Journal of Pharmaceutical and Biomedical Analysis 244 (2024) 116113

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.journals.elsevier.com/journal-of-pharmaceutical-and-biomedical-analysis

Enhancing breast cancer screening with urinary biomarkers and Random Forest supervised classification: A comprehensive investigation

Eugenio Alladio [a,b], Fulvia Trapani [a,b], Lorenzo Castellino [a,b], Marta Massano [a,b], Daniele Di Corcia [b], Alberto Salomone [a,b], Enrico Berrino [c,d], Riccardo Ponzone [d], Caterina Marchiò [c,d], Anna Sapino [c,d], Marco Vincenti [a,b,*]

[a] Department of Chemistry, University of Turin, Italy
[b] Centro Regionale Antidoping, Orbassano, TO, Italy
[c] Department of Medical Sciences, University of Turin, Turin, Italy
[d] Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy

ARTICLE INFO

Keywords:
breast cancer
urine biomarkers
sex hormones
steroids
supervised classification
machine-learning

ABSTRACT

Objectives: Urinary sex hormones are investigated as potential biomarkers for the early detection of breast cancer, aiming to evaluate their relevance and applicability, in combination with supervised machine-learning data analysis, toward the ultimate goal of extensive screening.
Methods: Sex hormones were determined on urine samples collected from 250 post-menopausal women (65 healthy - 185 with breast cancer, recruited among the clinical patients of Candiolo Cancer Institute FPO-IRCCS (Torino, Italy). Two analytical procedures based on UHPLC-MS/HRMS were developed and comprehensively validated to quantify 20 free and conjugated sex hormones from urine samples. The quantitative data were processed by seven machine learning algorithms. The efficiency of the resulting models was compared.
Results: Among the tested models aimed to relate urinary estrogen and androgen levels and the occurrence of breast cancer, Random Forest (RF) proved to underscore all the other supervised classification approaches, including Partial Least Squares – Discriminant Analysis (PLS-DA), in terms of effectiveness and robustness. The final optimized model built on only five biomarkers (testosterone-sulphate, alpha-estradiol, 4-methoxyestradiol, DHEA-sulphate, and epitestosterone-sulphate) achieved an approximate 98% diagnostic accuracy on replicated validation sets. To balance the less-represented population of healthy women, a Synthetic Minority Oversampling TEchnique (SMOTE) data oversampling approach was applied.
Conclusions: By means of tunable hyperparameters optimization, the RF algorithm showed great potential for early breast cancer detection, as it provides clear biomarkers ranking and their relative efficiency, allowing to ground the final diagnostic model on a restricted selection five steroid biomarkers only, as desirable for noninvasive tests with wide screening purposes.

1. Introduction

Breast cancer is the leading cause of cancer deaths for women and accounts for about 12% of new total diagnosed cases worldwide, according to the new Global Cancer statistics 2020 [1]. In patients' treatment and life expectancy, a fundamental role is played by early detection, requiring specific, efficient, and easily measurable biomarkers for routine screening. Steroids are among the candidate biomarkers, since epidemiological studies showed correlation between endogenous steroid hormone levels and the increased risk of developing

Abbreviations: RF, random forest; PLS-DA, partial least squares – discriminant analysis; SMOTE, synthetic minority over-sampling technique; UHPLC, ultra-high performance liquid chromatography; QTOF-HRMS, quadrupole/time-of-flight high-resolution mass spectrometry; DHEA, dehydroepiandrosterone; TBME, tert-butyl methyl ether; LOD, limit of detection; LOQ, limit of quantification; CV%, coefficient of variation; S/N, signal-to-noise ratio; ME, matrix effect; ML, machine learning; LR, logistic regression; NB, naive Bayes; K-NN, k-nearest neighbours; CART, classification and regression trees; SVM, support vector machine; ROC, receiver operating characteristic; PRC, precision-recall curve; AUC, area under the curve; SHAP, SHapley Additive exPlanations.

* Correspondence to: Department of Chemistry, University of Turin, Via Pietro Giuria, 7, Torino 10125, Italy.
E-mail address: marco.vincenti@unito.it (M. Vincenti).

https://doi.org/10.1016/j.jpba.2024.116113
Received 27 January 2024; Received in revised form 10 March 2024; Accepted 15 March 2024
Available online 20 March 2024

breast, ovarian and endometrial cancer [2–4].

Steroids are lipid-based hormones synthesized from cholesterol in endocrine glands and encompass androstanes, estrenes, and pregnanes, each characterized by unique structure and functions [2]. In the present study, we addressed our investigation toward androgens and estrogens, i.e., two groups of sex hormones belonging to the families of androstanes and estrenes, respectively. In particular, androgens are vital for male sexual development and growth, while estrogens impact female traits, energy metabolism, and mineral balance. Among estrogens, estrone, estriol and 17β-estradiol play crucial roles, the latter being essential for the development of secondary sexual features and mammary glands, while estriol is pivotal in pregnancy and urinary excretion [2–4]. Steroids' multifaceted functions underscore their physiological importance, as the hormone production varies with age and reproductive state.

Sex hormones also appear to play a significant role in the aetiology of breast cancer [4,5]. Several recent studies targeted the role of estrogen metabolites while clinical evidence already classifies the main free-circulating estrogens as candidate breast carcinogens [5,6]. Although the exact mechanism is not fully elucidated, estradiol, estrone and catecholic estrogens (the hydroxylated metabolites of estradiol and estrone) are thought to exert mitogenic and mutagenic effects by interacting with DNA and damaging it, resulting in cellular alterations that can degenerate into carcinogenesis [3]. Conversely, the potential role of androgens in breast cancer development is controversial: according to the literature, androgens stimulate the growth of breast cancer but also display growth-inhibitory properties [7]. High levels of androgens such as dehydroepiandrosterone (DHEA), testosterone, androstenedione, and some of their metabolites, are allegedly responsible for an increased risk of developing Estrogen-Receptor (ER)-positive breast cancers, either directly, by stimulating mammary cell proliferation, or indirectly, by acting as substrates in increased synthesis of estrogens, after aromatization reaction in the peripheral or adipose breast tissue [8,9]. Consistent results were obtained particularly for post-menopausal women whose hormone levels, especially estrogens, typically decrease and remain constant, with respect to pre-menopausal women, whose basal estrogen levels are relatively low and constant, enhancing the chance of detecting the occurrence of anomalous and potentially pathological levels [10]. Overall, a detailed analysis of estrogens and androgens profiles in women affected by breast cancer are likely to clarify the carcinogenesis mechanisms for both diagnostic and prognostic purposes.

We developed and validated a new analytical method based on ultra-high performance liquid chromatography (UHPLC) and time-of-flight high-resolution mass spectrometry (QTOF-HRMS) and devoted to the determination of large urinary steroid profiles, including androgens, estrogens, and their sulfate- and glucuronide-metabolites (20 targeted analytes). The method was subsequently applied to urine samples from post-menopausal women, 65 healthy and 185 with breast cancer, to investigate the relationships between estrogen and androgen levels and the incidence of breast cancer. The results were interpreted using several machine learning algorithms, whose efficiency in discriminating the subjects affected by breast cancer from healthy individuals was compared, to yield a final simple and effective classification model.

2. Materials and methods

2.1. Reagents and standards

All 20 steroid standards and 5 Internal Standards (IS) were purchased as pure powders from Sigma-Aldrich (Milan, Italy), Steraloids Inc. (Newport, RI, USA), or LGC Standards GmbH (Wesel, Germany). The list of the steroids and the internal standards is the following: 16α-hydroxyestrone, 4-methoxyestradiol, 16-epiestriol, 2-methoxyestradiol, 4-methoxyestrone, α-estradiol, β-estradiol, estriol, estrone, androstendiol, androstenedione, androsterone, dehydroepiandrosterone (DHEA), testosterone, DHEA-glucuronide (DHEA-G), DHEA-sulphate (DHEA-S), epitestosterone-glucuronide (epitestosterone-G), testosterone-glucuronide (testosterone-G), testosterone-sulphate (testosterone-S), estrone-β-glucuronide (estrone-β-G), 17β-estradiol-d4, estrone-d4, testosterone-d3, testosterone-glucuronide-d3, androsterone-sulphate-d4. Methanol, methyl tert-butyl ether (TBME), dansyl chloride, sodium hydroxide, sodium phosphate, sodium bicarbonate, synthetic urine and β-glucuronidase from *Escherichia coli* were provided by Sigma-Aldrich (Milan, Italy). Sodium hydroxide (reagent grade) and sodium acetate (reagent grade) were purchased from Fisher Scientific (Fair Lawn, NJ). Ultra-pure water was obtained using a Milli-Q® UF-Plus 6 apparatus (Millipore, Bedford, MA, USA). All stock standard solutions were prepared in methanol at 1 mg/mL and stored at −20 °C until used. Three working solution mixtures were prepared by dilution: one for estrogens at 500 μg/mL (MIX I), one for androgens and conjugated at 500 μg/mL, except androsterone added at the final concentration of 2500 μg/mL (MIX II) and one for internal standards at 500 μg/mL (ISTD). The solutions were stored at 4 °C. The internal standards were used for the different working solution mixtures as follows: 17β-estradiol-d4 and estrone-d4 for MIX I; testosterone-d3, testosterone-glucuronide-d3 and androsterone-sulphate-d4 for MIX II (Table 1).

2.2. Samples collection and pre-treatment

Urine samples were collected from 250 patients, including 65 (26%) from healthy, volunteer post-menopausal women and 185 (74%) from post-menopausal women with diagnosis of breast cancer performed on core biopsy samples, examined at the Candiolo Cancer Institute– FPO IRCCS (Torino, Italy). The urine samples, collected before surgical intervention, were stored at −80 °C. For method development and validation, spiked synthetic urine was fortified with two working solutions (MIX I and MIX II) at six concentration levels, as indicated in Table 1. Synthetic and real urine samples were processed with identical procedures and instrumental conditions.

Two aliquots were collected from each thawed urine sample. Androgens and steroid conjugates were determined directly on one aliquot (100 μL), after dilution with ultrapure water (1:2) and fortification with the IS mixture, without any pre-analytical treatment (dilute-and-shoot). The diluted aliquot was centrifuged at 13.3 rpm for 5 min, and 5 μL supernatant was injected into the UHPLC-MS/HRMS.

The urine aliquot used for estrogen determination (1.5 mL) was fortified with the IS solution at a final 5 ng/mL concentration. The pH was adjusted to 6.8–7.4 adding 0.5 mL phosphate buffer 0.1 M. Enzymatic hydrolysis was conducted by adding 25 μL β-glucuronidase and incubating the aliquot at 58 °C for 1 h. After cooling, 0.5 mL carbonate buffer 0.1 M and drops NaOH 1 M was added to reach a final 9–9.5 pH. Liquid-liquid extraction was performed with 2.5 mL of TBME; the mixture was shaken in a multi-mixer (10 min), centrifuged at 4000 rpm (5 min) and the organic supernatant was transferred into a glass tube. The extracts were dried under nitrogen at 70 °C. The residue was reconstituted with 50 μL carbonate buffer 0.1 M and 50 μL dansyl chloride 1 mg/mL in acetone, to convert the free estrogens into the corresponding dansyl derivatives. The reaction was allowed to proceed at 60 °C for 6 min. Lastly, 5 μL of the supernatant was injected into the UHPLC system.

2.3. Instrumentation

UHPLC separation was performed with a SCIEX ExionLC™ AC system equipped with a Phenomenex Kinetex C18 column (100 × 2.1 mm, 1.7 μm) maintained at 45 °C. The mobile phases consisted of water (A) and acetonitrile (B), both with formic acid 5 mM. The LC flow rate was set at 0.5 mL/min, and the mobile phase eluted under the following linear gradient conditions: (A:B, v-v) isocratic elution at 95:5 for 0.5 min, from 95:5–0:100 in 9 min, isocratic elution at 0:100 for 0.5 min and final re-equilibration for 1.5 min to the initial condition. The total run time was 11 min. All analyses were performed using a quadrupole/

Table 1
Working mixtures, target analytes, chemical formula, molecular weight (after derivatization with dansyl chloride), precursor ion mass, fragment ion mass, UHPLC-MS retention times, internal standards used for quantification and concentration levels of the target analytes used to build the calibration curves.

	Target analyte	Chemical formula	Molecular weight (after derivatization*)	Precursor ion	Fragment ion	Rt	Internal standard
MIX I	16α-hydroxyestrone	$C_{18}H_{22}O_3$	519.2*	520.2158	171.1048	6.93	estrone-d4
	4-methoxyestradiol	$C_{19}H_{26}O_3$	535.2*	536.2471	171.1048	7.68	17β-estradiol-d4
	16-epiestriol	$C_{18}H_{24}O_3$	521.2*	522.2314	171.1048	6.99	17β-estradiol-d4
	2-methoxyestradiol	$C_{19}H_{26}O_3$	535.2*	536.2471	171.1048	7.60	17β-estradiol-d4
	4-methoxyestrone	$C_{19}H_{24}O_3$	533.2*	534.2314	171.1048	7.86	estrone-d4
	α-estradiol	$C_{18}H_{24}O_2$	505.2*	506.2365	171.1048	7.83	17β-estradiol-d4
	β-estradiol	$C_{18}H_{24}O_2$	505.2*	506.2365	171.1048	7.71	17β-estradiol-d4
	estriol	$C_{18}H_{24}O_3$	521.2*	522.2314	171.1048	6.48	17β-estradiol-d4
	estrone	$C_{18}H_{22}O_2$	503.2*	504.2209	171.1048	7.86	estrone-d4
MIX II	androstenedione	$C_{19}H_{26}O_2$	286.2	287.2011	97.0653	5.15	testosterone-d3
	androsterone	$C_{19}H_{30}O_2$	290.2	291.2324	255.2113	5.72	testosterone-d3
	DHEA	$C_{19}H_{28}O_2$	288.2	289.2168	253.1956	5.20	testosterone-d3
	testosterone	$C_{19}H_{28}O_2$	288.2	289.2168	97.0653	4.96	testosterone-d3
	androstenediol	$C_{19}H_{30}O_2$	290.2	291.2324	151.1123	4.85	testosterone-d3
	estrone-β-glucuronide	$C_{24}H_{30}O_8$	446.2	445.2013	269.0662	3.84	testosterone-G-d3
	DHEA-glucuronide	$C_{25}H_{36}O_8$	464.2	463.2483	75.0082	4.07	testosterone-G-d3
	testosterone-glucuronide	$C_{25}H_{36}O_8$	464.2	463.2483	75.0082	3.95	testosterone-G-d3
	DHEA-sulphate	$C_{19}H_{28}O_5S$	368.2	367.1730	96.9596	4.69	androsterone-S-d4
	epitestosterone-sulphate	$C_{19}H_{28}O_5S$	368.2	367.1730	96.9596	4.24	androsterone-S-d4
	testosterone-sulphate	$C_{19}H_{28}O_5S$	368.2	367.1730	96.9596	4.16	androsterone-S-d4
Calibration level	CAL 1	CAL 2	CAL 3	CAL 4	CAL 5	CAL 6	
Mix I and II (ng/mL)	1	2	5	10	25	50	
Androsterone (ng/mL)	10	25	50	75	125	250	

time-of-flight SCIEX X500R QTOF mass spectrometer (Sciex, Darmstadt, Germany) equipped with a Turbo VTM ion source operating in electrospray positive-ion mode (for the ionization of androgens and estrogens) and in the negative-ion mode (for steroid conjugates). Data acquisition involved a preliminary TOF-MS high-resolution full scan followed by a SWATH™ acquisition protocol, which used a variable window setup: 12 windows covering the mass range from *m/z* 249.5–470.0 for androgens, 12 windows from *m/z* 244.9–470.0 for the conjugates, and 16 windows from *m/z* 441.0–760.0 for estrogens at 0.025 resolving power. The target analytes identification was based on the coincidence of their retention times, precursor ion and characteristic fragment ion *m/z* values (accepted mass error <5 ppm) (Table 1). Data were acquired using the SCIEX OS 1.5 Software.

2.4. Method validation

The validation strategy was based on a protocol recently published [11]. Nine independent replicated analyses prepared in synthetic blank urine at each concentration level (6 levels) were executed on three different days, resulting in three calibration data points collected on day 1, three on day 5, and three on day 12. This dataset of 54 analyses formed the groundwork on which the statistical evaluation of several validation parameters was based, including precision, accuracy, limit of detection (LOD), limit of quantification (LOQ), ion abundance ratio repeatability, selectivity, specificity, and carry-over (adding blank samples after the analysis of the samples spiked at the upper limit of quantification). Recovery, matrix effect and stability parameters were evaluated with further independent experiments. An *ad hoc* Excel® sheet was built in-house to adapt the routine developed by Desharnais et al. [12]. All the equations employed to compute the validation parameters can be found elsewhere [13].

2.4.1. Linearity

Calibration curves were generated from the peak-area ratio between the quantifier transition for each analyte and that of the corresponding ISTD; the ratio was then plotted on the y-axis against the hormone concentration, from which the curve that best fits and predicts the data distribution was computed with the support of statistical tests [14]. Three replicates of the calibration curve were prepared by spiking synthetic urine at 6 concentration levels in compound-specific ranges (Table 1) and assessed using a weighted regression model. The heteroscedasticity of the data-points distributions was checked to evaluate whether a weighting factor of $1/x$ or $1/x^2$ was needed in the calibration model calculation using least squares regression. Then, the order of the calibration model (linear vs. quadratic) was selected based on Mandel's and lack-of-fit tests, using a significance level of 95%.

2.4.2. Limit of detection (LOD) and limit of quantitation (LOQ)

The limit of detection (LOD) was determined as the minimal detectable analyte concentration in a sample, generating a signal substantially above background noise [15]; the Hubaux-Vos' algorithm, adjusted for heteroscedastic data using Currie's weighting correction [15] was used for LOD calculation. Calculated LOD values were verified experimentally by spiking blank matrices with the target analytes at the calculated LOD concentrations, confirming that a signal-to-noise ratio (S/N) exceeding 3 was obtained. The limit of quantitation (LOQ) was established as the lowest analyte concentration which could be quantified with predetermined and acceptable precision and accuracy [15]. Trueness was considered acceptable for bias% $<\pm20\%$.

2.4.3. Accuracy and precision

Accuracy represents the agreement between an analytical result and the accepted true or reference value, in our case the concentration obtained by spiking. Since our data collection involved the repetition of nine analytical sequences at six concentration levels, accuracy was estimated by back-calculation, i.e., the results from one sequence were elaborated using a calibration obtained from the remaining data-point sequences. In particular, intra-day accuracy was computed by excluding one calibration sequence at a time out of the three sequences of each validation day. This process was repeated over three validation

days, and the average bias (%) was determined. Then, inter-day accuracy was computed following a similar process by repeatedly using two-day calibration sequences for evaluating the third-day data. Intra- and inter-day accuracy was considered validated if below 20% for the averaged calibration levels. Single values were considered "good" in the range ±15% and "acceptable" in the range ±20%, while random values exceeding ±20% required attention but did not exclude validation if the primary condition (<20% for the average) was respected.

Precision, including both repeatability and reproducibility, was estimated from the coefficient of variation (CV%) for repeated determinations. Intra-day precision was assessed independently over three validation days using three daily replicates, while inter-day precision utilized all nine replicates back-calculated from independent calibration data.

2.4.4. Matrix effect and extraction recovery

Matrix effect (ME) [16] was assessed by comparing the experimental results obtained from synthetic blank urine samples and blank deionized water samples, equally spiked after the extraction step [11]. Three replicates at low, medium, and high concentration levels, corresponding to CAL 2, CAL 4 and CAL 5 were made. The ionization suppression or enhancement for each target analyte was expressed as the mean percentage ratio between the two measured signals.

The extraction recovery was determined by comparing the experimental results obtained from blank synthetic urine samples spiked before and after the pre-analytical sample handling and extraction steps (the second spiking was made before the derivatization step). As for the matrix effect, three replicates were made at low, intermediate, and high concentrations. The results were expressed as the mean percentage ratio between the two signals, with its uncertainty expressed as extraction repeatability (CV%) from the three replications.

2.5. Multivariate data analysis

2.5.1. Benchmarking

Seven different supervised classification algorithms were tested and their efficiency in the identification of valuable biomarkers was compared for discriminating the patients with breast cancer from the healthy ones. The tested machine learning (ML) algorithms included Logistic Regression (LR) [17], Naive Bayes (NB) [18], Partial Least Squares – Discriminant Analysis (PLS-DA), k-Nearest Neighbors (k-NN) [19], Classification and Regression Trees (CART) [20], Support Vector Machine (SVM) [21], and Random Forest (RF) [22], the latter yielding the most reliable output. A brief description of the tested ML models is reported in the Supplementary Material.

The data collected from real urine samples (250 samples × 20 analytes) were initially pre-processed by autoscaling all the features (analytes); then, 10-fold cross-validation was performed on the training data to avoid overfitting and get a more robust estimate of the models' performance during the benchmark process. Stratification was performed during the data-splitting process to ensure that the class distribution in the target variable was preserved in the split datasets. In particular, a randomized stratification approach was performed by sorting the instances in a random order firstly, and then assigning them to either the training or testing set while keeping the class proportions approximately the same. To stress the robustness of the developed models, the data-splitting process was repeated for ten times.

Accuracy and balanced accuracy, together with their standard deviation, were evaluated to compare the performance of the different models using the following formulas:

$$Sensitivity \quad (Recall) = \frac{TP}{TP + FN} \quad Specificity = \frac{TN}{TN + FP} \quad Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad Balanced \quad accuracy = \frac{Sensitivity + Specificity}{2}$$

where *TP* represents the number of patients positive to breast cancer that are correctly classified as positive (i.e., true positive), *TN* stands for the number of patients negative to breast cancer that are correctly classified as negative (i.e., true negative), *FP* is the number of patients negative to breast cancer that are incorrectly classified as positive (i.e., false positive), and *FN* is the number of patients positive to breast cancer that are incorrectly classified as negative (i.e., false negative) [23].

2.5.2. Random Forest (RF) modelling

Following the benchmark process, the selected RF approach was optimized to further improve its classification performance. The collected data were split into a training set and a test set with a split ratio equal to 0.8, providing a training set of 200 patients (i.e., 80% of the available data) and a test set of 50 patients (i.e., 20% of the available data). Stratification was performed during the data-splitting process.

Subsequently, a grid-search approach was performed to exhaustively inspect different combinations of RF hyperparameter values, train the RF models with each combination, and report their performance using the balanced accuracy metric [22]. The following RF hyperparameters were tuned in cross-validation with k=5, as follows:

- *max_features*: this parameter determines the maximum number of features to consider when looking for the best split in each decision tree of the RF model. Two types of settings were selected, such as *'n'*, which means that the total number of *n* features (in our dataset, n=20 free and conjugated sex hormones), and *'sqrt'*, which is equal to the square root of the total number of *n* features in the dataset (i.e., 4.47, in this case, round up to 5). *'sqrt'* was selected after tuning.
- *n_estimators*: this value defines the number of decision trees (estimators) used in the RF model. The values tested, such as 10, 100, and 1000, stand for different choices for the number of trees created and combined to make predictions. A number of estimators equal to 1000 was selected.
- *max_depth*: it controls the maximum depth of each decision tree in the RF model by limiting the number of nodes from the root to the deepest leaf of the tree. The values provided, such as 0, 2, 5, and 10, indicate different depth constraints for the trees. A maximum depth of 5 was selected.
- *min_samples_split*: it sets the minimum number of samples required to split an internal node of each decision tree. The values evaluated, such as 2, 5, and 10, represent different thresholds for the minimum number of samples required for a node to be split. The optimal number of split was set to 5.
- *min_samples_leaf*: this hyperparameter sets the minimum number of instances (samples) required at a leaf node of each decision tree. The tested values (i.e., 1, 2, and 4) show different thresholds for the minimum number of samples required at each leaf. The optimal minimum number of samples was set to 4.
- *bootstrap*: this parameter controls whether the RF model should build trees using bootstrapped samples. Setting it to *'True'* allows for bootstrapping, while setting it to *'False'* means that the entire dataset is used for building each tree [24]. Bootstrap was used in the benchmarking step of our RF models computation, not in the final refinement described in this chapter.

The optimal, final trained RF model was evaluated using the Receiver Operating Characteristic (ROC) curve, the precision-recall curve (PRC), and the confusion matrix.

Subsequently, the study explored the feature importance provided by the RF model in terms of Gini importance, which measures the total reduction of impurity (or weighted Gini impurity) across all decision

trees of the RF model attributed to a specific feature [25]. Therefore, the Gini impurity for each class is given by $1 - (p_0^2 + p_1^2)$, where p_0 is the probability of a negative patient being classified in either class defined for breast cancer, and p_1 is the corresponding probability for a positive patient. Higher Gini importance indicates greater predictive significance of the feature. The results provided by Gini feature importance were confirmed using a complementary approach known as SHAP (SHapley Additive exPlanations) Values. This method is rooted in the cooperative game theory developed by Shapley [26,27], which assigns a value to each player in a game based on their contribution to different coalitions and, in this context, offers a comprehensive understanding of the importance of a particular feature in the RF model [28] by quantifying how much each feature contributes to the overall decision-making process of the ensemble model.

2.5.3. Software

The data analysis was performed using Python version 3.10.8 with essential libraries, including numpy, pandas, scikit-learn [24], seaborn, matplotlib, and shap [28]. These libraries provided efficient tools for data manipulation, machine learning implementation, and visualization, ensuring a smooth and effective analysis workflow. The complete list of references is provided in the Supplementary Material.

3. Results and discussion

3.1. Method validation

3.1.1. Linearity

Residues and variances of calibration data points across low, medium, and high concentration levels revealed the occurrence of heteroscedastic distributions for all target analytes, making the introduction of weighting factors in the calibration (either $1/x$ or $1/x^2$) beneficial. Additionally, the statistical significance of the quadratic term in the calibration model was observed for all substances under evaluation. Consequently, a quadratic calibration model was chosen for all analytes combined with a weight of $1/x^2$, except for DHEA, testosterone, and androsterone, whose weights were $1/x$. The calibration curve equations used for each analyte are provided in Table S1 of the Supplementary Material.

3.1.2. LOD and LOQ

Calculated LOD values ranged from 0.50 ng/mL for estrone-β-G up to 3.4 for DHEA. The LOD value of 12.5 ng/mL calculated for androsterone partly depends on the different calibration range adopted to best fit its physiological urinary concentration. The calculated LOD values (weighted Hubaux-Vos method) were experimentally verified by spiking blank matrices at the approximate LOD concentrations for the different analytes, confirming a S/N ratio exceeding 3. Table 2 reports the mean values obtained for LOD. The mean values are calculated by averaging the six values obtained for the corresponding concentration levels, which are reported in detail in in Table S1 in the Supplementary Material.

With the exception of estriol, DHEA, and androsterone, all analytes exhibited LOD values falling below the lower limit of the calibration curve, which in turn was used as LOQ.

3.1.3. Accuracy and precision

The validation procedure adopted in the present study allows the calculation of precision and accuracy at all concentrations involved in the calibration process (6 calibration levels, except for 16α-hydroxyestrone with 5 calibration levels), not only at low, intermediate and high concentrations, as most validation protocols recommend. Table 2 reports the mean values obtained intraday and interday accuracy (expressed as bias %) and intraday and interday precision (expressed as CV%). The corresponding values for each concentration and target analyte are reported in Table S2 and Table S3 of the Supplementary Material.

All conjugated analytes showed intra-day and inter-day accuracy values below 15%. Among the free-form androgens, only DHEA and androsterone exhibited values exceeding 20% at the lowest calibration point. Among the remaining androgens, only four bias values resulted between 15% and 20%, while the other outcomes were all below 15%. Similarly, the analytical procedure for estrogens also yielded optimal bias% values (<15%), with only 16-epiestriol showing occasional bias exceeded the ±20% limit. Overall, the accuracy of both analytical

Table 2
Mean values obtained for LOD, intraday accuracy, interday accuracy, intraday precision, interday precision, matrix effect, and extraction recovery. The mean values are calculated by averaging the data obtained for the six (three for matrix effect and extraction recovery) concentration levels under study. The specific values for each concentration level are reported in Tables S2-S5 of the Supplementary Material.

		Mean values						
	Target analyte	LOD (ng/mL)	Intraday Accuracy (Bias %)	Interday Accuracy (Bias %)	Intraday Precision (CV%)	Interday Precision (CV%)	Matrix Effect %	Extraction Recovery %
MIX I	16α-hydroxyestrone	1.38	2.10%	10.76%	14.51%	1.29%	98.20%	81.70%
	4-methoxyestradiol	0.71	-0.67%	0.02%	6.62%	13.88%	100.20%	105.50%
	16-epiestriol	0.92	-5.95%	-0.76%	11.70%	19.63%	115.34%	102.70%
	2-methoxyestradiol	0.73	-0.50%	2.23%	6.21%	20.83%	113.76%	101.10%
	4-methoxyestrone	0.62	0.17%	0.14%	5.55%	14.53%	93.59%	78.50%
	α-estradiol	0.56	-0.11%	0.39%	5.16%	14.58%	90.92%	85.00%
	β-estradiol	0.70	0.04%	0.65%	5.09%	15.24%	97.80%	91.50%
	estriol	1.31	0.40%	4.71%	14.45%	32.64%	124.79%	109.30%
	estrone	0.74	-1.63%	-0.32%	5.46%	16.08%	76.71%	110.10%
MIX II	androstenedione	0.71	-3.74%	-0.12%	6.90%	7.99%	96.06%	
	androsterone	12.5	0.30%	3.10%	0.30%	3.10%	95.50%	
	DHEA	3.40	3.47%	6.29%	11.30%	13.05%	102.95%	
	testosterone	1.34	1.37%	-8.41%	5.83%	5.92%	103.15%	
	androstenediol	1.20	-2.00%	0.03%	8.69%	13.62%	88.01%	
	estrone-β-G	0.50	1.06%	0.26%	6.41%	8.71%	98.43%	
	DHEA-G	0.75	-1.93%	0.28%	4.00%	9.33%	96.01%	
	testosterone-G	0.84	-2.13%	0.73%	4.48%	11.21%	96.47%	
	DHEA-S	1.08	-0.57%	0.63%	5.79%	12.17%	89.66%	
	epitestosterone-S	1.16	0.88%	0.89%	5.72%	12.45%	95.58%	
	testosterone-S	0.81	-2.99%	0.17%	6.18%	9.48%	88.73%	

methods proved satisfactory for the determination of androgens, estrogens, and their conjugates, with the exception of the lowest calibration point for DHEA and androsterone.

The precision results reported in the Supplementary Material (Table 4 S) show optimal values for the conjugated steroids and androgens, with only three scattered values out of 108 slightly exceeding 20%. On average, higher variations were observed for estrogens, as a consequence of the more complex sample pre-treatment, also involving a derivatization step, in particular for estriol interday precision (21%< CV%<41%) and 2-methoxyestradiol interday precision (8%<CV%< 32%).

3.1.4. Matrix effect and extraction recovery

The results relative to the evaluation of the matrix effect for the studied analytes at low, medium, and high concentration levels are reported in Table 2 and Table S4 of the Supplementary Material. All the reported percentages are close to 100%, not evidencing any significant suppression or increase effect on the analytical signal. This was predictable, since limited effect from the of synthetic urine components is expected. Modest ion suppression was observed for 4-methoxyestrone and estrone, particularly at high concentrations.

Since the procedure used for androgens and conjugated did not involve any extraction, purification, or concentration steps, no extraction recovery was determined. For estrogens, the results obtained for the estimation of the extraction recovery are reported in Table 2 and Table S5 of the Supplementary Material. The mean percentage of recovery for the various estrogens varied in the 78% - 110% range. The overall mean value was 96.6%, with lower percentages (~93%-94%) observed for the lower concentrations. The lowest recoveries (~80%) were obtained for 4-methoxyestrone and 16α-hydroxyestrone. Overall, the recovery achieved by the procedure is satisfactory.

3.2. Benchmarking

In the domain of multivariate data analysis, gaining comprehensive insights into both data structure and algorithm performance is essential for addressing the supervised classification objectives. The use of a model selection and comparison process, commonly referred to as benchmarking, proves beneficial in shedding light on the estimated accuracy of various machine learning models. The results obtained from the benchmarking process provided a comprehensive comparison of the different supervised classification models selected. Fig. 1 shows the boxplots of the accuracies (A) and the balanced accuracies (B) provided by each model built using a 10-fold cross-validation strategy. In particular, the orange line represents each model's median (balanced) accuracy, and the circles outside the boxplots are anomalous data.

The median accuracy is homogeneously around 0.8 (Fig. 1A), even if the PLS-DA model exhibit the lowest accuracy while RF and CART models the highest. In contrast, Fig. 1B displays more varied outcomes concerning median balanced accuracy: PLS-DA and SVM models yielded low values (around 0.5–0.6), whereas the RF and NB models delivered superior accuracy around 0.85. The observed differences between accuracy and balanced accuracy results stem from a significant class disparity (185 positive patients vs. 65 negative patients) within our dataset, suggesting a different weighting for our binary classification problem. In this scenario, the use of balanced accuracy is known to mitigate the bias introduced by class distribution and deliver a more objective evaluation of the model by assessing accuracy for each individual class and then averaging the outcomes. In our case, where one class (positive patients) prevails over the other (negative patients), this method assigns equal importance to them, helping to unravel the strengths and limitations of different algorithms. Table S6 in the Supplementary Material summarises the collected results, showing the median accuracy and the median balanced accuracy values of each model, together with their standard deviation.

Among the best classifiers, we chose to further develop RF models rather than NB, since NB is prone to glitches stemming from collinearity among the variables, which is likely to introduce instability into the NB model and reduce its reliability. The higher instability of NB is evident in Fig. 1A by the high standard deviation of NB accuracy and in Fig. 1B by the occurrence of two outliers. In the case of NB, multicollinearity may lead to inflated standard errors and hinder the accurate estimation of model coefficients [27]. In contrast, RF is fundamentally resilient to the challenges posed by multicollinearity. In fact, RF is a powerful and versatile machine learning ensemble method, working on a multitude of decision trees during the training phase, each trained on a random subset of the data (bootstrap aggregation) and a random subset of features (feature randomness). The final prediction is made by aggregating the predictions of all individual trees, typically using majority voting [22].

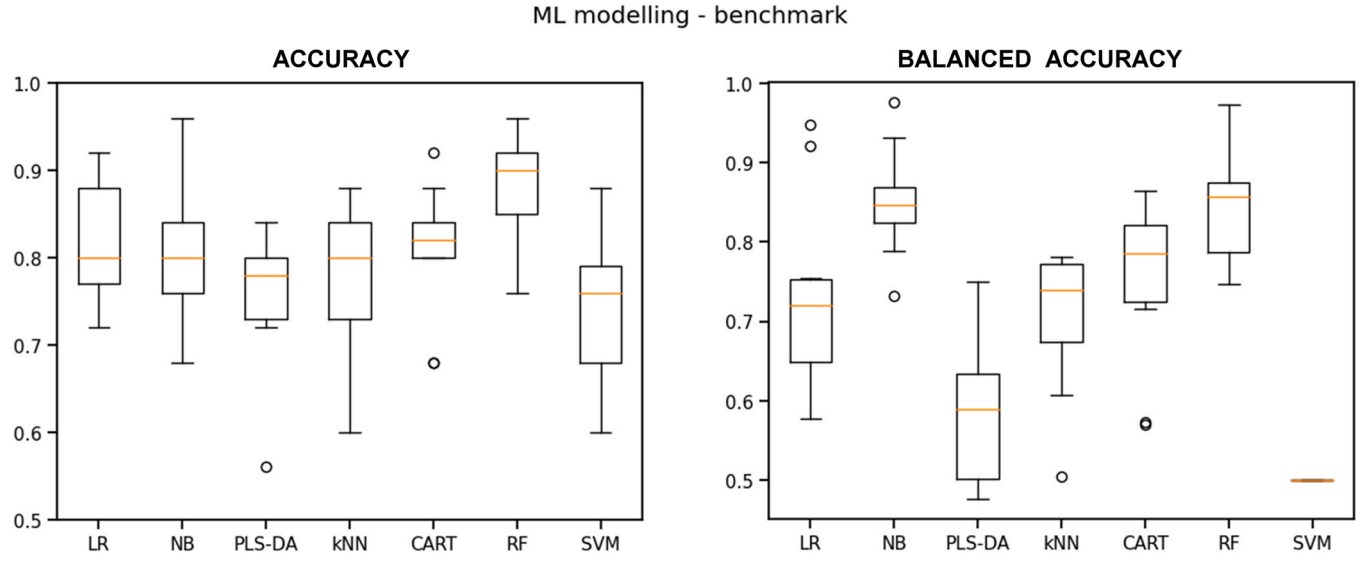

Fig. 1. Benchmark of supervised classification models including Logistic Regression (LR), Naive Bayes (NB), Partial Least Squares – Discriminant Analysis (PLS-DA), k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Classification and Regression Trees (CART), and Random Forest (RF). The y-axis shows the accuracy (A) and the balanced accuracy (B) metrics, the orange line represents the median balanced accuracy of each model, and the circles outside the boxplots represent the anomalous data (outliers).

3.3. Random Forest modelling

The RF model optimization involved a grid-search hyperparameter tuning step aimed to avoid over-fitting by setting appropriate constraints on the model's growth. Limiting the depth of the trees and increasing the minimum number of samples required to split a node ensures that the final RF model is expressive enough to learn from the data and make accurate predictions [22]. The balanced accuracy results obtained for each combination of hyperparameters (reported in Section 2.5.2) was used to identify the hyperparameter values that yielded the best performance.

The grid-search hyperparameter tuning resulted in a remarkable improvement of the Area Under the Curve (AUC) score on the test set (50 samples) which rose from 0.85 (benchmarking process) to 0.94, indicating a substantial enhancement of the model's predictive ability to distinguish the patients positive to breast cancer from negative ones. Among the crucial variables, limiting the maximum number of features to be considered in the pursuit of the best split in each decision tree (square root of the total = 5 ≅ $\sqrt{20}$) produced a significant improvement of the model, prevented overfitting, added randomness to the model, and yielded faster model computation. Setting the minimum number of samples to be selected for each leaf to 4 (i.e., at least 4 samples have to be selected for each leaf) also prevented the occurrence of overfitting by forcing the decision trees not to consider isolated outliers. The ROC curve and the confusion matrix obtained on the test set after the tuning process are reported in Fig. 2.

The confusion matrix reported in Fig. 2 shows the model's ability to distinguish most of the patients positive or negative to breast cancer, with the exception of three false positives and one false negative.

Fig. 3A shows the importance of different features in terms of weighted Gini importance: the features/variables are listed on the y-axis, and their importance is shown on the x-axis. Since the Gini importance of a feature is calculated by measuring how much the feature decreases the impurity of the trees in the RF model (averaged for all trees), the more a variable decreases the Gini mean impurity, the more influential the variable is in discriminating the patients under exam.

The most important feature reported in this plot is clearly testosterone-S, followed by alpha-estradiol, 4-methoxyestradiol-, DHEA-S, epitestosterone-S, and, to a lower extent, DHEA and androsterone. All these steroids have been previously associated with breast cancer [8]. The remaining features are apparently less important in classifying the patients and could be removed.

In Fig. 3B, the SHAP values of the features under investigation are illustrated as a beeswarm plot, offering a detailed viewpoint on the influence of each feature (sex hormone) on the instances (patients) under exam. In particular, this plot is designed to display an information-dense summary of how the features in a dataset impact the RF model's output. For each patient, the given explanation is represented by a single dot on each feature row, showing the variability and impact range that each feature has on the model's predictions. The x position of the dot is determined by the SHAP value of that feature, while the color display the original value of each feature (red for higher values, blue for low values). As an example, from Fig. 3B it results that testosterone-S is the most important feature and patients with low testosterone-S levels (blue) are more likely to be negative to breast cancer. Conversely, patients with high testosterone-S levels (red) are more likely to be positive to breast cancer. Features with broader distributions or more extreme values indicate higher variability in their contributions to predictions. A further explicative force plot for a positive patient is reported in Fig. 3C, showcasing the features that drive the prediction towards the maximum value (=1) from the expected base value (=0.5), the neutral prediction made by the model representing an equal likelihood of belonging to either class when no specific features are taken into account [28]. Again, the largest positive SHAP value is obtained for testosterone-S, whose high level means that this patient has a high prediction of being positive to breast cancer. The other features with the highest importance and positive SHAP values are 4-methoxyestradiol, epitestosterone-S, alpha-estradiol, and DHEA-S, while the only feature with a negative SHAP value is testosterone-G, meaning that a high level of testosterone-G is associated with a negative prediction to suffer from breast cancer.

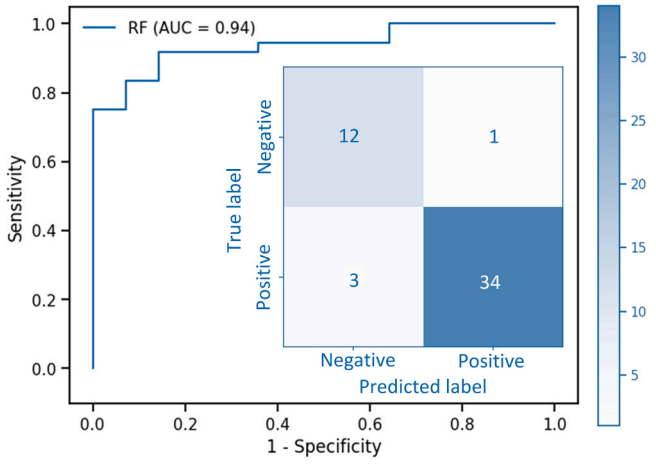

Fig. 2. ROC curve and confusion matrix obtained after the hyperparameter tuning process.

The univariate boxplots reported in Fig. 4 show the five model-shaping sex hormones that exhibit higher concentration values in the patients positive to breast cancer than in healthy women. These boxplots were obtained by considering all the patients available in the dataset. Testosterone-S and epitestosterone-S are phase-2 metabolites produced in the ovaries and adrenal glands, alpha-estradiol is produced in the ovaries, 4-methoxyestradiol is a metabolite of estradiol, and DHEA-S is a DHEA conjugate produced in the adrenal glands. Scattered studies have remarked that women with breast cancer are likely to have higher levels of these sex hormones than women without breast cancer [8,9], as our outcomes confirm. This suggests, again, that these sex hormones play an crucial role in the development or progression of breast cancer [5–7]. The univariate boxplots for all 20 targeted steroids comprising all 250 patients enrolled in this study, are reported in the Supplementary Material (Figure S1). A visual comparison among them provides a rough identification of the most (and the least) influential features, leading to an initial insight into the factors more likely to contribute to a multivariate model prediction. This preliminary knowledge facilitates both feature selection and dimensionality reduction, leading to more efficient and interpretable models [29].

Following the hints provided by the preceding data processing, a final RF model was built using the five most significant features evidenced in Figs. 3 and 4 (testosterone-S, alpha-estradiol, 4-methoxyestradiol, DHEA-S, and epitestosterone-S). The new RF model was tested on the same evaluation set, showing optimal and improved performance metrics. The related ROC curve displays an AUC value equal to 0.97. The confusion matrix reported in Fig. 5 A confirms that the latest RF five-features model is capable of identifying both positive and negative patients correctly, with only one false positive outcome. Comparable results (from 0 to 1 false positive results) were obtained for all ten data-sets obtained after using the stratified data-splitting strategy.

The reduced features set leads to a more efficient classification model with reduced computation time and improved interpretability. The features with negligible importance can be removed from the analytical procedure, reducing the costs, the analytical complexity, the dimensionality of the dataset, and simplifying the model without compromising its performance, but rather improving it, thus facilitating the decision-making process.

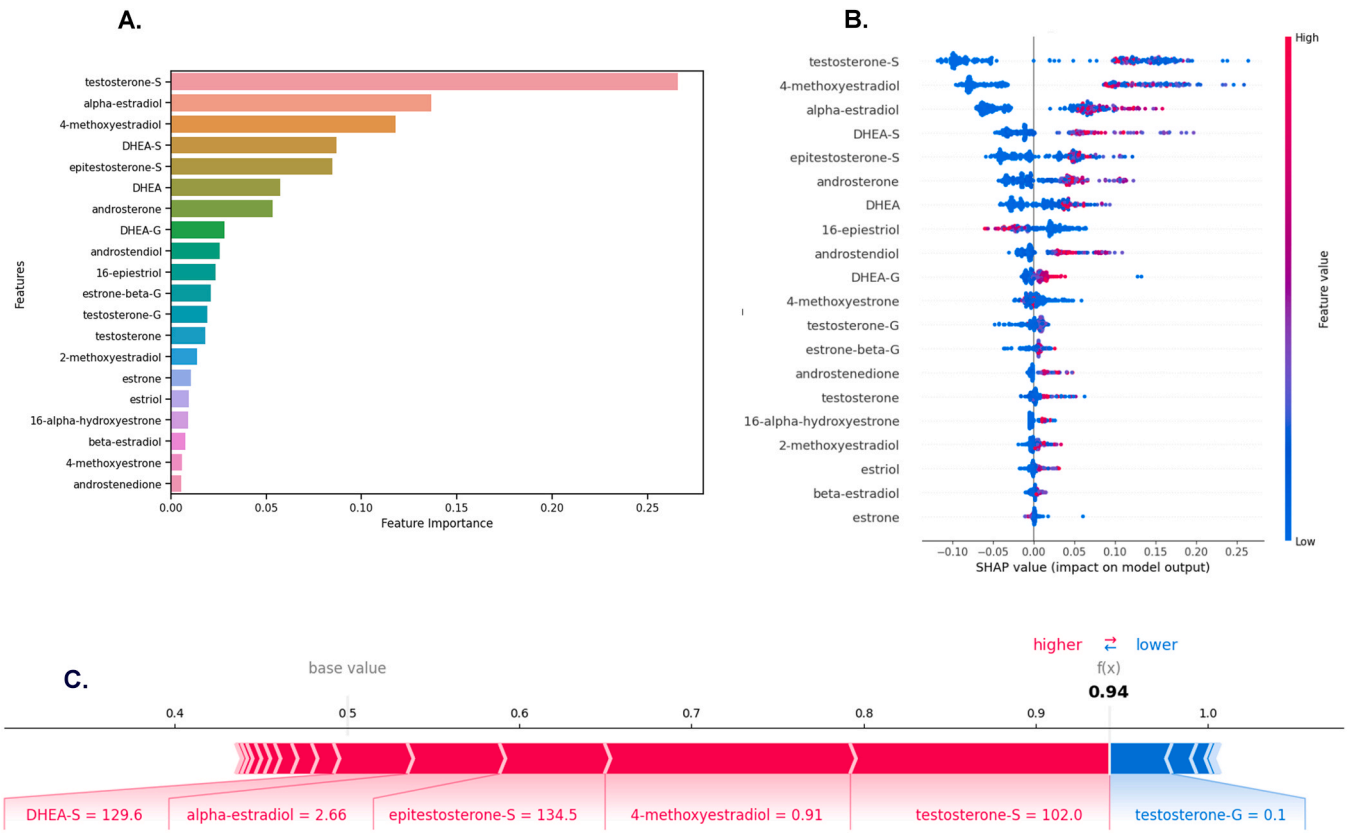

Fig. 3. Importance plots: (A) the variables are listed on the y-axis and their Gini importance is reported on the x-axis; (B) beeswarm plot showing the SHAP values and the features' impact on the RF model and instances; (C) force plot of a positive patient: the length and direction of the bar indicate the feature's impact on the prediction; longer bars have a more substantial influence towards the positive (red) or negative (blue) prediction.

3.4. RF and PLS-DA comparison

Further investigations were made to understand the reasons why RF models overperformed PLS-DA, which is generally considered the golden standard for binary classification. A possible explanation can be found in a Brereton and Lloyd study [30], reporting that PLS-DA may turn fragile when applied to class-imbalanced data. Intriguingly, this vulnerability may persist even when the PLS-DA model is calculated using a "weighted" approach. In our data, this phenomenon was already highlighted in Fig. 1, which depicts the pronounced alterations in the model performance when accuracy is computed with balanced approach (Fig. 1B) instead of a non-weighted (Fig. 1A). Indeed, RF and PLS-DA models provide diverging results in the precision-recall curve [30] and confusion matrices computed on the evaluation set (Fig. 5). Other studies support our findings: for example, Song [31] observed a superior efficiency of RF modelling in comparison with PLS-DA, when both models were tested on datasets with significant class imbalance, such as ECOLI and Glass datasets (references in the Supplementary Material). In our context, the inherent dataset imbalance depended on the recruited patients, arising from a dedicated cancer hospital, where the prevalence of breast cancer patients significantly outweighed the negative cases (non-cancer patients).

RF and PLS-DA models provide diverging results in the precision-recall curve (PRC) computed on the evaluation set. The RF model's superior performance compared with PLS-DA is highlighted by the confusion matrix reported in Fig. 5B, showing that PLS-DA wrongly classifies all negative patients as false positive. In the present case, it is more appropriate to represent the performance features by means of PRC curves (Fig. 5C, D) rather than ROC curves displaying sensitivity vs. 1-specificity [32], since ROC curves are susceptible to class imbalance, particularly when the minority class has small size. Notably, the "precision" of a specific class denotes the level of certainty in the model's predictions when assigning instances to that class. Conversely, the "recall" quantifies the model's ability to identify instances belonging to a given class effectively and emphasizes the classifier's performance concerning the smaller class.

The two class models were investigated further using the Synthetic Minority Oversampling Technique (SMOTE) method as a data simulation algorithm [33]. This method helps when there are unequal numbers of data points in the considered classes. SMOTE creates additional "artificial" data for the less represented class by shaping the new data points just to be similar to the existing ones [34]. By removing class imbalance, SMOTE improves the models and makes predictions more reliable. As a result, a balanced dataset (185 subjects per class) was created, and the RF and PLS-DA models were recalculated using the same 5 variables selected in Section 3.3. As reported in the PRC plot of Fig. 5D, the RF model continues to exhibit robust performance. Conversely, PLS-DA displays improvement with respect to the previous case (Fig. 5C), though it consistently lags behind RF's performance, thus remarking RF superiority in case of class-imbalanced datasets.

While PLS-DA may possibly provide optimal performance under conditions closer to ideal, the synergy between capabilities of SMOTE data simulation and RF ability to accommodate class imbalanced data highlights the efficacy of this strategy in addressing constrained conditions commonly encountered in clinical trials. In such contexts, an imbalanced distribution among collected patients is a common challenge that can be effectively mitigated using this approach [35].

4. Conclusions

Although preliminary, the results of the present research suggest that a limited number of steroid biomarkers in conjunction with adequate RF

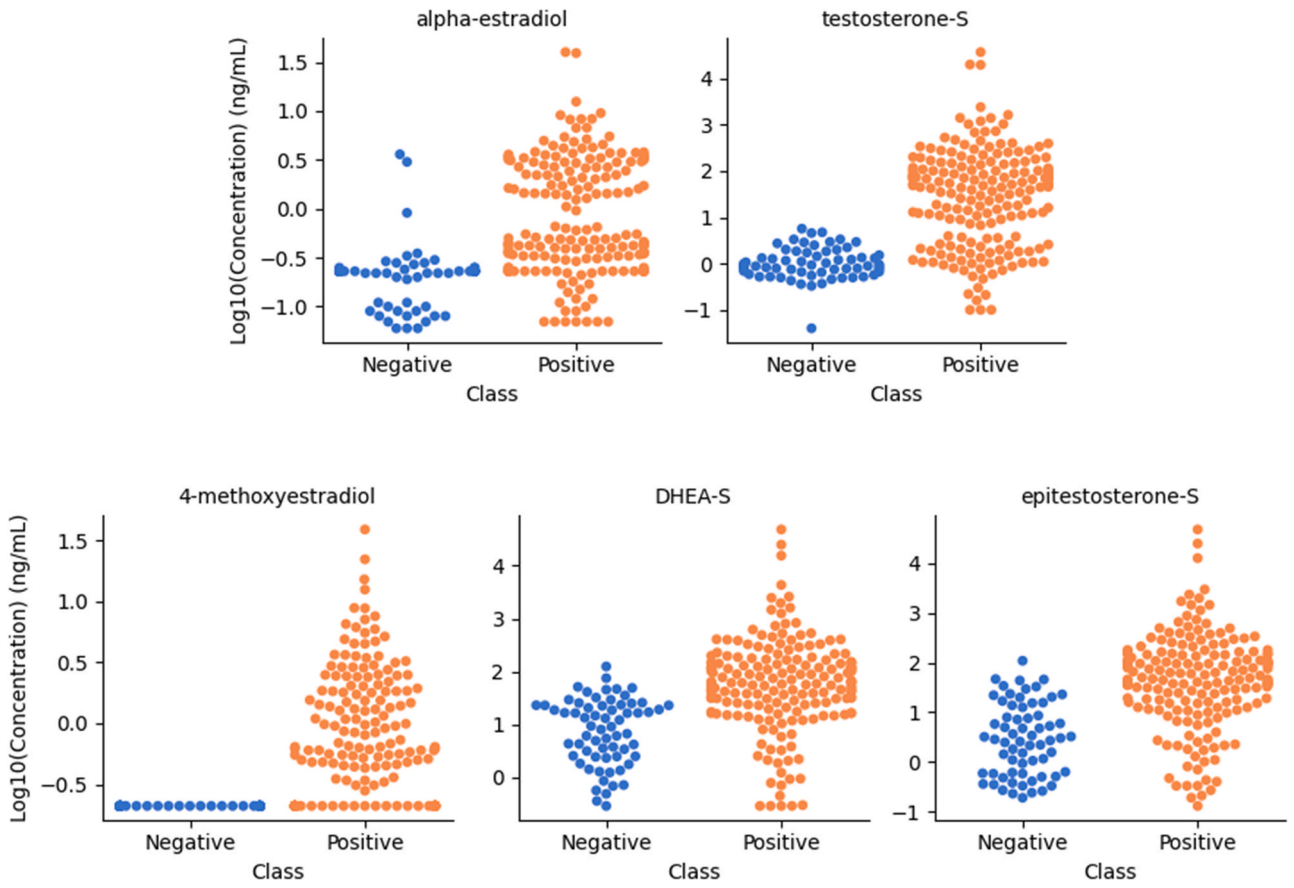

Fig. 4. Boxplots of the selected hormones showing the highest Gini importance. The y-axis shows the $\log_{10}$(concentration) values. Negative and positive classes identify healthy and breast cancer subjects, respectively.

modelling hold promise for early screening of breast cancer. The proposed protocol involves non-invasive sampling and provides improved precision with respect to the routine practice presently adopted. Similar protocols might be addressed to discriminate cancer sub-types and prognosticate cancer evolution, in response of surgical and pharmaceutical treatments, within an individually-specific precision therapy management.

Among the five selected biomarkers, the detection of two free estrogens requires a specific and quite complex analytical method in order to cope with their low urinary concentration. Their possible substitution in the model with their (or other) conjugate metabolites represents a forthcoming step of our research, so as to simplify the overall analytical procedure and extend it to a larger population. Indeed, conjugated steroids are relatively underexplored biomarkers in the existing scientific literature despite their potential significance.

Research ethics and informed consent

The present study protocol was discussed and approved by the ethical committee of Candiolo Cancer Institute– FPO IRCCS (protocol n. CE IRCCS 28/2019 - available upon request). The present research was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and observed ethical guidelines and international regulations, protecting the rights and well-being of the participants involved.

Informed consent was obtained from all patients enrolled in this study, indicating their voluntary agreement to be part of this research and to allow disclosure of the collected data for the advancement of scientific knowledge, together with their full awareness of the study's purpose, procedures, potential risks, and benefits. Moreover, each patient received a study information sheet indicating all the parameters that allowed to shape the pathological context and define the clinical and medical history of the patient.

Funding

This work was supported by (a) the Italian Ministry of Education, Universities, and Research - PRIN2017 project Prot. 2017Y2PAB8, entitled "Cutting Edge Analytical Chemistry Methodologies and Bio-Tools to Boost Precision Medicine in Hormone-Related Diseases," and (b) the Fondazione CRT - Grant Prot. RF=2022.1831 "Screening del carcinoma mammario mediante metabolomica urinaria, profilamento ormonale, machine learning".

CRediT authorship contribution statement

Caterina Marchiò: Writing – review & editing, Supervision, Resources, Data curation. **Marco Vincenti:** Writing – review & editing, Writing – original draft, Supervision, Resources, Conceptualization. **Anna Sapino:** Writing – review & editing, Supervision, Conceptualization. **Alberto Salomone:** Validation, Methodology. **Daniele Di Corcia:** Validation, Methodology. **Riccardo Ponzone:** Supervision. **Enrico Berrino:** Resources, Investigation. **Fulvia Trapani:** Validation, Investigation, Data curation. **Eugenio Alladio:** Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Marta Massano:** Validation, Investigation. **Lorenzo Castellino:** Software, Methodology, Formal analysis.

Declaration of Competing Interest

The authors declare the following financial interests/personal

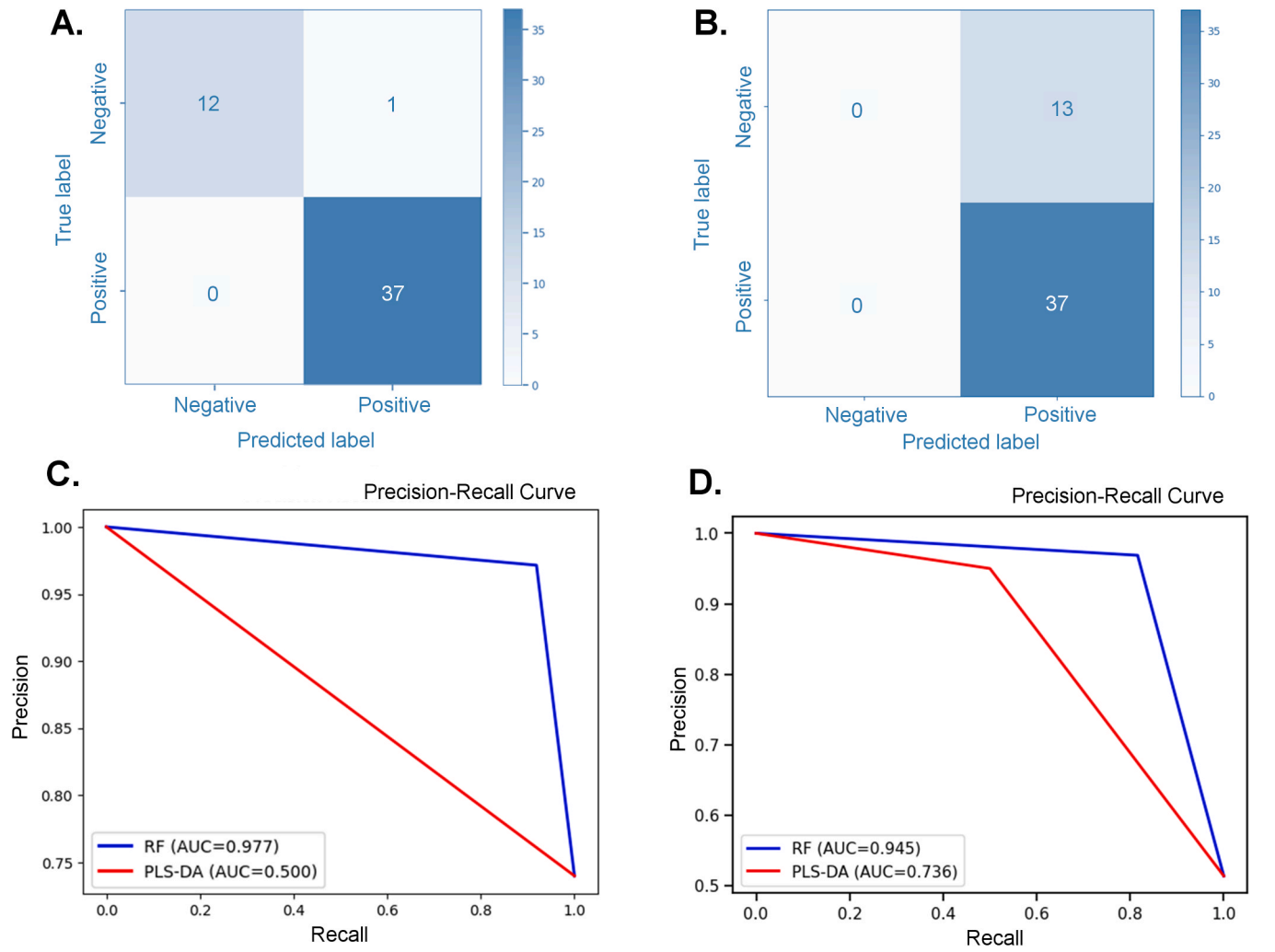

Fig. 5. (A) Confusion matrix of the final RF model; (B) Confusion matrix of the PLS-DA model; (C) PRC for RF (blue) and PLS-DA (red) models before SMOTE data simulation; (D) PRC for RF (blue) and PLS-DA (red) models after SMOTE data simulation. Negative and positive classes identify healthy and breast cancer subjects, respectively.

relationships which may be considered as potential competing interests: Marco Vincenti reports financial support was provided by CRT Foundation. Marco Vincenti reports financial support was provided by Italian Ministry of Education, Universities, and Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research acknowledges support from the Project CH4.0 under the MUR program "Dipartimenti di Eccellenza 2023–2027" (CUP D13C22003520001).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jpba.2024.116113.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, Ca. Cancer J. Clin. 71 (2021) 209–249, https://doi.org/10.3322/caac.21660.

[2] S.B. Brown, S.E. Hankinson, Endogenous estrogens and the risk of breast, endometrial, and ovarian cancers, Steroids 99 (2015) 8–10, https://doi.org/10.1016/j.steroids.2014.12.013.

[3] X. Xu, T.D. Veenstra, S.D. Fox, J.M. Roman, H.J. Issaq, R. Falk, J.E. Saavedra, L. K. Keefer, R.G. Ziegler, Measuring fifteen endogenous estrogens simultaneously in human urine by high-performance liquid chromatography-mass spectrometry, Anal. Chem. 77 (2005) 6646–6654, https://doi.org/10.1021/ac050697c.

[4] A.H. Eliassen, D. Spiegelman, X. Xu, L.K. Keefer, T.D. Veenstra, R.L. Barbieri, W. C. Willett, S.E. Hankinson, R.G. Ziegler, Urinary estrogens and estrogen metabolites and subsequent risk of breast cancer among premenopausal women, Cancer Res. 72 (2012) 696–706, https://doi.org/10.1158/0008-5472.CAN-11-2507.

[5] The Endogenous Hormones and Breast Cancer Collaborative Group, Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies, Cancer Knowl. Environ. 94 (2002) 606–616, https://doi.org/10.1093/jnci/94.8.606.

[6] J. Russo, I.H. Russo, The role of estrogen in the initiation of breast cancer, J. Steroid Biochem. Mol. Biol. 102 (2006) 89–96, https://doi.org/10.1016/j.jsbmb.2006.09.004.

[7] G. Secreto, A. Girombelli, V. Krogh, Androgen excess in breast cancer development: implications for prevention and treatment, Endocr. Relat. Cancer 26 (2019) R81–R94, https://doi.org/10.1530/ERC-18-0429.

[8] A.E. Drummond, C.T.V. Swain, K.A. Brown, S.C. Dixon-Suen, L. Boing, E.H. van Roekel, M.M. Moore, T.R. Gaunt, R.L. Milne, D.R. English, R.M. Martin, S.J. Lewis, B.M. Lynch, Linking physical activity to breast cancer via sex steroid hormones, part 2: the effect of sex steroid hormones on breast cancer risk, Cancer Epidemiol. Biomark. Prev. 31 (2022) 28–37, https://doi.org/10.1158/1055-9965.EPI-21-0438.

[9] J. Kotsopoulos, S.A. Narod, Androgens and breast cancer, Steroids 77 (2012) 1–9, https://doi.org/10.1016/j.steroids.2011.10.002.

[10] A. Allshouse, J. Pavlovic, N. Santoro, Menstrual cycle hormone changes associated with reproductive aging and how they may relate to symptoms, Obstet. Gynecol. Clin. North Am. 45 (2018) 613–628, https://doi.org/10.1016/j.ogc.2018.07.004.

[11] E. Alladio, E. Amante, C. Bozzolino, F. Seganti, A. Salomone, M. Vincenti, B. Desharnais, Effective validation of chromatographic analytical methods: the illustrative case of androgenic steroids, Talanta 215 (2020) 120867, https://doi.org/10.1016/j.talanta.2020.120867.
[12] B. Desharnais, F. Camirand-Lemyre, P. Mireault, C.D. Skinner, Procedure for the selection and validation of a calibration model I—description and application, J. Anal. Toxicol. (2017), https://doi.org/10.1093/jat/bkx001.
[13] E. Alladio, E. Amante, C. Bozzolino, F. Seganti, A. Salomone, M. Vincenti, B. Desharnais, Experimental and statistical protocol for the effective validation of chromatographic analytical methods, MethodsX 7 (2020) 100919, https://doi.org/10.1016/j.mex.2020.100919.
[14] H. Gu, G. Liu, J. Wang, A.-F. Aubry, M.E. Arnold, Selecting the correct weighting factors for linear and quadratic calibration curves with least-squares regression algorithm in bioanalytical LC-MS/MS assays and impacts of using incorrect weighting factors on curve stability, data quality, and assay perfo, Anal. Chem. 86 (2014) 8959–8966, https://doi.org/10.1021/ac5018265.
[15] L.A. Currie, Detection and quantification limits: origins and historical overview, Anal. Chim. Acta 391 (1999) 127–134, https://doi.org/10.1016/S0003-2670(99)00105-1.
[16] O. González, M.E. Blanco, G. Iriarte, L. Bartolomé, M.I. Maguregui, R.M. Alonso, Bioanalytical chromatographic method validation according to current regulations, with a special focus on the non-well defined parameters limit of quantification, robustness and matrix effect, J. Chromatogr. A. 1353 (2014) 10–27, https://doi.org/10.1016/j.chroma.2014.03.077.
[17] Y. Uwadaira, A. Shimotori, A. Ikehata, K. Fujie, Y. Nakata, H. Suzuki, H. Shimano, K. Hashimoto, Logistic regression analysis for identifying the factors affecting development of non-invasive blood glucose calibration model by near-infrared spectroscopy, Chemom. Intell. Lab. Syst. 148 (2015) 128–133, https://doi.org/10.1016/j.chemolab.2015.09.012.
[18] G.I. Webb, E. Keogh, R. Miikkulainen, Naïve Bayes, Encycl. Mach. Learn. 15 (2010) 713–714.
[19] R. Todeschini, k-nearest neighbour method: the influence of data transformations and metrics, Chemom. Intell. Lab. Syst. 6 (1989) 213–220, https://doi.org/10.1016/0169-7439(89)80086-3.
[20] I.E. Frank, S. Lanteri, Classification models: discriminant analysis, SIMCA, CART, Chemom. Intell. Lab. Syst. 5 (1989) 247–256, https://doi.org/10.1016/0169-7439(89)80052-8.
[21] J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, J.A.K. Suykens, A tutorial on support vector machine-based methods for classification problems in chemometrics, Anal. Chim. Acta 665 (2010) 129–145, https://doi.org/10.1016/j.aca.2010.03.030.
[22] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958, https://doi.org/10.1021/ci034160g.
[23] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemom. Intell. Lab. Syst. 174 (2018) 33–44, https://doi.org/10.1016/j.chemolab.2017.12.004.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning, Python, J. Mach. Learn. Res. 12 (2011) 2825–2830. ⟨http://jmlr.org/papers/v12/pedregosa11a.html⟩.
[25] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinforma. 10 (2009) 213, https://doi.org/10.1186/1471-2105-10-213.
[26] Y. Ning, M.E.H. Ong, B. Chakraborty, B.A. Goldstein, D.S.W. Ting, R. Vaughan, N. Liu, Shapley variable importance cloud for interpretable machine learning, Patterns 3 (2022) 100452, https://doi.org/10.1016/j.patter.2022.100452.
[27] Y. Kim, Y. Kim, Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models, Sustain. Cities Soc. 79 (2022) 103677, https://doi.org/10.1016/j.scs.2022.103677.
[28] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. (2017) 4766–4775.
[29] B.P.O. Lovatti, M.H.C. Nascimento, Á.C. Neto, E.V.R. Castro, P.R. Filgueiras, Use of random forest in the identification of important variables, Microchem. J. 145 (2019) 1129–1134, https://doi.org/10.1016/j.microc.2018.12.028.
[30] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213–225, https://doi.org/10.1002/cem.2609.
[31] W. Song, H. Wang, P. Maguire, O. Nibouche, Collaborative representation based classifier with partial least squares regression for the classification of spectral data, Chemom. Intell. Lab. Syst. 182 (2018) 79–86, https://doi.org/10.1016/j.chemolab.2018.08.011.
[32] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (2015) 1–21, https://doi.org/10.1371/journal.pone.0118432.
[33] H. He, Y. Ma, Imbalanced learning. Foundations, Algorithms and Applications, John Wiley & Sons, Inc, Hoboken, New Jersey, 2013.
[34] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique Nitesh, J. Artif. Intell. Res. 16 (2002) 321–357. ⟨https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp⟩.
[35] P.-Y. Zhou, A.K.C. Wong, Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement, BMC Med. Inform. Decis. Mak. 21 (2021) 16, https://doi.org/10.1186/s12911-020-01356-y.