396 (49.25)	378 (49.54)	18 (43.90)	
Female	408 (50.75)	385 (50.46)	23 (56.10)	
Age - yr (SD)	50.94 ± (15.77)	50.01 ± (15.47)	68.16 ± (10.47)	< 0.01
BMI - kg/m^2 (SD)	26.51 ± (4.34)	26.47 ± (4.34)	27.22 ± (4.23)	0.18
Blood sugar - mmol/L (SD)	4.94 ± (0.78)	4.91 ± (0.72)	5.43 ± (1.43)	< 0.01
Total to low cholesterol - ratio (SD)	3.87 ± (1.03)	3.87 ± (1.04)	3.77 ± (0.88)	0.85
Current smoker - no. (%)				0.78
No	643 (79.98)	609 (79.82)	34 (82.93)	
Yes	161 (20.02)	154 (20.18)	7 (17.07)	
Current alcohol intake - no. (%)				0.32
No	248 (30.85)	232 (30.41)	16 (39.02)	
Yes	556 (69.15)	531 (69.59)	25 (60.98)	
Heart event history - no. (%)				< 0.01
No	741 (92.16)	709 (92.92)	32 (78.05)	
Yes	63 (7.84)	54 (7.08)	9 (21.95)	
Serum creatine - μmol/L (SD)	84.00 ± (15.85)	83.97 ± (15.90)	84.56 ± (14.94)	0.73
Hypertension treatment - no. (%)				< 0.01
No	593 (73.76)	575 (75.36)	18 (43.90)	
Yes	211 (26.24)	188 (24.64)	23 (56.10)	
Mean blood pressure - mmHg (SD)	96.28 ± (10.73)	96.13 ± (10.81)	99.23 ± (8.49)	0.02

Abbreviations: SD: standard deviation, BMI: body mass index. Categorical variables were presented with the number of each category and the corresponding percentage. Numeric variables were presented with mean and standard deviation.

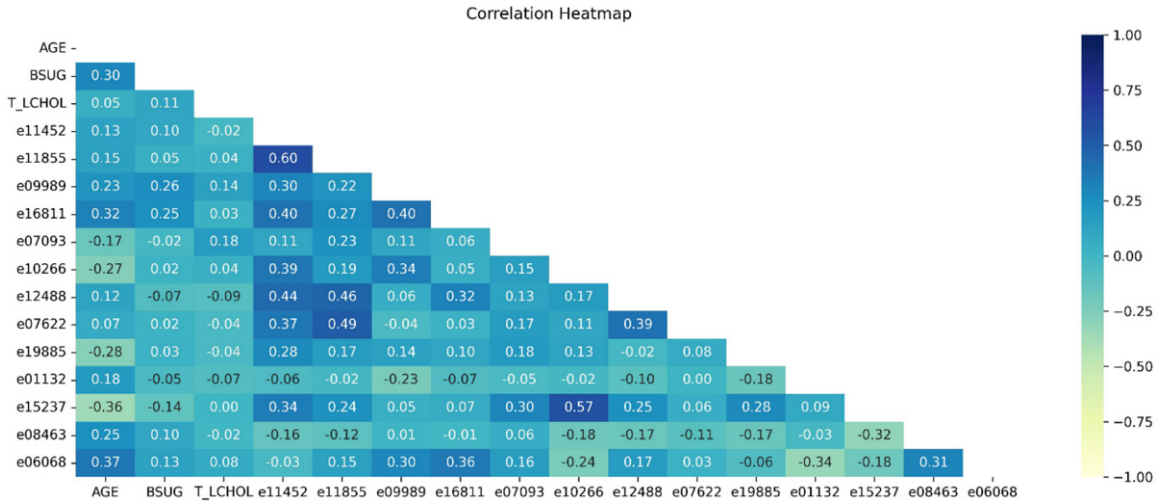


Figure 2. Pearson correlation matrix of the selected proteome features and clinical features.

structed by support vector machine showed significantly better prediction over the clinical model ($P < 0.05$, **Figure 3**). These findings were further supported by other metrics (**Table 2**). Algorithm-wise comparison in each input modality shows consistent performance in clinical data and different performance among algorithms in proteome and combined models (Supplementary Figure 2).

We represented all characteristics used in the combined model along with the SHAP summary

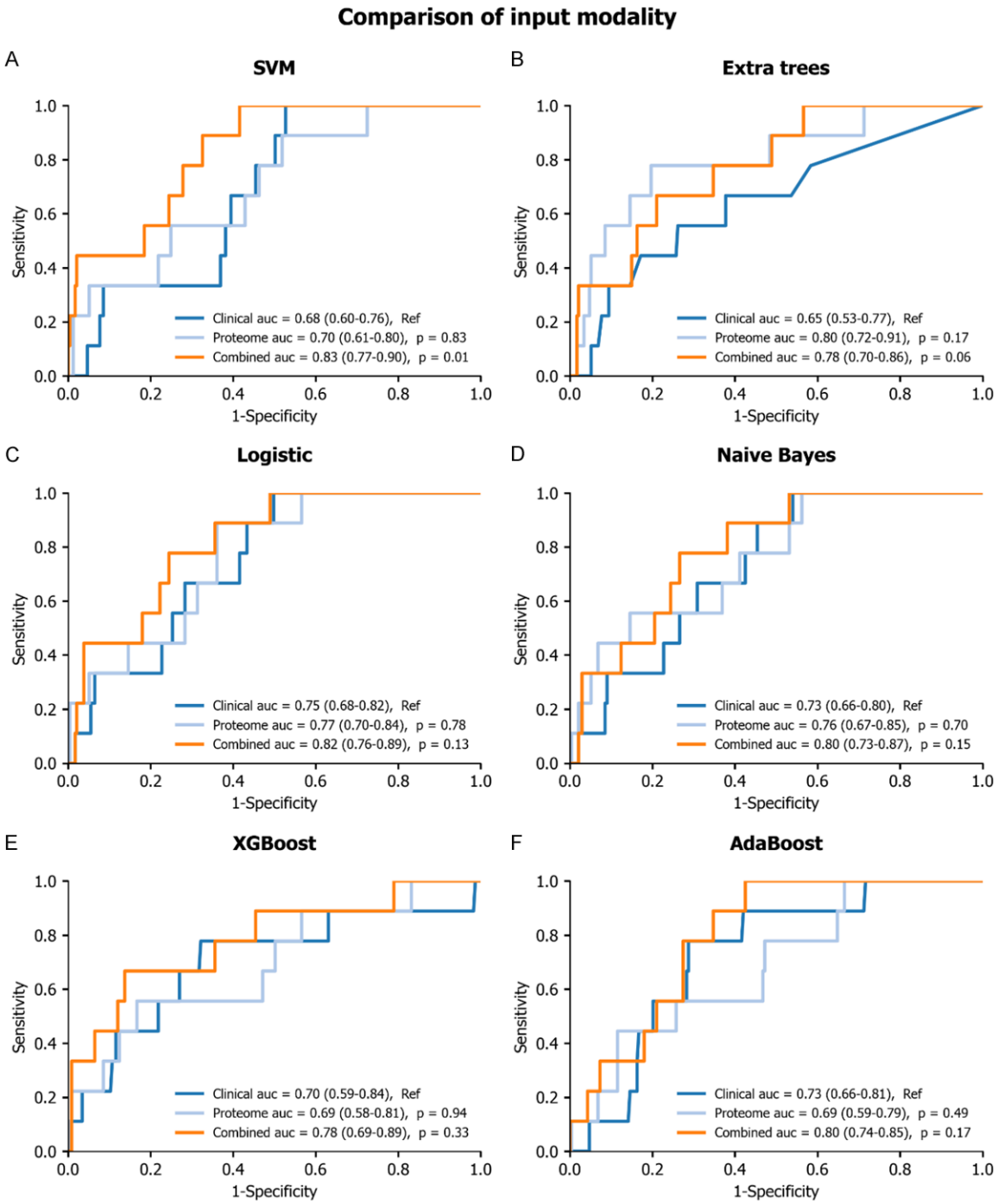


Figure 3. Receiver-operating-characteristic (ROC) curves for prediction of cancer diagnosis by either clinical model, proteome model or combined model via different algorithms. Models by support vector machine, extra trees, logistic regression, naïve Bayes, XGBoost and AdaBoost were built. The proteome model and combined model were compared with the clinical model with the Delong test. Abbreviations: Ref: reference, SVM: support vector machine.

plot of the support vector machine algorithm in the test dataset to determine which features had the most impact on the prediction model (**Figure 4**). The summary plot indicates the relationship between the value of a feature and the impact on the prediction. The decision plot's straight vertical line marks the model's base value. The coloured line is the prediction.

Table 2. Summary of comparison of different modalities for each algorithm

Algorithms	Input	ACC	AUC	Specificity	Sensitivity	F1 score
AdaBoost	Clinical	0.89 (0.88-0.91)	0.73 (0.66-0.81)	0.92 (0.91-0.94)	0.11 (0.00-0.20)	0.91 (0.89-0.93)
AdaBoost	Proteome	0.91 (0.90-0.93)	0.69 (0.59-0.79)	0.94 (0.93-0.96)	0.11 (0.00-0.20)	0.92 (0.91-0.94)
AdaBoost	Combined	0.93 (0.91-0.94)	0.80 (0.74-0.85)	0.96 (0.95-0.97)	0.11 (0.00-0.20)	0.93 (0.91-0.95)
Extra trees	Clinical	0.94 (0.92-0.95)	0.65 (0.53-0.77)	0.97 (0.97-0.98)	0.00 (0.00-0.00)	0.93 (0.91-0.95)
Extra trees	Proteome	0.96 (0.95-0.98)	0.80 (0.72-0.91)	1.00 (1.00-1.00)	0.00 (0.00-0.00)	0.94 (0.93-0.97)
Extra trees	Combined	0.95 (0.94-0.97)	0.78 (0.70-0.86)	0.99 (0.99-1.00)	0.00 (0.00-0.00)	0.94 (0.93-0.96)
Logistic	Clinical	0.76 (0.74-0.79)	0.75 (0.68-0.82)	0.78 (0.75-0.80)	0.33 (0.14-0.50)	0.83 (0.81-0.86)
Logistic	Proteome	0.77 (0.75-0.80)	0.77 (0.70-0.84)	0.79 (0.76-0.81)	0.44 (0.25-0.67)	0.84 (0.82-0.86)
Logistic	Combined	0.82 (0.80-0.85)	0.82 (0.76-0.89)	0.84 (0.81-0.86)	0.44 (0.25-0.67)	0.87 (0.85-0.89)
Naive Bayes	Clinical	0.85 (0.82-0.87)	0.73 (0.66-0.80)	0.87 (0.85-0.89)	0.33 (0.14-0.50)	0.89 (0.87-0.91)
Naive Bayes	Proteome	0.84 (0.82-0.87)	0.76 (0.67-0.85)	0.85 (0.83-0.88)	0.44 (0.25-0.67)	0.88 (0.87-0.90)
Naive Bayes	Combined	0.89 (0.88-0.91)	0.80 (0.73-0.87)	0.91 (0.90-0.94)	0.33 (0.14-0.50)	0.91 (0.90-0.93)
SVM	Clinical	0.89 (0.88-0.91)	0.68 (0.60-0.76)	0.91 (0.90-0.93)	0.33 (0.14-0.50)	0.91 (0.90-0.93)
SVM	Proteome	0.92 (0.91-0.94)	0.70 (0.61-0.80)	0.94 (0.93-0.96)	0.33 (0.14-0.50)	0.93 (0.92-0.95)
SVM	Combined	0.95 (0.94-0.97)	0.83 (0.77-0.90)	0.97 (0.97-0.98)	0.44 (0.25-0.67)	0.96 (0.94-0.97)
XGBoost	Clinical	0.90 (0.88-0.92)	0.70 (0.59-0.84)	0.92 (0.91-0.94)	0.22 (0.00-0.33)	0.92 (0.90-0.93)
XGBoost	Proteome	0.93 (0.92-0.95)	0.69 (0.58-0.81)	0.96 (0.95-0.97)	0.22 (0.00-0.40)	0.94 (0.92-0.96)
XGBoost	Combined	0.92 (0.91-0.94)	0.78 (0.69-0.89)	0.94 (0.93-0.96)	0.33 (0.14-0.50)	0.93 (0.92-0.95)

Abbreviations: ACC: accuracy, AUC: area under the curve, SVM: support vector machine.

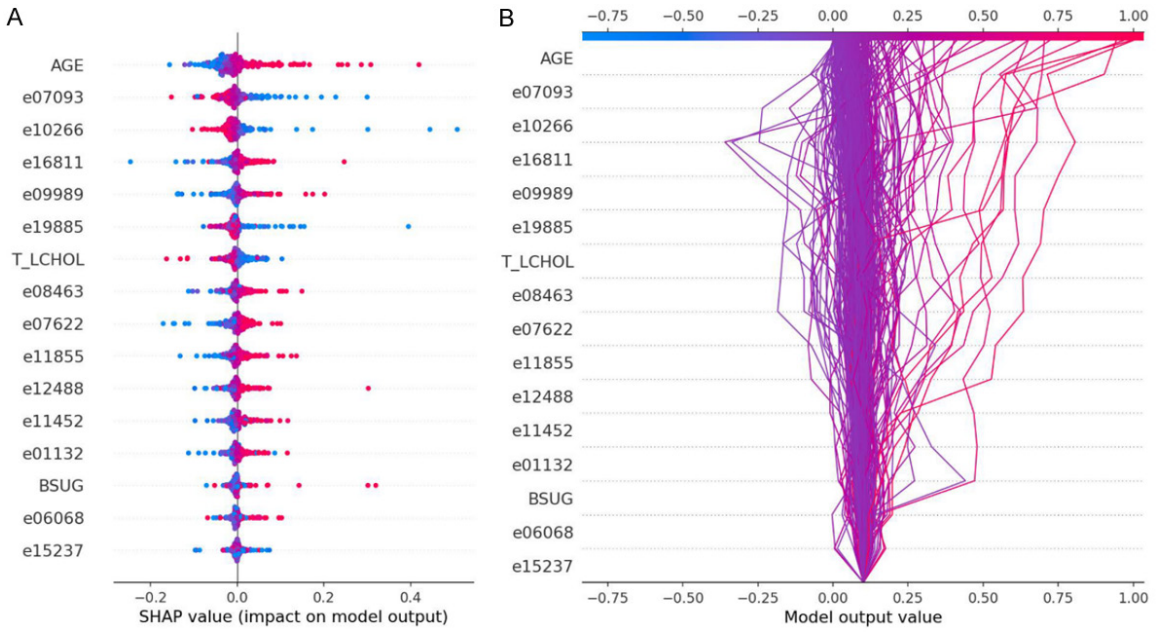


Figure 4. Explainability of the support vector machine in the combined model by the SHAP value plot and decision plot. The SHAP summary plot (A) and decision plot (B) for support vector machine in combined model.

Feature values are printed next to the prediction line for reference. Starting at the bottom of the plot, the prediction line shows how the SHAP values accumulate from the base value to arrive at the model's final score at the top of the plot. Decision plots are a literal representation of SHAP values, making them easy to interpret. Age is listed as the most important feature, which is positively associated with the probability of cancer diagnosis, whereas the

proteome biomarker (e07093) shows a negative association (Supplementary Figure 3; **Figure 4A**). Similarly, variables in the top rows contribute more to the models' output as a probability of being classified as cancer cases, as shown in the decision plot (**Figure 4B**).

Predictive role of proteome biomarkers in a longitudinal setting

After excluding participants with a cancer history, 757 participants were retained in the longitudinal cohort, among which 57 positive cancer diagnoses were observed. The follow-up time was a median of 9.11 years [interquartile range (IQR): 7.50-9.82]. Univariate Cox regression showed that 48 peptides were significantly associated with cancer diagnosis in the future (Supplementary Table 3).

The multivariable Cox regression model for each of the proteome biomarkers was adjusted for clinical variables (sex, age, body mass index (BMI), the status of a current smoker, the status of current alcohol intake, serum creatine, blood sugar, history of a cardiovascular event, administration of antihypertensive drugs, mean blood pressure and ratio of total cholesterol to LDL cholesterol) and the proportional hazard ratio assumption was met. Multivariable Cox regression showed that e03142, e05044, e08442 and e18831 are independent risk factors (Supplementary Table 3). These peptides were annotated as matrix metalloproteinase-24, collagen alpha-1(I) chain, collagen alpha-2(I) chain, and collagen alpha-1(I) chain.

Discussion

The population-based study demonstrates ML models for the prediction of cancer in a general population by urinary proteome. Specifically, it meets its preset aims: 1. Demonstration of extra predictive power from proteome profiling; 2. Demonstration of the predictive value of proteome in longitudinal risk of developing cancer.

The urinary proteome profiling is an informative, and non-invasive tool for the classification of a wide spectrum of diseases. Swensen et al. developed a database of urinary proteome profiling from various medical conditions, including healthy individuals (n=10), renal transplant recipients with acute rejection (n=10) or stable grafts (n=10), patients with non-specific proteinuria (n=10), and prostate cancer (n=5) patients. Here, they discovered hundreds of disease-related proteins, and for instance, serpin B3, renin receptor, and periostin are biomarkers for renal failure and prostate cancer, respectively [22]. To ensure the simplicity of the model and easiness of translational application, we adopted critical feature selection by LASSO which identifies 13 peptides. However, the detailed cause-effect association between identified peptides and cancer development remained unexplored in the current population. Among these 13 peptides, the majority are from the collagen family. Collagen is identified as a biomarker of non-cancer diseases like diabetes-related kidney disease, diastolic left ventricular dysfunction, liver fibrosis, interstitial fibrosis and tubular atrophy in chronic kidney disease, and bone resorption and bisphosphonate treatment in kidney transplant patients [23-27]. From a biological perspective, collagen is a major component of the extracellular matrix in normal tissue and cancer microenvironment. In addition, collagen, interacting with cancer cells as well as inflammatory cells, contributes to the proliferation, invasion, metastasis, treatment resistance, anti-cancer immunity regulation, hypoxia regulation and so forth [28]. In line with previous publications, collagen has been identified as a prognostic biomarker for a variety of cancer types [28]. For example, the higher mRNA and protein level of type I collagen, measured by polymerase chain reaction and immunohistochemistry respectively, are associated with a poorer prognosis in non-muscle invasive bladder cancers [29]. Additionally, urine proteome reveals that ANXA11, CDC42, NAPA and SLC25A4 were positively associated with the risk of gastric lesion progression into malignancy, with an AUC (95% confidence interval) of 0.92 (0.83-1.00), based on a case-control study of 255 cases from Linqu, China, a high-risk area for gastric cancer [30]. Urinary peptide signature represents a biomarker for the diagnosis of colorectal cancer and the development of liver metastases. Collagen was identified as a predictive biomarker for the diagnosis of colorectal liver metastases in a Western Europe population [31]. However, no peptide related to collagen was included in the signature, based on a case-control study of 657 healthy control and 993 colorectal cancer patients in an Asian population [32]. The dis-

parity among these results, together with our data, may be attributed to the different races of participants. The findings here, together with other studies, highlight the potential role as well as the heterogeneity of collagen in carcinogenesis, calling for a multi-center trial consisting of multiracial participants.

Given that different ML models perform differently depending on the training dataset characteristics, six ML algorithms were trained with hyperparameter tuning to achieve the best performance. Here, the support vector machine outperforms others and detects a significant difference. Although other algorithms can illustrate the difference between different input modalities, the difference is not statistically significant. This may be attributed to the relatively weaker stratification power of these algorithms and the relatively small sample size of the current study.

Among commonly collected clinical variables (sex, age, blood tests, cardiovascular treatment and so on) that were reported to be associated with cancer development [33-35], only age, blood sugar level and lipid profile were retained after LASSO regression here. The disparity between our results and other publications may be explained partially by different participant characteristics (generally older population here), and data collection and processing strategies. The derived clinical model shows comparable performance with the proteome-based model. Interestingly, the combinational model has produced a significantly higher AUC value than any single modality, which indicates that the urinary proteome can provide extra and complementary predictive information than clinical variables. Since the clinical variables can be easily obtained, the implementation of the combined model will not require extra examination or tests. Furthermore, the superior performance of the combined model also implies the necessity of the clinical information, which hopefully may boost the performance of the proteome-only pipeline [12, 36]. The SHAP summary plot for the support vector machine with combined input exhibited some similar predictors known to be associated with cancer development and additionally, our plot revealed additional novel predictors from urinary proteome.

Compared with case-control studies, this study utilized data from a healthy cohort, representing the generalizability of the conclusion here. However, for a population-based setting, one of the practical questions is the lower incidence of events of interest, namely class imbalance, which may lead to insufficient training of ML models. Based on our data (around 5% of cancer diagnoses), we performed random duplication for upsampling. In the model validation process, accuracy was not adopted here due to the class imbalance.

The study has the following limitations. Firstly, the result here is based on a single cohort, with one detection method and further external validation is suggested to confirm the generalizability of the ML models proposed. Secondly, detailed information on cancer including, staging, histology is missing, which limits the subgroup analyses. Thirdly, the exclusion of cases with urological malignancy, which may have an impact on the urinary proteome, was not done here as the main aim was to establish a cancer signature for all. Additionally, the total number of cancer cases is relatively limited for subgroup analyses.

In conclusion, this study demonstrated the feasibility of the prediction of cancer diagnosis by urinary proteome empowered by fine-tuned machine learning algorithms. A further validation with larger sample sizes in a multi-centre setting may assist the clinical screening of cancer for general purposes in the future.

Acknowledgements

We thank all colleagues for their contribution to the data collection and maintenance of the dataset. The work was supported by the European Research Area Net for Cardiovascular Diseases Funding (JTC2017-046-PROACT) and KU Leuven Funding (STG-18-00379).

Written informed consent was obtained from participants.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Zhenyu Zhang, Studies Coordinating Centre, Research Unit Hypertension and Cardiovascular Epidemiology, Department of Cardiovascular Sciences, KU Leuven, University of Leuven, Campus Sint Rafaël, Kapucijnenvoer 7, Block H, Box 7001, BE-3000 Leuven, Belgium. Tel: +32-16193398; Fax: +32-16193398; E-mail: zhenyu.zhang@kuleuven.be

References

[1] Wang S, Liu Y, Feng Y, Zhang J, Swinnen J, Li Y and Ni Y. A review on curability of cancers: more efforts for novel therapeutic options are needed. Cancers (Basel) 2019; 11: 1782.

[2] Siegel RL, Miller KD, Wagle NS and Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023; 73: 17-48.

[3] Ahn JH, Kang CK, Kim EM, Kim AR and Kim A. Proteomics for early detection of non-muscle-invasive bladder cancer: clinically useful urine protein biomarkers. Life (Basel) 2022; 12: 395.

[4] Wang P, Shi Y, Zhang J, Shou J, Zhang M, Zou D, Liang Y, Li J, Tan Y, Zhang M, Bi X, Zhou L, Ci W and Li X. UCseek: ultrasensitive early detection and recurrence monitoring of urothelial carcinoma by shallow-depth genome-wide bisulfite sequencing of urinary sediment DNA. EBioMedicine 2023; 89: 104437.

[5] Chen TK, Knicely DH and Grams ME. Chronic kidney disease diagnosis and management: a review. JAMA 2019; 322: 1294-1304.

[6] Vanarsa K, Castillo J, Wang L, Lee KH, Pedroza C, Lotan Y and Mohan C. Comprehensive proteomics and platform validation of urinary biomarkers for bladder cancer diagnosis and staging. BMC Med 2023; 21: 133.

[7] Thomas S, Hao L, Ricke WA and Li L. Biomarker discovery in mass spectrometry-based urinary proteomics. Proteomics Clin Appl 2016; 10: 358-370.

[8] Lin L, Yu Q, Zheng J, Cai Z and Tian R. Fast quantitative urinary proteomic profiling workflow for biomarker discovery in kidney cancer. Clin Proteomics 2018; 15: 42.

[9] Bergamini S, Caramaschi S, Monari E, Martorana E, Salviato T, Mangogna A, Balduit A, Tomasi A, Canu P and Bellei E. Urinary proteomic profiles of prostate cancer with different risk of progression and correlation with histopathological features. Ann Diagn Pathol 2021; 51: 151704.

[10] Shao C, Zhao M, Chen X, Sun H, Yang Y, Xiao X, Guo Z, Liu X, Lv Y, Chen X, Sun W, Wu D and Gao Y. Comprehensive analysis of individual variation in the urinary proteome revealed significant gender differences. Mol Cell Proteomics 2019; 18: 1110-1122.

[11] Penn DJ, Zala SM and Luzynski KC. Regulation of sexually dimorphic expression of major urinary proteins. Front Physiol 2022; 13: 822073.

[12] Ni M, Zhou J, Zhu Z, Yuan J, Gong W, Zhu J, Zheng Z and Zhao H. A novel classifier based on urinary proteomics for distinguishing between benign and malignant ovarian tumors. Front Cell Dev Biol 2021; 9: 712196.

[13] Wei D, Melgarejo JD, Thijs L, Temmerman X, Vanassche T, Van Aelst L, Janssens S, Staessen JA, Verhamme P and Zhang ZY. Urinary proteomic profile of arterial stiffness is associated with mortality and cardiovascular outcomes. J Am Heart Assoc 2022; 11: e024769.

[14] Lundberg SM, Erion GG and Lee SI. Consistent Individualized Feature Attribution for Tree Ensembles [Internet], 2019. [cited 2023 May 21]. Available from: http://arxiv.org/abs/1802.03888.

[15] Wittke S, Mischak H, Walden M, Kolch W, Rädler T and Wiedemann K. Discovery of biomarkers in human urine and cerebrospinal fluid by capillary electrophoresis coupled to mass spectrometry: towards new diagnostic and therapeutic approaches. Electrophoresis 2005; 26: 1476-1487.

[16] Theodorescu D, Wittke S, Ross MM, Walden M, Conaway M, Just I, Mischak H and Frierson HF. Discovery and validation of new protein biomarkers for urothelial cancer: a prospective analysis. Lancet Oncol 2006; 7: 230-240.

[17] Jantos-Siwy J, Schiffer E, Brand K, Schumann G, Rossing K, Delles C, Mischak H and Metzger J. Quantitative urinary proteome analysis for biomarker evaluation in chronic kidney disease. J Proteome Res 2009; 8: 268-281.

[18] Schiess R, Mueller LN, Schmidt A, Mueller M, Wollscheid B and Aebersold R. Analysis of cell surface proteome changes via label-free, quantitative mass spectrometry. Mol Cell Proteomics 2009; 8: 624-638.

[19] Stalmach A, Albalat A, Mullen W and Mischak H. Recent advances in capillary electrophoresis coupled to mass spectrometry for clinical proteomic applications. Electrophoresis 2013; 34: 1452-64. Wiley Online Library [Internet] [cited 2023 May 21] Available from: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/elps.201200708.

[20] Scikit-learn: machine learning in python [Internet] [cited 2023 May 21]. Available from: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

[21] Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Journals & Magazine | IEEE Xplore [Internet] [cited 2023 May 21]. Available from: https://ieeexplore.ieee.org/document/6851192/authors#authors.

[22] Swensen AC, He J, Fang AC, Ye Y, Nicora CD, Shi T, Liu AY, Sigdel TK, Sarwal MM and Qian WJ. A comprehensive urine proteome database generated from patients with various renal conditions and prostate cancer. Front Med (Lausanne) 2021; 8: 548212.

[23] Marx D, Anglicheau D, Caillard S, Moulin B, Kochman A, Mischak H, Pejchinowski M, Latosinska A, Bienaimé F, Prié D, Marquet P, Perrin P, Gwinner W and Metzger J. Urinary collagen-derived peptides as sensitive markers for bone resorption and bisphosphonate treatment in kidney transplant patients [Internet], 2022 2022.02.15.22270979. [cited 2023 May 21]. Available from: https://www.medrxiv.org/content/10.1101/2022.02.15.22270979v1.

[24] Catanese L, Siwy J, Mavrogeorgis E, Amann K, Mischak H, Beige J and Rupprecht H. A novel urinary proteomics classifier for non-invasive evaluation of interstitial fibrosis and tubular atrophy in chronic kidney disease. Proteomes 2021; 9: 32.

[25] Bannaga AS, Metzger J, Kyrou I, Voigtländer T, Book T, Melgarejo J, Latosinska A, Pejchinovski M, Staessen JA, Mischak H, Manns MP and Arasaradnam RP. Discovery, validation and sequencing of urinary peptides for diagnosis of liver fibrosis-A multicentre study. EBioMedicine 2020; 62: 103083.

[26] Fan G, Gong T, Lin Y, Wang J, Sun L, Wei H, Yang X, Liu Z, Li X, Zhao L, Song L, He J, Liu H, Li X, Liu L, Li A, Lu Q, Zou D, Wen J, Xia Y, Wu L, Huang H, Zhang Y, Xie W, Huang J, Luo L, Wu L, He L, Liang Q, Chen Q, Chen G, Bai M, Qin J, Ni X, Tang X and Wang Y. Urine proteomics identifies biomarkers for diabetic kidney disease at different stages. Clin Proteomics 2021; 18: 32.

[27] Zhang ZY, Nkuipou-Kenfack E and Staessen JA. Urinary peptidomic biomarker for personalized prevention and treatment of diastolic left ventricular dysfunction. Proteomics Clin Appl 2019; 13: e1800174.

[28] Xu S, Xu H, Wang W, Li S, Li H, Li T, Zhang W, Yu X and Liu L. The role of collagen in cancer: from bench to bedside. J Transl Med 2019; 17: 309.

[29] Brooks M, Mo Q, Krasnow R, Ho PL, Lee YC, Xiao J, Kurtova A, Lerner S, Godoy G, Jian W, Castro P, Chen F, Rowley D, Ittmann M and Chan KS. Positive association of collagen type I with non-muscle invasive bladder cancer progression. Oncotarget 2016; 7: 82609-82619.

[30] Fan H, Li X, Li ZW, Zheng NR, Cao LH, Liu ZC, Liu MW, Li K, Wu WH, Li ZX, Zhou T, Zhang Y, Liu WD, Zhang LF, You WC, Wang Y, Wu J, Pan KF, Qin J and Li WQ. Urine proteomic signatures predicting the progression from premalignancy to malignant gastric cancer. EBioMedicine 2022; 86: 104340.

[31] van Huizen NA, van Rosmalen J, Dekker LJM, Coebergh van den Braak RRJ, Verhoef C, IJzermans JNM and Luider TM. Identification of a collagen marker in urine improves the detection of colorectal liver metastases. J Proteome Res 2020; 19: 153-160.

[32] Sun Y, Guo Z, Liu X, Yang L, Jing Z, Cai M, Zheng Z, Shao C, Zhang Y, Sun H, Wang L, Wang M, Li J, Tian L, Han Y, Zou S, Gao J, Zhao Y, Nan P, Xie X, Liu F, Zhou L, Sun W and Zhao X. Noninvasive urinary protein signatures associated with colorectal cancer diagnosis and metastasis. Nat Commun 2022; 13: 2757.

[33] Kim HI, Lim H and Moon A. Sex differences in cancer: epidemiology, genetics and therapy. Biomol Ther (Seoul) 2018; 26: 335-342.

[34] Wang Y, Wang Y, Han X, Sun J, Li C, Adhikari BK, Zhang J, Miao X and Chen Z. Cardio-oncology: a myriad of relationships between cardiovascular disease and cancer. Front Cardiovasc Med 2022; 9: 727487.

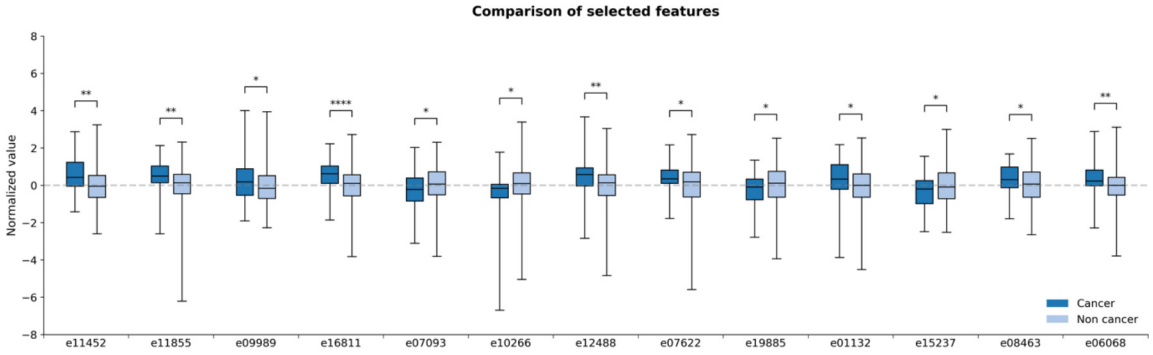
[35] White MC, Holman DM, Boehm JE, Peipins LA, Grossman M and Henley SJ. Age and cancer risk: a potentially modifiable relationship. Am J Prev Med 2014; 46 Suppl 1: S7-15.

[36] Zhang C, Leng W, Sun C, Lu T, Chen Z, Men X, Wang Y, Wang G, Zhen B and Qin J. Urine proteome profiling predicts lung cancer from control cases and other tumors. EBioMedicine 2018; 30: 120-128.

Supplementary Table 1. Baseline demographic and clinical characteristics of the participants

Categories	All (n=804)	Train (n=562)	Test (n=242)	*P* value
SEX - no. (%)				0.96
Male	396 (49.25)	276 (49.11)	120 (49.59)	
Female	408 (50.75)	286 (50.89)	122 (50.41)	
Age - yr (SD)	50.94 ± (15.77)	51.43 ± (15.81)	49.80 ± (15.62)	0.23
BMI - kg/m^2 (SD)	26.51 ± (4.34)	26.58 ± (4.33)	26.34 ± (4.36)	0.33
Blood sugar - mmol/L (SD)	4.94 ± (0.78)	4.83 ± (0.52)	5.03 ± (0.63)	0.26
Total to low cholesterol - ratio (SD)	3.87 ± (1.03)	3.92 ± (1.05)	3.74 ± (0.96)	0.03
Current smoker - no. (%)				0.70
No	643 (79.98)	452 (80.43)	191 (78.93)	
Yes	161 (20.02)	110 (19.57)	51 (21.07)	
Current alcohol intake - no. (%)				0.49
No	248 (30.85)	178 (31.67)	70 (28.93)	
Yes	556 (69.15)	384 (68.33)	172 (71.07)	
Heart event history - no. (%)				1.00
No	741 (92.16)	518 (92.17)	223 (92.15)	
Yes	63 (7.84)	44 (7.83)	19 (7.85)	
Serum creatine - μmol/L (SD)	84.00 ± (15.85)	84.20 ± (14.17)	83.70 ± (15.52)	0.52
Hypertension treatment - no. (%)				0.86
No	593 (73.76)	416 (74.02)	177 (73.14)	
Yes	211 (26.24)	146 (25.98)	65 (26.86)	
Mean blood pressure - mmHg (SD)	96.28 ± (10.73)	96.37 ± (10.48)	96.07 ± (11.28)	0.61

Abbreviations: SD: standard deviation, BMI: body mass index. Categorical variables were presented with the number of each category and the corresponding percentage. Numeric variables were presented with mean and standard deviation.

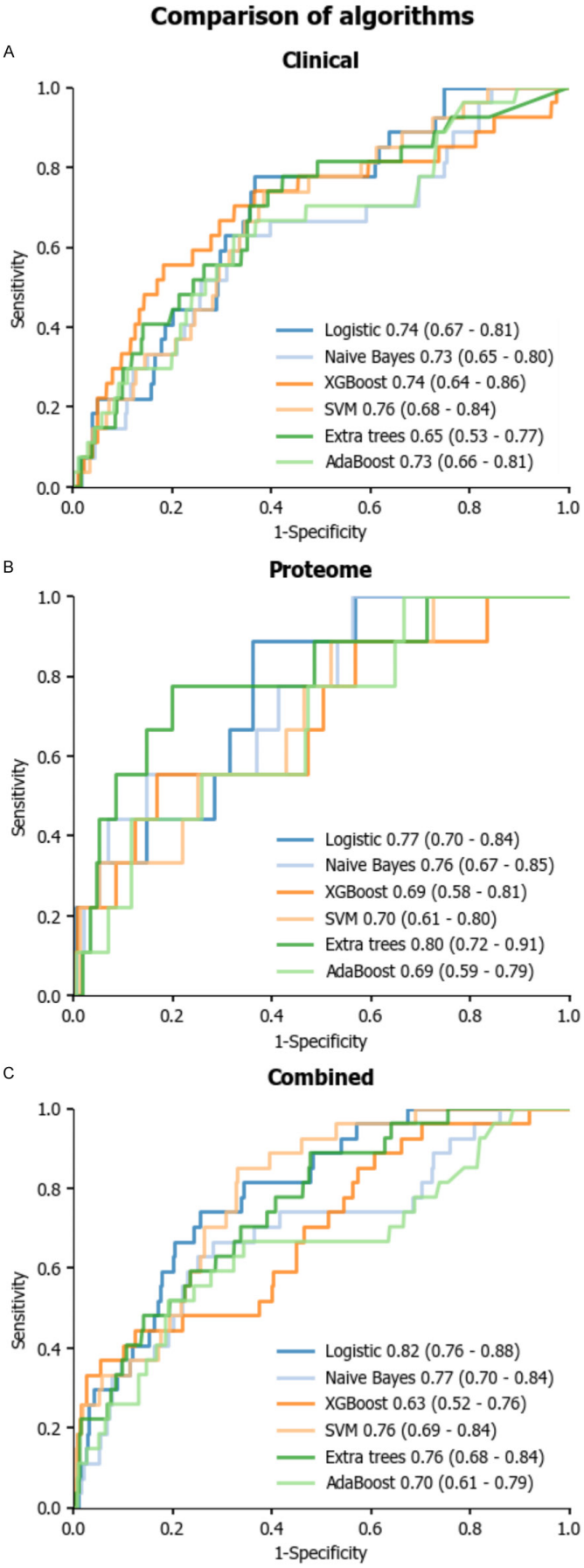


Supplementary Figure 1. Comparison of normalized values of selected features between cancer and non-cancer groups. Abbreviations: *: $P \leq 0.05$, **: $P \leq 0.01$, ****: $P \leq 0.0001$.

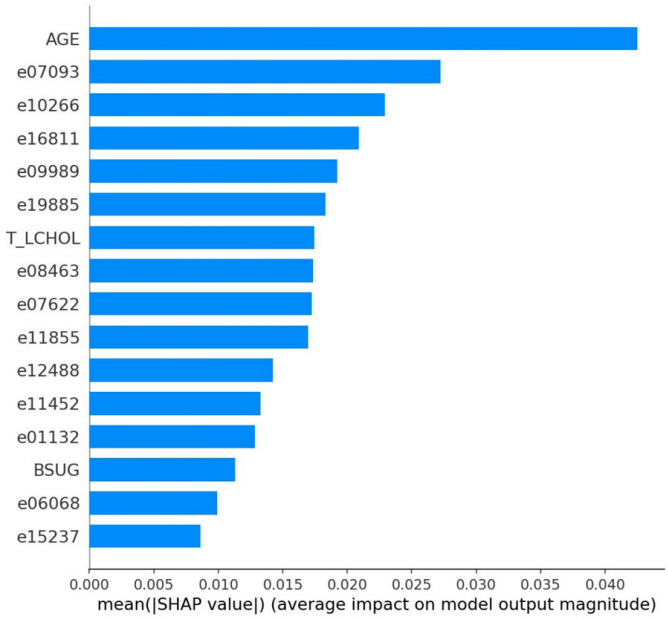
Supplementary Table 2. Tuned hyperparameters for different models under various input modalities

Algorithms	Parameter	Clinical	Proteome	Combined
SVM	C	1	1	0.1
	class_weight	Balanced	Balanced	Balanced
	gamma	0.1	auto	0.1
	kernel	poly	poly	Poly
Extra trees	class_weight	Balanced	Balanced	Balanced
	criterion	gini	gini	gini
	max_depth	20	10	10
	n_estimators	50	50	50
Naïve Bayes	var_smoothing	0.811	1.000	1.000
XGBoost	learning_rate	0.2	0.1	0.2
	max_depth	5	4	3
	scale_pos_weight	100	10	99
AdaBoost	learning_rate	1	1	1
	n_estimators	300	300	300

Abbreviation: SVM: support vector machine.


Comparison of algorithms
A
Clinical
Sensitivity
1-Specificity
Logistic 0.74 (0.67 - 0.81)
Naive Bayes 0.73 (0.65 - 0.80)
XGBoost 0.74 (0.64 - 0.86)
SVM 0.76 (0.68 - 0.84)
Extra trees 0.65 (0.53 - 0.77)
AdaBoost 0.73 (0.66 - 0.81)
B
Proteome
Sensitivity
1-Specificity
Logistic 0.77 (0.70 - 0.84)
Naive Bayes 0.76 (0.67 - 0.85)
XGBoost 0.69 (0.58 - 0.81)
SVM 0.70 (0.61 - 0.80)
Extra trees 0.80 (0.72 - 0.91)
AdaBoost 0.69 (0.59 - 0.79)
C
Combined
Sensitivity
1-Specificity
Logistic 0.82 (0.76 - 0.88)
Naive Bayes 0.77 (0.70 - 0.84)
XGBoost 0.63 (0.52 - 0.76)
SVM 0.76 (0.69 - 0.84)
Extra trees 0.76 (0.68 - 0.84)
AdaBoost 0.70 (0.61 - 0.79)

Supplementary Figure 2. Receiver-operating-characteristic (ROC) curves for prediction of cancer diagnosis by each algorithm via different input modalities. Models by support vector machine, extra trees, logistic regression, naïve Bayes, XGBoost and AdaBoost were built. The proteome model and combined model were compared with the clinical model with the Delong test. Abbreviations: Ref: reference, SVM: support vector machine.



Supplementary Figure 3. Feature importance by SHAP. Abbreviations: BSUG: blood sugar, T-LCHOL: total to low cholesterol ratio, SHAP: Shapley Additive Explanations.

Supplementary Table 3. Multivariate Cox regression for proteome data in prediction of a future cancer diagnosis

Peptide	Protein Name	Abbreviations	HR (95% CI)	*P* value
e00340	Matrix Gla protein	MGP	1.0 (1.0 ± 1.0)	0.45
e01132	Matrix Gla Protein	MGP	1.0 (1.0 ± 1.0)	0.09
e01274	POTE ankyrin domain family member F	POTEF	1.0 (1.0 ± 1.0)	0.58
e03016	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.98
e03142	Matrix metalloproteinase-24	MMP24	1.0 (1.0 ± 1.0)	0.02
e03180	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.64
e03248	Protocadherin-9	PCDH9	1.0 (1.0 ± 1.0)	0.11
e04419	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.42
e05044	Collagen alpha-1(I) chain	COL1A1	0.99 (0.99 ± 1.0)	0.02
e05074	Collagen alpha-2(I) chain	COL1A2	1.0 (1.0 ± 1.0)	0.05
e05560	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.74
e05830	Collagen alpha-1(XXII) chain	COL22A1	1.0 (1.0 ± 1.0)	0.37
e06154	Collagen alpha-2(I) chain	COL1A2	1.0 (1.0 ± 1.0)	0.35
e06213	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.43
e06288	Collagen alpha-2(I) chain	COL1A2	1.0 (1.0 ± 1.0)	0.22
e06650	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.23
e06733	Fibrinogen alpha chain	FGA	1.0 (1.0 ± 1.0)	0.91
e06839	Collagen alpha-2(I) chain	COL1A2	1.0 (0.99 ± 1.0)	0.15
e06961	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.52
e07098	Collagen alpha-1(V) chain	COL5A1	1.0 (1.0 ± 1.0)	0.52
e07132	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.09
e07513	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.34
e07678	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.74
e08188	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.69
e08442	Collagen alpha-2(I) chain	COL1A2	1.0 (1.0 ± 1.0)	0.02
e09449	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.13
e09697	Collagen alpha-1(XXV) chain	COL25A1	1.0 (1.0 ± 1.0)	0.45
e10771	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.07
e11008	Collagen alpha-2(V) chain	COL5A2	1.0 (1.0 ± 1.0)	0.07
e11073	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.75
e11325	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.53
e11641	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.06
e11753	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.46
e11780	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.17
e12851	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.87
e12949	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.81
e12986	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.59
e13065	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.61
e13707	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.44
e14204	Collagen alpha-2(XI) chain	COL11A2	1.0 (1.0 ± 1.0)	0.43
e14837	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.05
e15323	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.26
e17280	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.65
e17856	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.46
e18831	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0
e18864	Collagen alpha-1(II) chain	COL2A1	1.0 (1.0 ± 1.0)	0.58
e18867	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.85
e19885	Collagen alpha-1(I) chain	COL1A1	1.0 (1.0 ± 1.0)	0.61
e20065	Collagen alpha-1(III) chain	COL3A1	1.0 (1.0 ± 1.0)	0.91