Original Article
Male reproductive health and infertility

pISSN: 2287-4208 / eISSN: 2287-4690
World J Mens Health Published online Oct 23, 2025
https://doi.org/10.5534/wjmh.250096

The World Journal of MEN's HEALTH

Machine Learning Evaluation of Semen Analysis Could Reveal New Infertility-Related Markers: A Pilot Study

Daniele Santi[1,2,3], Carlotta Pozza[4], Giorgia Spaggiari[2,3], Daniele Gianfrilli[4], Emilia Sbardella[4], Donatella Paoli[5], Laura Roli[6], Maria Cristina De Santis[6], Marco Bonomi[7,8], Tommaso Trenti[6], Andrea M. Isidori[4], Manuela Simoni[1,2]

[1]Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy, [2]Unit of Endocrinology, Department of Medical Specialties, Azienda Ospedaliero-Universitaria of Modena, Modena, Italy, [3]Unit of Andrology and Sexual Medicine of the Unit of Endocrinology, Department of Medical Specialties, Azienda Ospedaliero-Universitaria of Modena, Modena, Italy, [4]Department of Experimental Medicine, "Sapienza" University of Rome, Roma, Italy, [5]Laboratory of Seminology - "Loredana Gandini" Sperm Bank, Department of Experimental Medicine, "Sapienza" University of Rome, Roma, Italy, [6]Department of Laboratory Medicine and Pathology, Azienda USL of Modena, Modena, Italy, [7]Department of Medical Biotechnology and Translational Medicine, University of Milan, Milan, Italy, [8]Department of Endocrine and Metabolic Diseases, IRCCS Istituto Auxologico Italiano, Milan, Italy

Purpose: To perform a pilot study aiming at evaluating whether machine learning could be a useful model to evaluate semen analysis, improving the diagnostic work-up of male partner of infertile couples.
Materials and Methods: A retrospective observational study was conducted using real-world data on male evaluated in routine andrological clinical practice at two Italian tertiary centers. The study utilized two distinct datasets: the first (UNIROMA) encompassed three distinct variables, including semen analysis, sex hormones, and testicular ultrasound parameters. The second dataset (UNIMORE) was constructed incorporating semen analysis, sex hormones, biochemical examinations, and parameters related to environmental pollution. The XGBoost analysis, as part of machine learning techniques, was applied separately to each dataset, as the two datasets did not share a significant overlap in terms of variables.
Results: The UNIROMA dataset comprised 2,334 male subjects. The XGBoost analysis exhibited the highest accuracy (area under the curve [AUC], 0.987) in predicting patients with azoospermia compared to other categories. Remarkably, our analysis revealed that among the most influential predictive variables, follicle-stimulating hormone serum levels (F-score=492.0), inhibin B serum levels (F-score=261), and bitesticular volume (F-score=253.0) stood out. The UNIMORE dataset consisted of 11,981 records. The XGBoost analysis demonstrated a good predictive accuracy (AUC, 0.668), especially for identifying the azoospermia group. Notably, the most crucial predictive variables were environmental pollution parameters (PM10, F-score=361; NO_2, F-score=299) and biochemical data (white blood cells, F-score=326; red blood cells, F-score=299).
Conclusions: This pilot study applies machine learning to two extensive datasets, suggesting that changes in semen analysis may be linked to other variables, such as testicular ultrasound characteristics, red blood cell count, and environmental pollution.

Keywords: Infertility, male; Semen analysis; Testis; Ultrasonography

Received: Mar 13, 2025 **Revised:** May 29, 2025 **Accepted:** Jun 30, 2025 **Published online** Oct 23, 2025
Correspondence to: Daniele Santi https://orcid.org/0000-0001-6607-7105
Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Via Giardini 1355, Modena 41126, Italy.
Tel: +39-593961816, **Fax:** +39-593961433, **E-mail:** daniele.santi@unimore.it

The World Journal of MEN's HEALTH

INTRODUCTION

Couple infertility is a worldwide clinical challenge, affecting at least 15% of couples attempting to conceive through unprotected sexual intercourse [1]. In this context, the male partner contributes to approximately 50% of all cases [2]. Despite significant advancements in the diagnostic work-up of male infertility [3], around 40% of infertile men remain have unexplained etiology, forming the group known as male idiopathic infertility [4]. This cohort of men represents an intriguing epidemiological group that should be further engaged in specific clinical trials. Given the unknown potential etiological factors in idiopathic infertility, specific treatments have not yet been established. Consequently, various empirical attempts, both hormonal and non-hormonal treatments, have been applied, albeit with concerning results for current clinical applications [5-9]. However, the low success rates reported in the literature for these empirical treatments may be attributed to the significant heterogeneity among men falling under the definition of male idiopathic infertility. Indeed, given the lack of knowledge, male idiopathic infertility necessarily includes men with infertility due to different, yet undiagnosed, causes, which underscores the need for further diagnostic development.

Recently, we employed machine learning algorithms to analyze a comprehensive andrological dataset. This groundbreaking effort revealed, for the first time, a significant, yet previously hidden, connection between hematological and semen parameters. This discovery suggests potential intra-individual links that could provide valuable insights into male infertility [10]. This success serves as a prime example of how artificial intelligence can be harnessed to advance our understanding of the factors contributing to male infertility [11,12]. Machine learning, a subfield of artificial intelligence, is instrumental in identifying latent connections between input and output variables to develop automated algorithms [13,14]. This approach is founded on computer-based statistics and self-learning, utilizing algorithms to predict relationships among a multitude of concurrently analyzed variables [15]. These algorithms evolve, becoming 'adaptive' systems as they learn from increasingly extensive datasets, enabling them to handle vast quantities of information and illuminate previously uncharted territories, including challenging conditions such as male infertility [16]. Machine learning has found applications across various domains within andrology, ranging from initial semen analysis to facilitating clinical decision-making [14]. However, the full clinical implementation of these algorithms is still on the horizon. The true value of applying artificial intelligence in male infertility lies in the generation of high-quality evidence, the aggregation of copious amounts of data through diverse approaches, the establishment of interconnected networks, and the potential to reshape medical practice [14]. Consequently, the enigmatic realm of male idiopathic infertility represents a fertile ground for the application of artificial intelligence. This pursuit aims to identify potential subgroups among infertile men. Such a discovery could prove immensely beneficial for future clinical development, as each subgroup of infertile men might receive tailored, and possibly highly effective, therapies.

With these in mind, the present study was designed to implement machine learning in a big data model. In particular, this pilot study was designed with the aim at evaluating whether machine learning could be a useful model to evaluate semen analysis, improving the diagnostic work-up of male partner of infertile couples. It involved the analysis of two extensive datasets containing parameters from two tertiary Italian centers, encompassing semen analysis. Given that semen analysis remains the gold standard for assessing male fertility, it was chosen as the foundation for generating big data to be processed by machine learning algorithms.

MATERIALS AND METHODS

1. Study design

A retrospective, observational pilot study based on real-world data collected during routine andrological clinical practice was carried out in two Italian tertiary centers. The institutional review board of the Azienda Ospedaliero-Universitaria of Modena approved the study conduction (protocol number 2400P).

The first dataset was created considering patients referred for testicular ultrasound at University of Rome (Italy) (UNIROMA) and who performed at least one conventional semen analysis, from 2005 to 2019. The second dataset was created considering all semen analyses performed at University of Modena and Reggio Emilia (Italy) (UNIMORE), from January 2010 to December 2022, regardless of the reason for testing, which ranged from routine screening to infertility evaluation

in couples.

In order to avoid false positive results due to improper sample size [17] or selection biases of potentially informative predictors [18], the study design did not foresee any patients' exclusion criteria for enrollment. Thus, men enrolled in the study were either infertile, fertile or with unknown infertility status, undergoing seminal examinations for any clinical reason.

All semen samples were collected within hospital at room temperature. Semen analyses were performed according to the World Health Organization (WHO) manual edition used by the laboratory at the time of semen collection. In particular, the UNIROMA semen analyses were evaluated using the WHO manual IV edition until 2010 and the V edition for the following years. Otherwise, UNIMORE used the WHO manual V edition until 2021 and the VI edition for the last year of observation.

2. Datasets creation

Two large datasets were built.

The first one (UNIROMA dataset) collected three different categories of variables for each man: (1) semen analysis, (2) sex hormones, and (3) testicular ultrasound parameters (Table 1). Among all patients evaluated at University of Rome, only patients with all three variables available were included in the final dataset. Part of the data was previously reported [19].

The second dataset (UNIMORE dataset) included all laboratory examinations performed in the Province of Modena, which currently accounted for more than 700,000 inhabitants. Table 1 reported the variables included in the UNIMORE dataset, divided in four different categories: (1) semen analysis, (2) hormonal data, (3) biochemical examinations, and (4) environmental pollution related parameters. The latter were introduced to increase the knowledge of potential influencing factors on semen parameters. These data are publicly available at https://dati.arpae.it/dataset/qualita-dell-aria-rete-di-monitoraggio. The description of the geo-localization process was previously reported [20,21]. Environmental data were collected daily from various stations located across different sites in the Province of Modena. These data were sent to the Data Processing Centre of the Emilia-Romagna Environmental Protection Agency (ARPA), which managed the entire moni-

Table 1. Parameters included in the two datasets considered

	UNIROMA dataset	UNIMORE dataset
Anthropometrical variables	Age (years)	Age (years)
	Residential address	Residential address
Semen analysis	Days of abstinence	Days of abstinence
	Semen volume (mL)	Semen volume (mL)
	Semen pH	Semen pH
	Total sperm number (million)	Total sperm number (million)
	Sperm concentration (million/mL)	Sperm concentration (million/mL)
	Sperm progressive motility (%)	Sperm progressive motility (%)
	Normal forms (%)	Normal forms (%)
	Abnormal forms (%)	Abnormal forms (%)
	Leucocytes (million/mL)	Leucocytes (million/mL)
Hormonal examinations	Total testosterone (ng/mL)	Total testosterone (ng/mL)
	Bioavailable testosterone (nmol/L)	
	Free testosterone (nmol/L)	
	FSH (IU/L)	FSH (IU/L)
	LH (IU/L)	LH (IU/L)
	Estradiol (pg/mL)	Estradiol (pg/mL)
	SHBG (nmol/L)	
	Inhibin B (pg/mL)	
		PSA (ng/mL)
		TSH (microIU/mL)
		fT3 (pg/mL)
		fT4 (pg/mL)
		Prolactin (ng/mL)

Table 1. Continued

	UNIROMA dataset	UNIMORE dataset
Biochemical examinations		Red blood cells (millions/mm^3) Hematocrit (%) Red cell distribution width (cv%) Mean corpuscular volume (fl) Hemoglobin (g/dL) Mean corpuscular hemoglobin (pg) Mean corpuscular hemoglobin concentration (g/dL) White blood cells ($10^3/mm^3$) Neutrophil granulocytes ($10^3/mm^3$) Basophilic granulocytes ($10^3/mm^3$) Eosinophil granulocytes ($10^3/mm^3$) Lymphocytes ($10^3/mm^3$) Monocytes ($10^3/mm^3$) Platelets ($10^3/mm^3$) Mean platelet volume (fl) VES (mm) C-reactive protein (mg/dL) Alanine transaminase (U/L) Aspartate transaminase (U/L) Total bilirubin (mg/dL) Fractioned bilirubin (mg/dL) Gamma-glutamyl transferase (GGT) (U/L) Alkaline phosphatase (U/L) Creatinine (mg/dL) Urea (mg/dL) Uric acid (mg/dL) Sodium (mEq/L) Potassium (mEq/L) Glucose (mg/dL) Glycated hemoglobin (%) Total cholesterol (mg/dL) HDL cholesterol (mg/dL) LDL cholesterol (mg/dL) Triglycerides (mg/dL)
Testicular ultrasound	Right testicular volume (mL) Left testicular volume (mL) Bitesticular volume (mL) Asymmetry Score_text (echotexture) Score_reflect (echogenicity) Score_micro (microlithiasis) Left varicocele Right varicocele Score varicocele Bilateral varicocele Lesions	
Environmental pollution		PM10 PM2.5 O_3 NO_2

The first collected at the University of Rome 'La Sapienza' (UNIROMA dataset), the second at the University of Modena and Reggio Emilia (UNIMORE dataset).

FSH: follicle-stimulating hormone, LH: luteinizing hormone, SHBG: sex hormone-binding globulin, PSA: prostate-specific antigen, TSH: thyroid-stimulating hormone, VES: erythrocyte sedimentation rate, HDL: high-density lipoprotein, LDL: low-density lipoprotein, NO: nitric oxide; PM: particulate matter.

toring network. From these databases, environmental parameters (*e.g.*, PM10, O_3, NO_2, and NO) recorded during the 30 and 90 days preceding the semen analysis were included in the dataset as mean values. The sites for environmental data collection were separately registered based on their geographical distribution. These data were then linked to each patient by identifying the station closest to their residential address.

3. Statistical analysis

Statistical analyses were performed considering each dataset alone, since the two datasets did not match for most variables.

The first step provided the bivariate correlation analysis, including all variables available in each dataset. After evaluation of continuous data distribution, correlation analyses were performed by Pearson or Rho's Spearman, according to normal or not normal data distribution, respectively. A strong correlation between two variables was defined when coefficient of correlation was higher than 0.75.

The second step consisted of the application of principal component analysis (PCA), to analyze the dataset with a large number of features, reducing it to a smaller number through a decomposition in a space of reduced dimensions and a noise component. This method allowed an overview of data of which correlation is not known, in order to simplify the graphical representation of the clusters [22]. PCA were applied including all parameters available for each dataset.

The third step was the detection of potential predictors of semen analysis alteration. To this purpose, machine learning was performed using the eXtreme Gradient Boosting (XGBoost) technique [23]. XGBoost is a powerful algorithm, part of the ensemble methods that obtain an accurate classifier starting from poorly skilled models gradually made more and more efficient. It could also capture non-linear patterns and apply regularization methods to avoid overfitting. The selection of XGBoost among potential available algorithm was performed after performing other classifiers (*e.g.*, deep neural network) and considering the characteristics of datasets collected. Indeed, XGBoost fitted with specific scenario, in which was requested high level of accuracy, large datasets, highly scalable, high variety of feature types, and adaptation to unbalanced classes. The XGBoost pre-processing step included a normalization for numeric variables and an encoding for categorical ones. In both cases an Imputer was provided to fill any missing values with the closest neighbor value, in case of numerical features, and with the most frequent value in case of categorical features. For the training pipeline, a 5-fold cross-validation and a randomized fine tuning of some of the hyper-parameters typical of an XGBoost classifier were used randomly selected data within datasets. Before application, the entire dataset was divided in three classes, according to semen analysis characteristics. In particular, we defined subjects with normozoospermia, when sperm concentration, progressive and total sperm motility and normal morphology were above the 5th centile suggested by WHO. The second group was formed by subjects with altered semen analysis parameters. This group included all men with sperm concentration, sperm motility and sperm morphology below the 5th centile suggested by WHO manual edition used when the semen analysis was performed. The third group consisted of men with azoospermia (*i.e.*, no sperms detectable at semen analysis). To address the multi-class problem, which refers to the task of classifying data into more than two classes, both One *versus* Rest (OvR) and One *versus* One (OvO) approaches were performed. These strategies transformed a multi-class problem into a set of binary problems, each characterized by the considered class as opposed to the group comprising all the remaining ones (OvR) or to each of the other classes, taken individually (OvO). Different subsets of starting features (cases) are considered and an XGBoost classifier is trained to predict the class that identifies the condition (target) of a patient starting from the values assumed by these features.

To XGBoost analysis, the relevance of each variable was evaluated by feature importance. This is a statistical approach technique used to determine the relative importance of each feature in a machine learning model. The interpretation and calculation of feature importance can vary depending on the model used. For tree-based models like decision trees and random forests, feature importance is typically calculated based on how much each feature contributes to reducing impurity or splitting the data. The importance was measured by the total reduction in impurity or the total decrease in the Gini index. The higher the value, the more important the feature was considered to be in making predictions. In XGBoost, which is also a tree-based algorithm, feature importance in multi-class

classification models is generally calculated using different methods, such as the "weight," "gain," and "cover" metrics. The "weight" metric, which has been used in this work, represents the number of times a feature appears in a tree across all boosting rounds. It considers both the number of times a feature is used to make splits and the depth of those splits. Features that are more frequently used for splits and have deeper splits tend to have higher F scores, indicating their greater importance in the model.

When machine learning was applied to identify categories of semen analysis alterations, the parameter used for classifying male patients was excluded from the analysis to avoid potential confounding results.

All calculations, data pre-processing, exploratory data analysis, modeling and visualizations were performed with *ad hoc* python scripts and Jupiter notebooks (Linux Ubuntu 20.04.6 LTS [Focal Fossa], Python 3.10.11, with support from OpenAI ChatGPT).

RESULTS

1. UNIROMA dataset

The UNIROMA dataset consisted of 2,334 male subjects (Table 2). A first cleaning step was performed applying data frame reduction for numerical variables, maintaining only those patients in which all three categories of variables (Table 1) were reported. A total of 1,036 subjects were considered (mean age 32.0±10.2 years). As expected, the main outcomes (*i.e.*, semen analysis parameters) showed a wide variability in the entire cohort, ranging from azoospermia (subjects in which no sperms have been detected) to potential normal fertility (subjects in which all semen parameters were above WHO thresholds).

Data distribution was evaluated by qualitative normality test, generating quantile-quantile plots for each variable. Although several variables showed a normal distribution, the vast majority did not follow a Gaussian distribution (Supplement Fig. 1), thus non-parametric statistical analyses were applied.

2. Correlation analysis

Bivariate correlation analyses were performed considering each variable alone (Fig. 1). A strong correlation (Rho>0.75) was detected among variables of the same categories. Indeed, semen analysis parameters showed a positive correlation only among each other (Fig. 1). Similarly, hormonal analyses correlated with each other, as well as ultrasound variables, as expected (Fig. 1). This analysis did not highlight significant correlations connecting semen parameters to either ultrasound or hormonal variables.

3. Clustering analysis

PCA analysis was applied to the final UNIROMA dataset, showing that the first three clusters retained more than 80% of data variability (Supplement Fig. 2). Considering these three principal components and applying a linear clustering algorithm such as K-means, a data separation has been obtained, although the clustering quality was still poor, as suggested by the silhouette plot (Supplement Fig. 3).

4. Machine learning analysis

Since conventional statistical analyses did not describe potential associations among data, machine learning was applied, using the supervised approach. The XGBoost analysis showed an overall good predictive accuracy (area under the curve [AUC], 0.950) (Fig. 2A). In particular, the highest accuracy (AUC, 0.987) was detected to predict patients with azoospermia compared to other classes in

Table 2. Semen analysis parameters in 2,334 male subjects evaluated in the UNIROMA dataset and in 8,095 subjects included in UNIMORE dataset

	UNIROMA dataset	UNIMORE dataset
Days of abstinence	4.00±1.00 (1.0, 7.0)	3.00±1.50 (1.0, 9.0)
Semen volume (mL)	1.25±1.78 (0.5, 13.0)	2.98±1.66 (0.4, 11.5)
Sperm concentration (million/mL)	17.56±42.32 (0.0, 600.0)	161.64±199.88 (0.0, 3200.0)
Total sperm number (millions)	50.12±126.13 (0.0, 1980.0)	58.29±68.34 (0.0, 900.0)
Progressive sperm motility (%)	12.16±19.84 (0.0, 65.0)	32.92±22.73 (0.0, 88.0)
Typical forms (%)	6.70±11.12 (0.0, 47.0)	5.10±10.45 (0.0, 100.0)
Leucocyte number (million/mL)	0.31±0.69 (0.0, 15.0)	0.43±1.87 (0.0, 93.0)

Values are presented as mean±standard deviation (minimum and maximum).

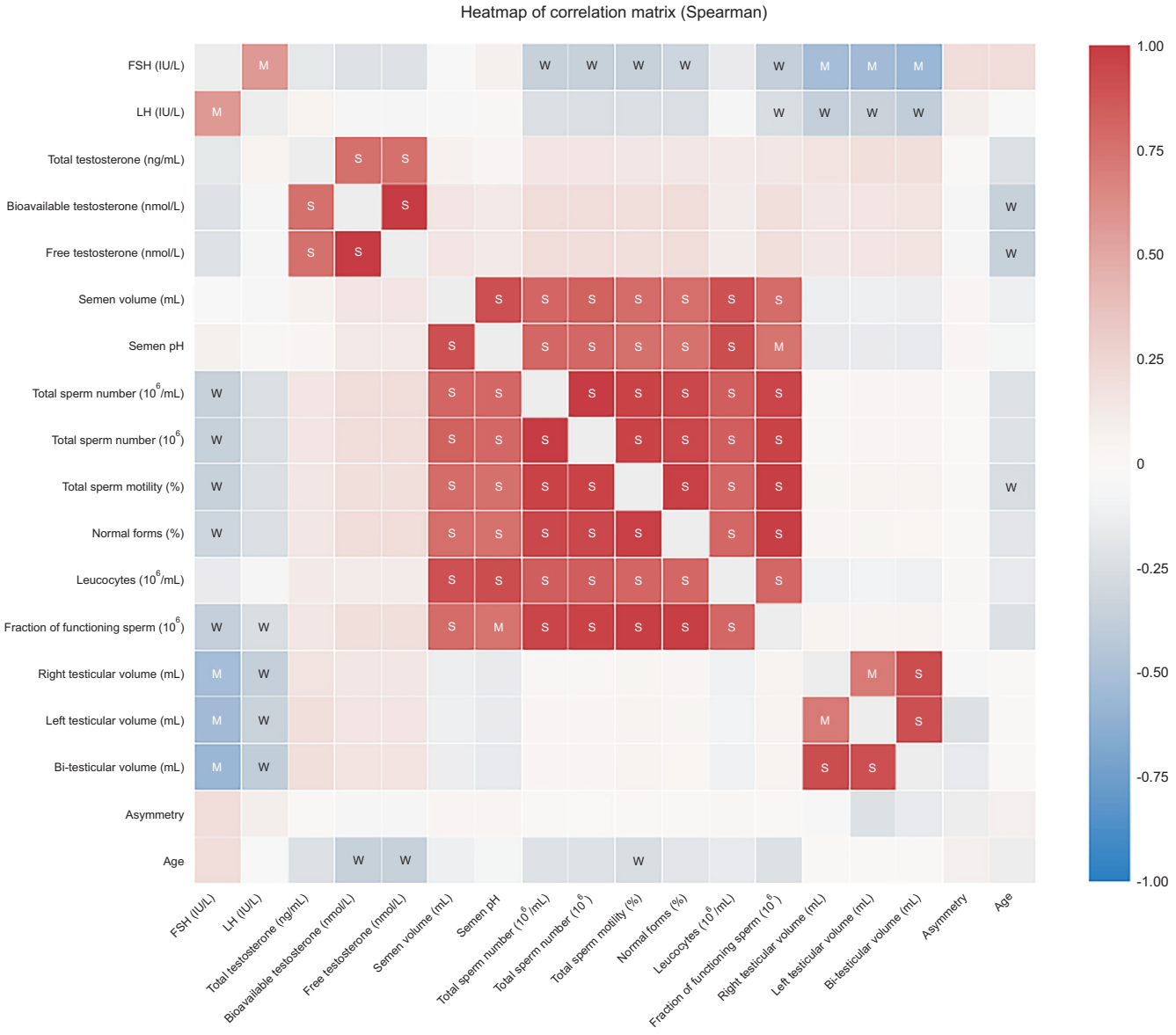

Fig. 1. Heat-map of correlation analyses performed by Rho's Spearman, considering the UNIROME dataset. Positive correlations are displayed in red and negative correlations in blue color. Color intensity is proportional to the correlation coefficients. Statistical significant correlations are reported as strong (S) when p<0.001, and moderate (M) when p<0.05. FSH: follicle-stimulating hormone, LH: luteinizing hormone. The red color indicates a positive correlation, the blue a negative one.

OvR approach (Fig. 2B). The accuracy decreased to 0.907 at predicting subjects with altered semen analysis (Fig. 2B). Similarly, the OvO approach showed a good accuracy, although lowest levels were obtained when subjects with altered semen parameters were compared to those with normozoospermia in the OvO approach (AUC, 0.742) (Fig. 2C). Interestingly, among the most relevant predictive variables, our analysis showed follicle-stimulating hormone (FSH) serum levels (F-score=492.0), inhibin B serum levels (F-score=261) and bitesticular volume (F-score=253.0) (Fig. 3).

5. UNIMORE dataset

The UNIMORE dataset consisted of 11,981 records, belonging to 8,095 subjects. Indeed, several patients underwent semen analysis more than one time during the time frame interval of datasets generation. Data pre-processing consisted of the evaluation of data consistency, removing outliers and missing data, obtaining a total of 10,870 records (mean age 35.0±9.2 years). Outliers were identified by the machine learning approach, defining outliers based on deviations from the majority data distribution.

The World Journal of MEN's HEALTH

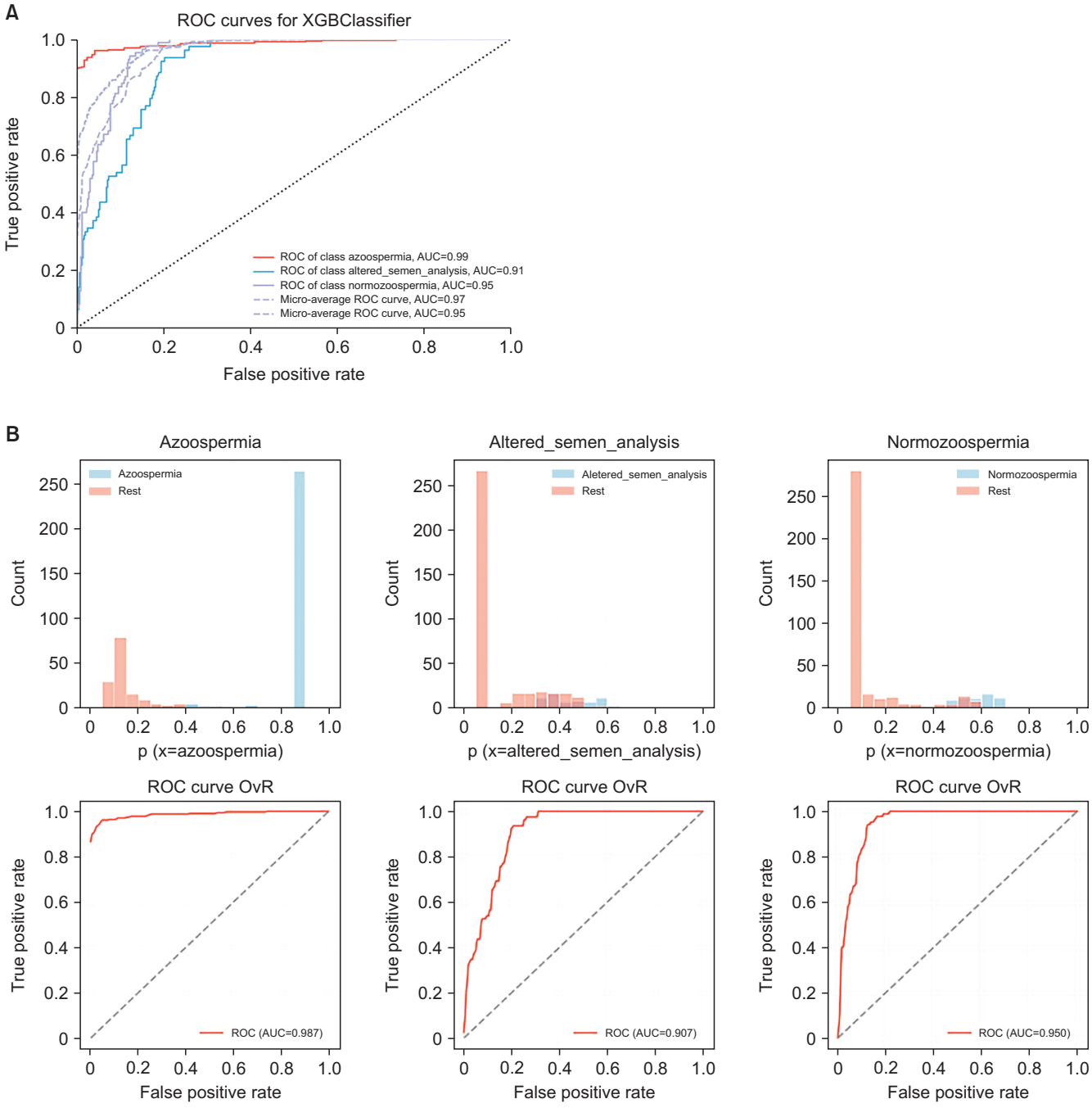

Fig. 2. Classification analysis performed with XGBoost analysis on UNIROMA dataset, subdividing subjects according to semen analysis and considering hormonal and testicular ultrasound data (A), applying the one *vs.* rest (OvR) (B) and the one *vs.* one (OvO) (C) approaches. ROC: receiver operating characteristic, AUC: area under the curve.

Several variables showed a normal distribution, whereas the vast majority did not follow a Gaussian distribution (Supplement Fig. 4).

6. Correlation analysis

Bivariate correlation analyses, considering each variable *per se*, showed a strong correlation (Rho>0.75) among variables within the same category (Fig. 4), such as biochemical examination, semen analyses and environmental parameters (Table 1). On the contrary, external correlations between different characteristics (inter-correlations) are not detected (Fig. 4), such as those obtained in UNIROMA dataset.

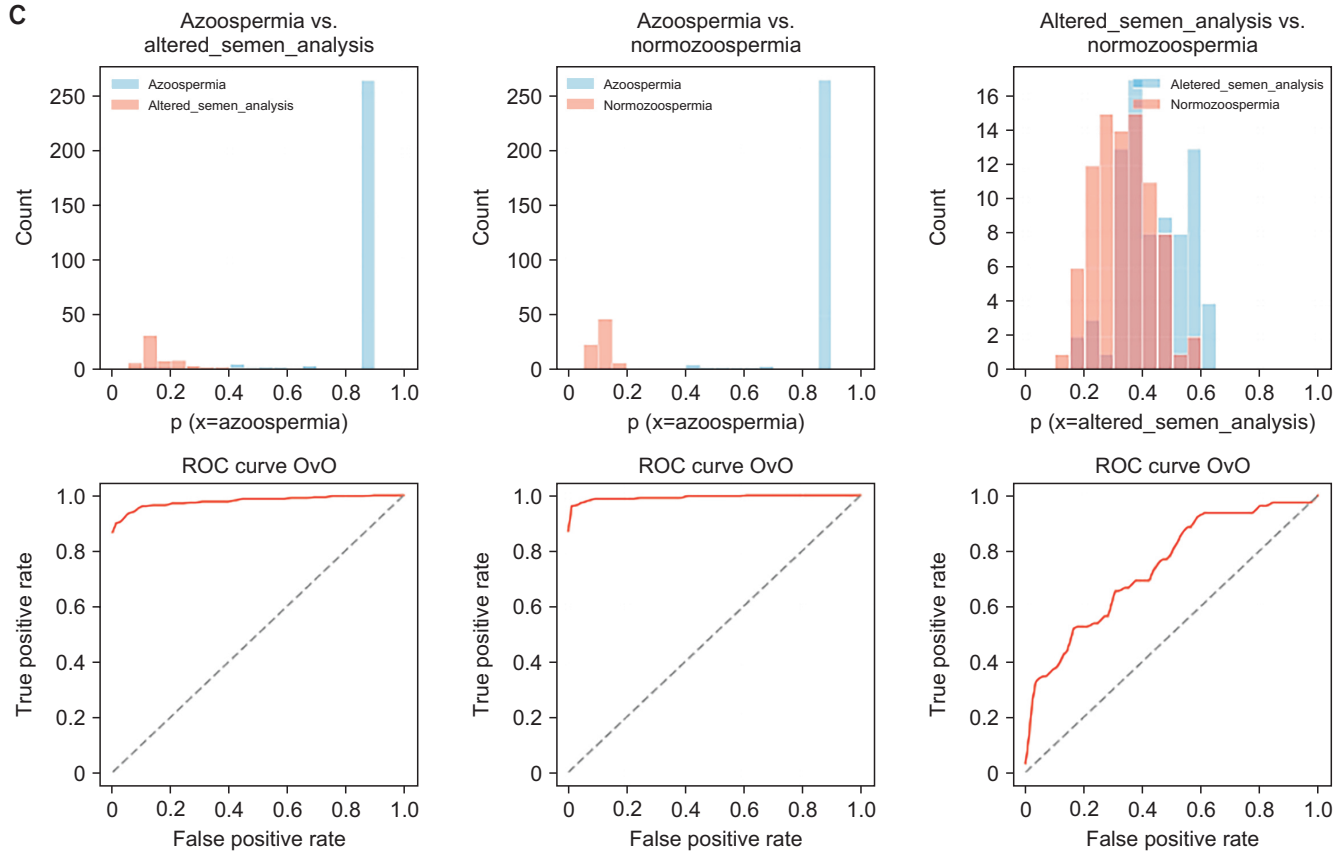

Fig. 2. Continued.

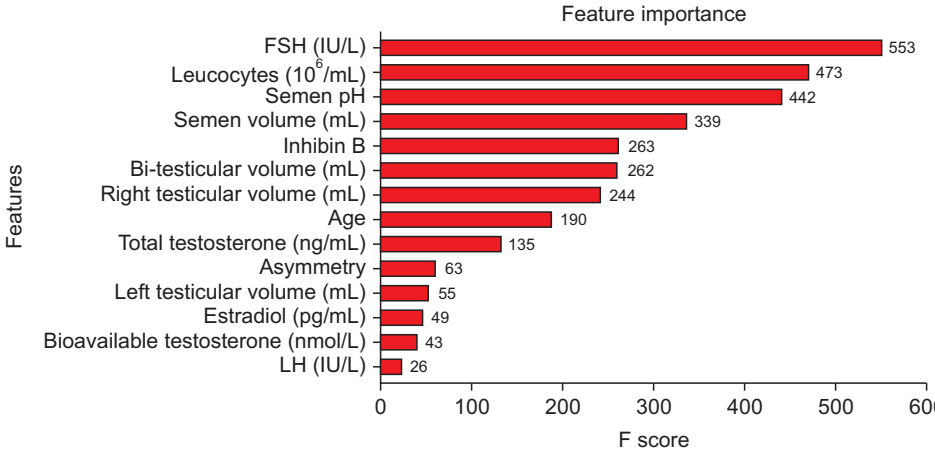

Fig. 3. Features relevance detected by classification analysis performed with XGBoost analysis. FSH: follicle-stimulating hormone, LH: luteinizing hormone.

7. Clustering analysis

A clustering analysis was carried out to highlight whether data, without any external intervention, allowed an intrinsic structure to emerge, using K-means analysis. Moving k from two to five, no results detected well-defined, separate, and balanced clusters. A similar result was obtained applying PCA, showing that seven components were required to explain the 80% of data variance (Supplement Fig. 5).

K-means linear clustering algorithm together with PCA were applied in unsupervised manner, showing that features appeared to separate better than using individual analyses alone (Supplement Fig. 6). However, as demonstrated for the UNIROMA dataset, the quality of this clustering remained insufficient (Supplement Fig. 6), suggesting that data are not linearly

The World Journal of MEN's HEALTH

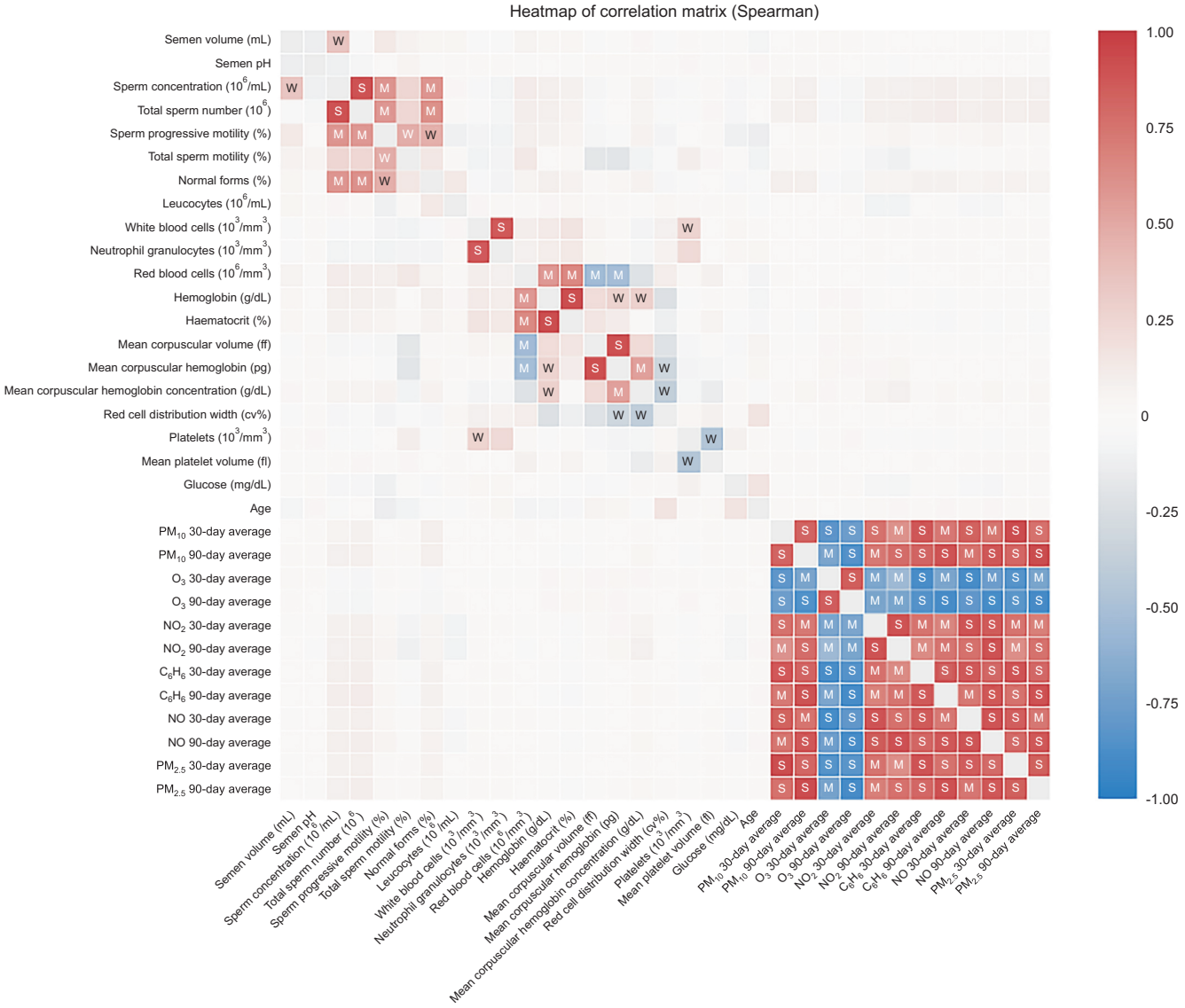

Fig. 4. Heat-map of correlation analyses performed by Rho's Spearman on the UNIMORE dataset. Statistical significant correlations are reported as strong (S) when p<0.001, and moderate (M) when p<0.05. NO: nitric oxide, PM: particulate matter. The red color indicates a positive correlation, the blue a negative one.

separable, and non-linearity should be considered for further modeling techniques adopted.

8. Machine learning analysis

The entire dataset was divided in three classes, according to semen analysis parameters (see Materials and Methods section). The XGBoost analysis showed a fairly good predictive accuracy (AUC: OvO=0.870, OvR=0.668) (Fig. 5A). In details, the azoospermia group was better identified (AUC, 0.699) with the OvR approach compared to the altered semen quality group, which showed the poorest accuracy detected (AUC, 0.631) (Fig. 5B). The OvO approach did not significantly improve the accuracy of the methods, confirming the best performance in the discrimination between azoospermia and altered semen analysis (AUC, 0.680) (Fig. 5C).

Interestingly, among the most relevant predictive variables, our analysis showed that both environmental pollution parameters (PM10, F-score=361 and NO_2 F-score=299) and biochemical data (white blood cells, F-score=326, and red blood cells F-score=299) (Fig. 6) were the most important.

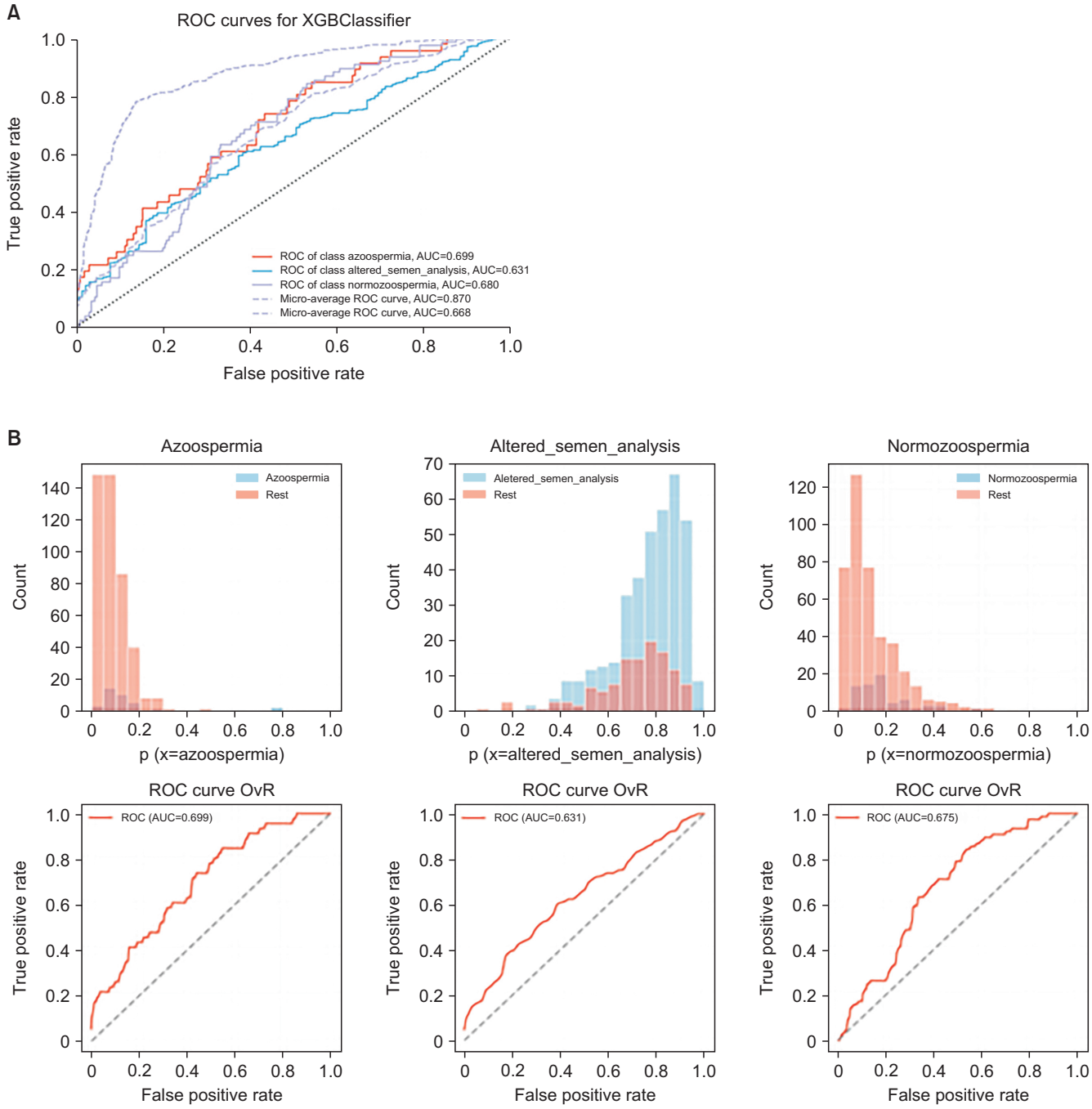

Fig. 5. XGBoost analysis performed to predict subjects classification in UNIMORE dataset (A), applying the one *vs*. rest (OvR) (B) and the one *vs*. one (OvO) (C) approaches. ROC: receiver operating characteristic, AUC: area under the curve.

DISCUSSION

Here, we explore the potential application of artificial intelligence, specifically machine learning, in advancing our understanding of idiopathic male infertility. Through a big data approach, we achieve several key objectives. Firstly, we validate the significant predictive role of FSH and testicular volume in defining male infertility. Secondly, we shed light on the detrimental impact of environmental pollution on sperm production. Lastly, we provide further confirmation of a previously hinted connection, documented in existing literature, between two human tissues characterized by consistent and robust proliferative activity—the spermatogenic and hematopoietic tissues.

Big data analysis enables the examination of large

The World Journal of MEN's HEALTH

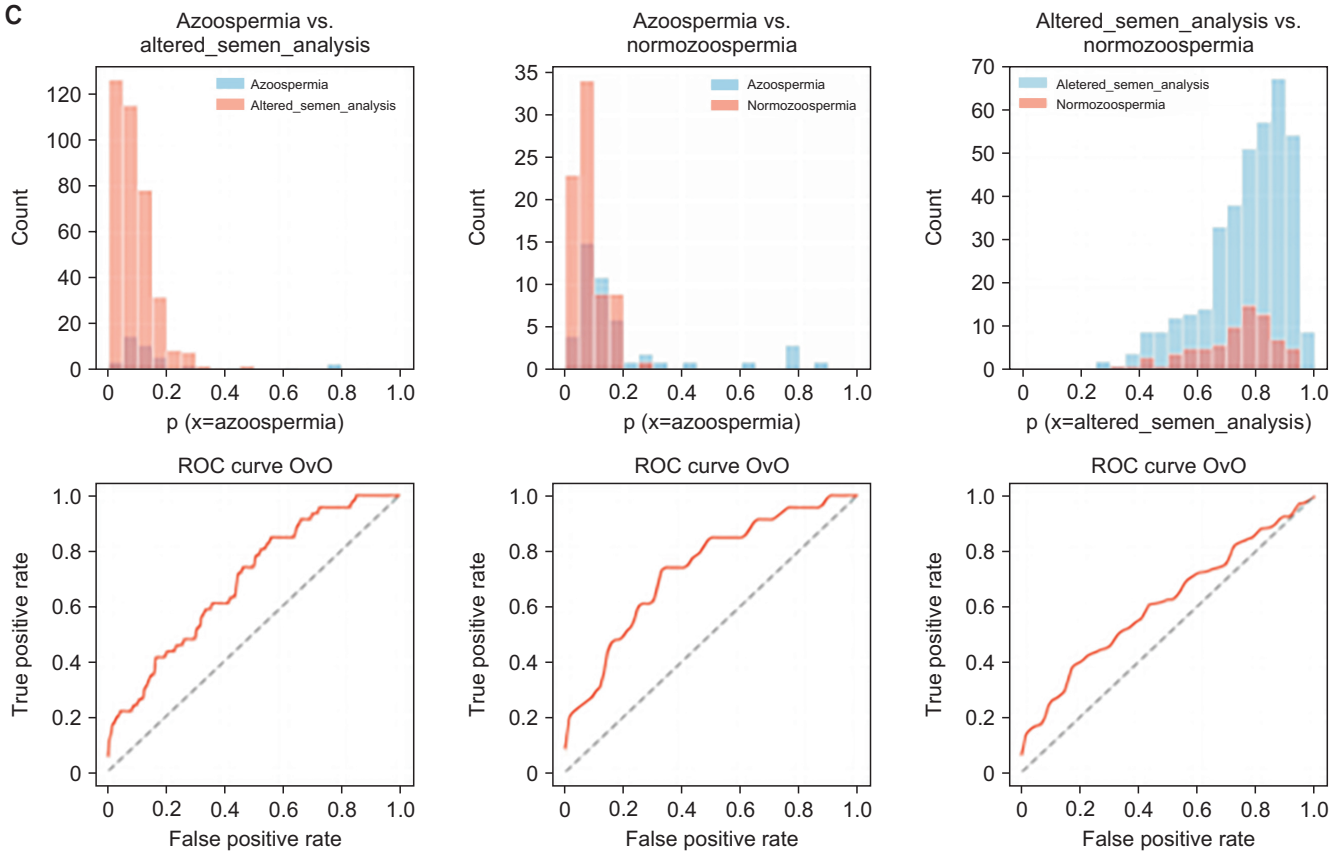

Fig. 5. Continued.

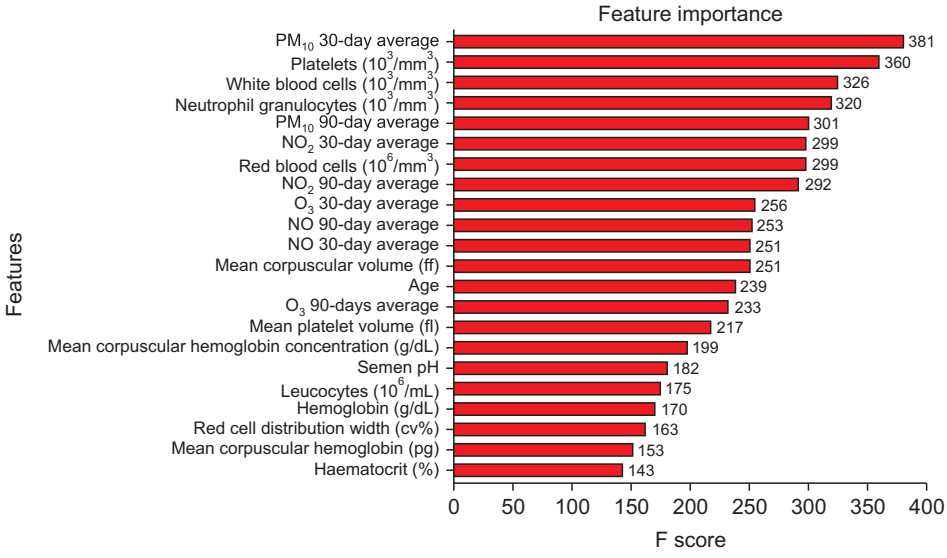

Fig. 6. Features relevance detected by classification analysis performed with XGBoost analysis. NO: nitric oxide, PM: particulate matter.

amounts of data, detecting unexpected new associations. However, such discoveries are only possible when data is analyzed in a meticulous and often complex manner using unconventional statistical approaches. In this context, our study demonstrated that semen analysis parameters, which are traditionally considered the gold standard for assessing male fertility, do not exhibit significant correlations with testicular ultrasound, hormonal levels, or biochemical variables when analyzed using bivariate correlations. Similarly, employing PCA, our data reinforced the intricate nature of male fertility, as it did not reveal any potential data

reduction. Therefore, it might be inferred that there are no parameters directly or indirectly influencing sperm production in humans when considering it as a whole. However, when we applied machine learning, the landscape changed dramatically. Supervised XGBoost analyses proved effective in predicting the fertility status of men included in our two datasets. Upon dividing the two cohorts into three categories based on semen analysis results, namely azoospermia, altered semen analysis, and normozoospermia, we found that machine learning could accurately predict this classification. Notably, subjects with azoospermia were consistently distinguished from the others, regardless of the dataset, suggesting that this group of men is more homogenous and thus easier to recognize. This outcome can be partly explained by the inherent limitations of semen analysis, which cannot effectively differentiate between fertile and infertile men. It is widely acknowledged that there is substantial overlap in conventional semen parameters between fertile and infertile men, failing to provide insights into the functional capacity of sperm to fertilize an oocyte. Consequently, a significant number of men with semen parameters below the WHO 5th centile can achieve pregnancy spontaneously or through assisted reproduction. Thus, conventional semen analysis has inherent limitations in diagnosing male fertility or infertility. Our machine learning approach highlights this challenge and demonstrates a partial overlap in cases where sperm are present. Azoospermic individuals appear to be distinctly different and easily distinguishable. Therefore, grouping men with both azoospermia and altered semen analysis into the same "male infertility" category appears to be a diagnostic pitfall. These two categories are distinct, and men with altered semen analysis appear to be more comparable to those with normozoospermia rather than azoospermia. Obviously, this result must be carefully considered in light also of the inherent pitfalls of semen analysis per se, as well as acknowledging that the high separability in that case is largely driven by extreme values in semen parameters (*e.g.*, sperm concentration=0) and may impact the true diagnostic complexity seen in more ambiguous cases.

In our study, machine learning works with different accuracy depending on the dataset considered, high for UNIROMA, fair for UNIMORE. This difference probably depends on the different nature of data introduced in final datasets. In particular, UNIROMA considered parameters strictly related to the testicular function, such as hypothalamic-pituitary-gonadal axis hormones and testicular ultrasound parameters. Thus, when the male fertility/infertility is evaluated with these parameters, the predictive accuracy is higher. On the contrary, UNIMORE considered data far from the testicular gland, *i.e.* environment parameters and biochemical data. Interestingly, a prediction is possible also considering these data, although with a slightly reduced accuracy. Entering the machine learning models generated and considering the first example, *i.e.* UNIROMA dataset, highlights that semen analysis parameters are mainly predicted by FSH and inhibin B serum levels. These hormones are markers of Sertoli cells function which are considered one of the most complex cell types involved in the complex and highly orchestrated biological process, such as spermatogenesis [24,25]. Sertoli cells are the target cell of FSH action, which is fundamental, together with testosterone, for a qualitatively and quantitatively normal sperm production [26]. Thus, the machine learning approach confirms the predictive role of Sertoli cells' functional markers in the evaluation on male fertility/infertility diagnosis. In the literature, FSH and inhibin B have demonstrated moderate predictive power for distinguishing obstructive from non-obstructive azoospermia, with varying degrees of accuracy depending on the thresholds and models used. Traditional statistical approaches have shown that elevated FSH and low inhibin B levels are generally associated with impaired spermatogenesis, but their sensitivity and specificity remain suboptimal. Recent AI-based studies have attempted to improve the predictive performance of these markers by integrating them with other clinical variables [27,28]. However, these models still face limitations due to the intrinsic biological variability and the lack of standardized cut-off values. Moreover, this machine learning approach highlighted the relevance of several parameters reported in semen analysis, such as leukocytes, semen pH, and volume. These findings underscore the role of accessory glands in determining overall semen quality and emphasize the need for a comprehensive evaluation of semen production, rather than focusing solely on sperm-related parameters [29]. Furthermore, the diagnostic evaluation of male fertility/infertility could be accounted for by the varying characteristics of the two populations. The UNIROMA dataset comprised specifically chosen men from tertiary centers for andrological

concerns, whereas UNIMORE consists of unselected subjects, likely encompassing both fertile and infertile men. It is noteworthy that the predictive significance of hormones, such as FSH and inhibin B, seems more pronounced within the former scenario, whereas it appears attenuated when considering the general population. This discrepancy, rather than indicating a lack of generalizability, highlights the importance of context-specific modeling and the need for multi-center datasets to enhance robustness.

Bitesticular volume calculated at testicular ultrasound is the third main predictor of male fertility/infertility in the UNIROMA dataset. This result confirms that testicular volume is the most relevant parameter obtained by testicular ultrasound, predicting male fertility/infertility. Indeed, spermatogenesis occurs within seminiferous tubules, representing about 80% to 90% of the total testicular volume [30]. The higher is testicular volume, the better is the sperm production. Here, we confirm that testicular ultrasound is the gold standard in assessing testicular volume [31], entering the predicting model, able to discriminate azoospermic patients from those with normal or altered semen analysis. However, among all parameters detectable at ultrasound examination, testicular volume seems to be the only sufficiently strong variable to predict male fertility/infertility. Indeed, several works in the literature tried to evaluate other testicular characteristics on ultrasound at supporting the prognostic significance of its findings in relation to spermatogenesis and testicular endocrine function. Accordingly, a testicular ultrasound score was recently built, combining testicular ultrasound, endocrine, and sperm parameters to provide not only morphological but also functional information of the testis [19]. This score is simple and informative, but probably applicable only to the specific population in which it was created [19]. Indeed, increasing the sample size and including more parameters, its predictive power was lost. Similarly, we applied radiomics on testicular ultrasound variables, showing that from a single ultrasound snapshot we could extract important predictive information on testicular function [32]. However, this 'omics' approach still requires further application in clinical practice [33,34]. With this in mind, our study confirms the relevance of testicular ultrasound in the first andrological assessment for fertility/infertility issues [3].

When other variables are included, far from the testicular function (UNIMORE dataset), we demonstrated that the most relevant predictor of patients' classification were environmental pollution (*i.e.*, PM10 and NO_2), and biochemical variables (*i.e.* white and red blood cells counts). Thus, here we first confirm the potential hidden link between spermatogenic and hematopoietic tissues [10], confirming previous suggestion of a specular regulatory mechanisms at the basis of the two systems [35,36]. Several animal models suggested a common origins between spermatogonial stem cells' niche and both hematopoietic stem cells [36] and white component of blood cells (*i.e.*, lymphocytes and granulocytes) [37-39]. Similarly, our machine learning results confirm the previously suggested association between environmental pollution and sperm production [20,21]. Despite this association has been repeatedly reported [40-42], the mechanism by which several chemicals, free in the air, could affect spermatogenesis is still largely unknown. Previous machine learning approaches highlighted the correlation between air pollution measurements and the percentage of typical and atypical forms, sperm concentration and motility [21]. The higher air pollution is, the higher progressive sperm motility reduction is detected, suggesting a possible effect of pollution on semen quality [43-48]. Obviously, machine learning results are not able to define a cause-effect relationship [49], but they should be intended as a spy of hidden associations that require further deep evaluations. At present, these tools are primarily being developed as diagnostic aids to support clinical decision-making, for example, in stratifying patients, predicting outcomes such as sperm retrieval success, or guiding the need for invasive procedures. In this context, artificial intelligence can help synthesize complex multidimensional datasets that are often challenging to interpret using traditional methods. However, an equally promising avenue lies in the ability of artificial intelligence to uncover novel predictive markers or patterns that may not be evident through conventional statistical analysis. By analyzing large datasets with minimal a priori assumptions, machine learning models may highlight new combinations of features or interactions that are biologically relevant but previously unrecognized. This exploratory capacity opens the door for hypothesis generation and further validation in both clinical and experimental studies.

Our study highlights significant limitations. The first main limit of the study is intrinsic on the pilot

study nature. Indeed, we applied machine learning to two extensive datasets although we are not able to perform external validation so far. The result of this pilot application requires further validation in other large cohorts of subjects. Then, machine learning was employed on two distinct populations. Specifically, UNIROMA comprised selectively chosen patients with infertility issues, whereas UNIMORE encompassed a more heterogeneous group likely to include both fertile and infertile men. Therefore, it's imperative to interpret our findings in view of these notable distinctions. Secondly, the UNIMORE dataset comprised subjects undergoing semen analysis for various clinical reasons, making it unavailable for subgroup analysis based on patients' characteristics. Moreover, only patients with all parameters required for the machine learning application were included in both datasets. This inevitably represents a selection bias that cannot be mitigated by any statistical adjustments. Moreover, our study explores the potential application of machine learning to distinguish between fertile and infertile men. However, this goal is extremely challenging due to the inherent limitations of semen analysis itself. Indeed, no semen analysis parameter can definitively classify a patient as fertile or infertile. Nevertheless, these parameters must be used, at least in research settings, to create subgroups of infertile men, thereby enabling classifications that are useful for statistical analysis [50]. Finally, when a study aims to evaluate couple infertility, pregnancy and live birth rates are the most appropriate endpoints for statistical analysis. However, these data are entirely absent from our datasets, as not all men included were evaluated for infertility, and only male patient data were considered. Therefore, our results must be interpreted with these limitations in mind.

CONCLUSIONS

In conclusion, our study underscores the potential future utility of machine learning in the diagnosis of male infertility. Indeed, this analysis highlights that this statistical approach could be used for a more efficient categorization of male infertility than the conventional methods applied so far. The analysis of two extensive datasets clearly reveals that alterations in semen analysis may be associated with other variables, including testicular ultrasound characteristics, as well as red and white blood cell count and environmental pollution. Future studies, utilizing larger datasets and incorporating more parameters, could enhance the accuracy of the diagnostic process for male infertility.

Conflict of Interest

The authors have nothing to disclose.

Funding

This work was supported by MIUR, Ministero dell'Istruzione dell'Università with the "Progetti di Rilevante Interesse Nazionale" (PRIN2018).

Acknowledgements

None.

Author Contribution

Conceptualization: DS. Data curation: CP, GS, DG, ES, DP, LR, MCDS, TT. Formal analysis: DS. Funding acquisition: MS, DS. Methodology: MS, DS. Writing – original draft: DS. Writing – review & editing: CP, GS, MB, AMI, MS, DS.

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.5534/wjmh.250096.

REFERENCES

1. Sharlip ID, Jarow JP, Belker AM, Lipshultz LI, Sigman M, Thomas AJ, et al. Best practice policies for male infertility. Fertil Steril 2002;77:873-82.
2. Agarwal A, Mulgund A, Hamada A, Chyatte MR. A unique view on male infertility around the globe. Reprod Biol Endocrinol 2015;13:37.
3. Pallotti F, Barbonetti A, Rastrelli G, Santi D, Corona G, Lombardo F. The impact of male factors and their correct and early diagnosis in the infertile couple's pathway: 2021 perspectives. J Endocrinol Invest 2022;45:1807-22.
4. Ventimiglia E, Pozzi E, Capogrosso P, Boeri L, Alfano M, Cazzaniga W, et al. Extensive assessment of underlying etiological factors in primary infertile men reduces the proportion of men with idiopathic infertility. Front Endocrinol (Lausanne) 2021;12:801125.

5. Huijben M, Huijsmans RLN, Lock M, de Kemp VF, de Kort LMO, van Breda J. Clomiphene citrate for male infertility: a systematic review and meta-analysis. Andrology 2023;11:987-96.
6. Santi D, Granata ARM, Simoni M. FSH treatment of male idiopathic infertility improves pregnancy rate: a meta-analysis. Endocr Connect 2015;4:R46-58.
7. Guo B, Li JJ, Ma YL, Zhao YT, Liu JG. Efficacy and safety of letrozole or anastrozole in the treatment of male infertility with low testosterone-estradiol ratio: A meta-analysis and systematic review. Andrology 2022;10:894-909.
8. Magill RG, MacDonald SM. Male infertility and the human microbiome. Front Reprod Health 2023;5:1166201.
9. de Ligny WR, Fleischer K, Grens H, Braat DDM, de Bruin JP. The lack of evidence behind over-the-counter antioxidant supplements for male fertility patients: a scoping review. Hum Reprod Open 2023;2023:hoad020.
10. Santi D, Spaggiari G, Casonati A, Casarini L, Grassi R, Vecchi B, et al. Multilevel approach to male fertility by machine learning highlights a hidden link between haematological and spermatogenetic cells. Andrology 2020;8:1021-9.
11. Jodar M, Sendler E, Krawetz SA. The protein and transcript profiles of human semen. Cell Tissue Res 2016;363:85-96.
12. Brohi RD, Huo LJ. Posttranslational modifications in spermatozoa and effects on male fertility and sperm viability. Omics 2017;21:245-56.
13. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019;19:64.
14. Ghayda RA, Cannarella R, Calogero AE, Shah R, Rambhatla A, Zohdy W, et al. Artificial intelligence in andrology: from semen analysis to image diagnostics. World J Mens Health 2024;42:39-61.
15. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216-9.
16. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 2016;316:2402-10.
17. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14:e0224365.
18. Krautenbacher N, Theis FJ, Fuchs C. Correcting classifiers for sample selection bias in two-phase case-control studies. Comput Math Methods Med 2017;2017:7847531.
19. Pozza C, Kanakis G, Carlomagno F, Lemma A, Pofi R, Tenuta M, et al. Testicular ultrasound score: a new proposal for a scoring system to predict testicular function. Andrology 2020;8:1051-63.
20. Santi D, Vezzani S, Granata AR, Roli L, De Santis MC, Ongaro C, et al. Sperm quality and environment: a retrospective, cohort study in a Northern province of Italy. Environ Res 2016;150:144-53.
21. Santi D, Magnani E, Michelangeli M, Grassi R, Vecchi B, Pedroni G, et al. Seasonal variation of semen parameters correlates with environmental temperature and air pollution: A big data analysis over 6 years. Environ Pollut 2018;235:806-13.
22. Tipping ME, Bishop CM. Mixtures of probabilistic principal component analyzers. Neural Comput 1999;11:443-82.
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Paper presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA, USA. p. 785-94.
24. Yin Y, Ma J, Lu X, Yan S, Jiang Q, Wu D, et al. FSH promotes immature porcine Sertoli cell proliferation by activating the CCR7/Ras-ERK signaling axis. Reproduction 2023;165:593-603.
25. You X, Chen Q, Yuan D, Zhang C, Zhao H. Common markers of testicular Sertoli cells. Expert Rev Mol Diagn 2021;21:613-26.
26. Santi D, Crépieux P, Reiter E, Spaggiari G, Brigante G, Casarini L, et al. Follicle-stimulating hormone (FSH) action on spermatogenesis: a focus on physiological and therapeutic roles. J Clin Med 2020;9:1014.
27. Bachelot G, Dhombres F, Sermondade N, Haj Hamid R, Berthaut I, Frydman V, et al. A machine learning approach for the prediction of testicular sperm extraction in nonobstructive azoospermia: algorithm development and validation study. J Med Internet Res 2023;25:e44047.
28. Zeadna A, Khateeb N, Rokach L, Lior Y, Har-Vardi I, Harlev A, et al. Prediction of sperm extraction in non-obstructive azoospermia patients: a machine-learning perspective. Hum Reprod 2020;35:1505-14.
29. Dutta S, Bocu K, Agarwal A. Role of leukocytospermia in the management of male infertility: decoding a mystery for the busy clinicians. World J Mens Health 2025;43:465-76.
30. Mirochnik B, Bhargava P, Dighe MK, Kanth N. Ultrasound evaluation of scrotal pathology. Radiol Clin North Am 2012;50:317-32, vi.
31. Dogra VS, Gottlieb RH, Oka M, Rubens DJ. Sonography of the scrotum. Radiology 2003;227:18-36.
32. De Santi B, Spaggiari G, Granata AR, Romeo M, Molinari F, Simoni M, et al. From subjective to objective: A pilot study on testicular radiomics analysis as a measure of gonadal function. Andrology 2022;10:505-17.

33. Campbell K, Suarez Arbelaez MC, Ghomeshi A, Ibrahim E, Roy S, Singh P, et al. Next-generation sequencing analysis of semen microbiome taxonomy in men with nonobstructive azoospermia vs. fertile controls: a pilot study. F S Sci 2023;4:257-64.
34. Chatziparasidou A, Sarafidou T, Kyrgiafini MA, Moutou K, Markantoni M, Giannoulis T, et al. Unraveling the genetic basis of azoospermia: transcriptome profiling analyses in a Greek population. F S Sci 2025;6:16-29.
35. Shirazi R, Zarnani AH, Soleimani M, Nayernia K, Ragerdi Kashani I. Differentiation of bone marrow-derived stage-specific embryonic antigen 1 positive pluripotent stem cells into male germ cells. Microsc Res Tech 2017;80:430-40.
36. Köse S, Yersal N, Önen S, Korkusuz P. Comparison of hematopoietic and spermatogonial stem cell niches from the regenerative medicine aspect. Adv Exp Med Biol 2018;1107:15-40.
37. Fraczek M, Kurpisz M. Cytokines in the male reproductive tract and their role in infertility disorders. J Reprod Immunol 2015;108:98-104.
38. Aykan S, Canat L, Gönültaş S, Atalay HA, Altunrende F. Are There Relationships between Seminal Parameters and the Neutrophil-to-Lymphocyte Ratio or the Platelet-to-Lymphocyte Ratio? World J Mens Health 2017;35:51-6.
39. Hamada A, Esteves SC, Nizza M, Agarwal A. Unexplained male infertility: diagnosis and management. Int Braz J Urol 2012;38:576-94.
40. Carlsen E, Giwercman A, Keiding N, Skakkebaek NE. Evidence for decreasing quality of semen during past 50 years. Bmj 1992;305:609-13.
41. Forti G, Serio M. Male infertility: is its rising incidence due to better methods of detection or an increasing frequency? Hum Reprod 1993;8:1153-4.
42. Jensen TK, Carlsen E, Jørgensen N, Berthelsen JG, Keiding N, Christensen K, et al. Poor semen quality may contribute to recent decline in fertility rates. Hum Reprod 2002;17:1437-40.
43. De Rosa M, Zarrilli S, Paesano L, Carbone U, Boggia B, Petretta M, et al. Traffic pollutants affect fertility in men. Hum Reprod 2003;18:1055-61.
44. Mendiola J, Jørgensen N, Andersson AM, Stahlhut RW, Liu F, Swan SH. Reproductive parameters in young men living in Rochester, New York. Fertil Steril 2014;101:1064-71.
45. Viskum S, Rabjerg L, Jørgensen PJ, Grandjean P. Improvement in semen quality associated with decreasing occupational lead exposure. Am J Ind Med 1999;35:257-63.
46. Fathi Najafi T, Latifnejad Roudsari R, Namvar F, Ghavami Ghanbarabadi V, Hadizadeh Talasaz Z, Esmaeli M. Air pollution and quality of sperm: a meta-analysis. Iran Red Crescent Med J 2015;17:e26930.
47. Guven A, Kayikci A, Cam K, Arbak P, Balbay O, Cam M. Alterations in semen parameters of toll collectors working at motorways: does diesel exposure induce detrimental effects on semen? Andrologia 2008;40:346-51.
48. Hammoud A, Carrell DT, Gibson M, Sanderson M, Parker-Jones K, Peterson CM. Decreased sperm motility is associated with air pollution in Salt Lake City. Fertil Steril 2010;93:1875-9.
49. Gentile I, Vezzoli V, Martone S, Totaro MG, Bonomi M, Persani L, et al. Short-term exposure to benzo(a)pyrene causes disruption of GnRH network in zebrafish embryos. Int J Mol Sci 2023;24:6913.
50. Björndahl L, Esteves SC, Ferlin A, Jørgensen N, O'Flaherty C. Improving standard practices in studies using results from basic human semen examination. Andrology 2023;11:1225-31.