Journal of
Clinical Medicine

MDPI

Article

Machine Learning Predictive Models for Evaluating Risk Factors Affecting Sperm Count: Predictions Based on Health Screening Indicators

Hung-Hsiang Huang [1,†], Shang-Ju Hsieh [1,†], Ming-Shu Chen [2,†], Mao-Jhen Jhou [3], Tzu-Chi Liu [3], Hsiang-Li Shen [3], Chih-Te Yang [4], Chung-Chih Hung [5], Ya-Yen Yu [6] and Chi-Jie Lu [3,7,8,*]

1 Division of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei City 220, Taiwan
2 Department of Healthcare Administration, College of Healthcare & Management, Asia Eastern University of Science and Technology, New Taipei City 220, Taiwan
3 Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City 242, Taiwan
4 Department of Business Administration, Tamkang University, New Taipei City 251, Taiwan
5 Department of Laboratory Medicine, Taipei Hospital, Ministry of Health and Welfare, New Taipei City 242, Taiwan
6 Department of Medical Laboratory, Chang-Hua Hospital, Ministry of Health and Welfare, Chang Hua County 513, Taiwan
7 Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City 242, Taiwan
8 Department of Information Management, Fu Jen Catholic University, New Taipei City 242, Taiwan
* Correspondence: 059099@mail.fju.edu.tw; Tel.: +886-2-2905-2973
† These authors contributed equally to this work.

Abstract: In many countries, especially developed nations, the fertility rate and birth rate have continually declined. Taiwan's fertility rate has paralleled this trend and reached its nadir in 2022. Therefore, the government uses many strategies to encourage more married couples to have children. However, couples marrying at an older age may have declining physical status, as well as hypertension and other metabolic syndrome symptoms, in addition to possibly being overweight, which have been the focus of the studies for their influences on male and female gamete quality. Many previous studies based on infertile people are not truly representative of the general population. This study proposed a framework using five machine learning (ML) predictive algorithms—random forest, stochastic gradient boosting, least absolute shrinkage and selection operator regression, ridge regression, and extreme gradient boosting—to identify the major risk factors affecting male sperm count based on a major health screening database in Taiwan. Unlike traditional multiple linear regression, ML algorithms do not need statistical assumptions and can capture non-linear relationships or complex interactions between dependent and independent variables to generate promising performance. We analyzed annual health screening data of 1375 males from 2010 to 2017, including data on health screening indicators, sourced from the MJ Group, a major health screening center in Taiwan. The symmetric mean absolute percentage error, relative absolute error, root relative squared error, and root mean squared error were used as performance evaluation metrics. Our results show that sleep time (ST), alpha-fetoprotein (AFP), body fat (BF), systolic blood pressure (SBP), and blood urea nitrogen (BUN) are the top five risk factors associated with sperm count. ST is a known risk factor influencing reproductive hormone balance, which can affect spermatogenesis and final sperm count. BF and SBP are risk factors associated with metabolic syndrome, another known risk factor of altered male reproductive hormone systems. However, AFP has not been the focus of previous studies on male fertility or semen quality. BUN, the index for kidney function, is also identified as a risk factor by our established ML model. Our results support previous findings that metabolic syndrome has negative impacts on sperm count and semen quality. Sleep duration also has an impact on sperm generation in the testes. AFP and BUN are two novel risk factors linked to sperm counts. These findings could help healthcare personnel and law makers create strategies for creating environments to increase the country's fertility rate. This study should also be of value to follow-up research.

Citation: Huang, H.-H.; Hsieh, S.-J.; Chen, M.-S.; Jhou, M.-J.; Liu, T.-C.; Shen, H.-L.; Yang, C.-T.; Hung, C.-C.; Yu, Y.-Y.; Lu, C.-J. Machine Learning Predictive Models for Evaluating Risk Factors Affecting Sperm Count: Predictions Based on Health Screening Indicators. *J. Clin. Med.* **2023**, *12*, 1220. https://doi.org/10.3390/jcm12031220

Academic Editor: Emad Ibrahim

Received: 20 December 2022
Revised: 13 January 2023
Accepted: 1 February 2023
Published: 3 February 2023

Keywords: health screening indicator; machine learning; male reproductive health; sperm quality; sleep time

1. Introduction

Population aging is one of the by-products of a country's economic development. It increases the burden on the younger generation and diminishes the time available to raise the next generation. Fertility and birth rates have been continually declining in many countries. Taiwan's fertility rate reached its lowest point in 2022 at 1.08 children born per woman, which is lower than the 2.1 needed to maintain the population [1]. Therefore, it is crucial to ensure that married couples wanting to raise the next generation are able to conceive successfully. However, around 15–20% of couples are unable to conceive within one year of unprotected intercourse. Male factors contribute to 50% of all infertile cases. Although advances in assisted reproductive techniques (ART) help many couples to conceive successfully, the success rate of ARTs still depends on semen quality [2].

The decline in fertility has coincided with the falling trend in semen quality in recent years. Sperm count and sperm concentration, two determinants of semen quality, were found to be declining in a meta-analysis of 61 studies published between 1938 and 1990 comparing men with no history of infertility [3]. Since that finding, multiple studies have confirmed this worrying trend of decreasing sperm count and sperm density. In a more recent study, the proportion of men with normal total motile sperm count (>15 million) was found to have declined by about 10% over the past 16 years [4]. Although this trend was found within the subfertile male population, it implies that more couples need ARTs to help them to conceive.

Many risk factors, ranging from the patient's genetic background [5], maternal exposure [6], environmental pollutants [7], metabolic syndrome (MetS) [8], and obesity [9] to the patient's lifestyle [10], have been recognized to affect sperm count. Sperm count is further associated with sperm quality and could determine male fertility [11]. However, the extent of the influence of these factors on semen quality remains to be clearly determined due to the inability to design an experiment to account for all possible confounding factors. In addition, many previous study populations were recruited from infertility centers and their conclusions were not representative of the general population. Therefore, to gain more insights into the interplay between these factors and male fertility in the general population, we are the first study to analyze the annual health screening data, the MJ health-check-up-based population database (MJPD), from a major health screening center in Taiwan. The MJPD is widely used in the healthcare/medical informatics studies [12]. Patients with metabolic syndrome, hyperlipidemia, or different lifestyles were considered and used in this study to analyze the impacts of these risk factors on semen count.

Most of the existing studies usually utilized traditional multiple linear regression (MLR) to analyze the relationship between risk factors and sperm count [13–15]. MLR assumes that the dependent variable should be linearly correlated with independent variables and that collinearity should not occur between independent variables [16–18]. However, the use of MLR has limitations when the data may have non-linear relationships or complex interactions between variables [16]. Machine learning (ML) methods are data-driven algorithms and do not require statistical assumptions. They can capture non-linear relationships between variables or those with complex interactions [19–22]. As ML methods can handle collinearity more effectively than MLR and generate promising performances, they have been widely used for prediction issues in the field of healthcare/medical informatics, while MLR is used as a baseline for comparison [23–26]. However, only a few studies have utilized ML for sperm-count-related research [27–29].The five effective ML methods with different modeling mechanisms, namely, random forest (RF), stochastic gradient boosting (SGB), least absolute shrinkage and selection operator regression (Lasso), ridge regression (Ridge), and extreme gradient boosting (XGBoost), are used in this study since they have

been successfully utilized in many healthcare or medical informatics studies to provide promising results [24,25,30–39]. Thus, this study aims to construct a framework based on RF, SGB, Lasso, Ridge, and XGBoost prediction models to identify the major risk factors affecting male sperm count in order to provide more sperm-count-related research that utilizes ML in the field of reproductive biology.

2. Materials and Methods

2.1. Data Material

The process for identifying subjects in this study consisted of scrutinizing health screening indicators and questionnaire records of 71,108 members of the MJPD for the period 2005–2017. The study selected 30 health screening indicators and questionnaire variables relevant to the investigation. As there might have been multiple annual screening data for each member in the database, only the most recent annual record of the subject was analyzed. Subjects who lacked data on the main study variables were excluded, leaving 30,255 individuals who met the study eligibility criteria. We excluded 6 subjects who were older than 50 years and not evenly distributed in the study groups and 28,874 non-male subjects for whom sperm counts or motility tests were not performed in their annual health examination. We finally identified 1375 eligible male subjects, of whom 686 (49.89%) were married and 619 (45.02%) were unmarried, with an average age of 33.22 $\pm$ 4.36 years.

In Taiwan, many studies using the MJPD are listed on the website (http://www.mjhrf.org/main/page/resource/en/#resource07; accessed on 1 October 2022). The MJPD includes data collected from four MJ clinics that provide health screening to the center's members. All the datasets used were authorized by MJ Health Research Foundation (Approval No.: MJHRF-2016005A). The data application procedures are described at http://www.mjhrf.org/main/page/release1/en/#release01(accessed on 1 October 2022). The MJPD is accessible to academic researchers upon request. The protocol of this study was evaluated for ethical issues regarding the use of data in the database and was deemed acceptable by the Research Ethics Review Committee of Far Eastern Memorial Hospital (FEMH-IRB-107127-E, Protocol Version 1, 15 February 2022) and the MJ Health Research Foundation; it was approved by ClinicalTrials.gov (ID: NCT05225454). The study was conducted according to the guidelines of the Declaration of Helsinki, and all data were anonymized before analysis in accordance with the ethics requirements of the institutional review board.

Figure 1 illustrates the sperm count distribution in different age groups in the sample, while Figure 2 shows the subject identification process for selecting the sample in this study. Table 1 provides the sample attributes of the subjects, including descriptive statistics of the independent and dependent variables. Figure 3 presents the correlation coefficients between 20 numerical independent variables and sperm count using Pearson correlation analysis. It can be seen from Figure 3 that a total of 3 risk factors have a positive linear correlation with the dependent variable, namely, UA, HDL-C, and AFP. A total of 16 risk factors have a negative linear correlation with the dependent variable, namely Age, BMI, BF, WC, WHR, SBP, DBP, Hb, FPG, SGOT, SGPT, BUN, e-GFR, TG, T-Cho, LDL-C, and C/H. Hb has no linear correlation with the dependent variable. Although all of the numerical independent variables do not have a strong linear correlation with the dependent variable, there may be non-linear relationships or complex interactions between variables. Therefore, the five ML predictive algorithms were used in this study as they can analyze data with non-linear relationships or complex interactions between variables [19–22].

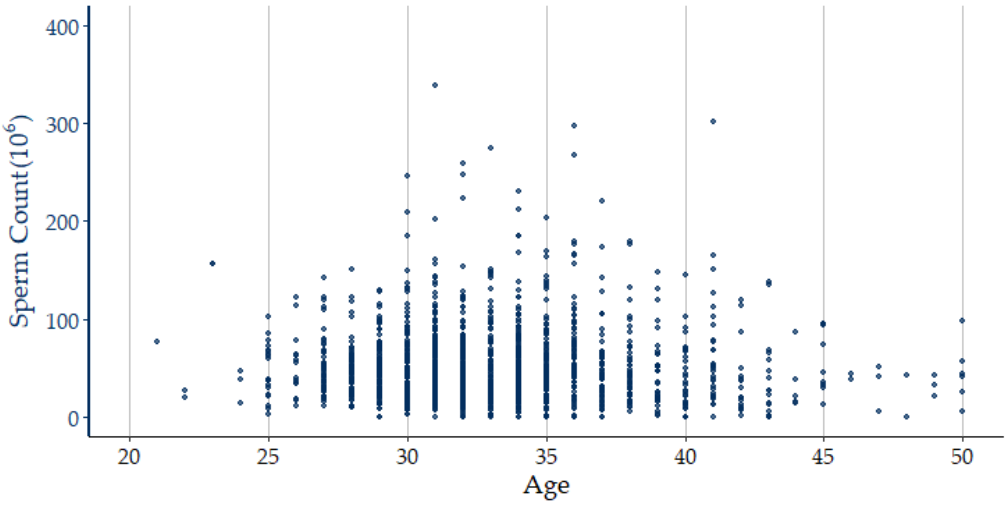

Figure 1. Sperm count distribution by age.

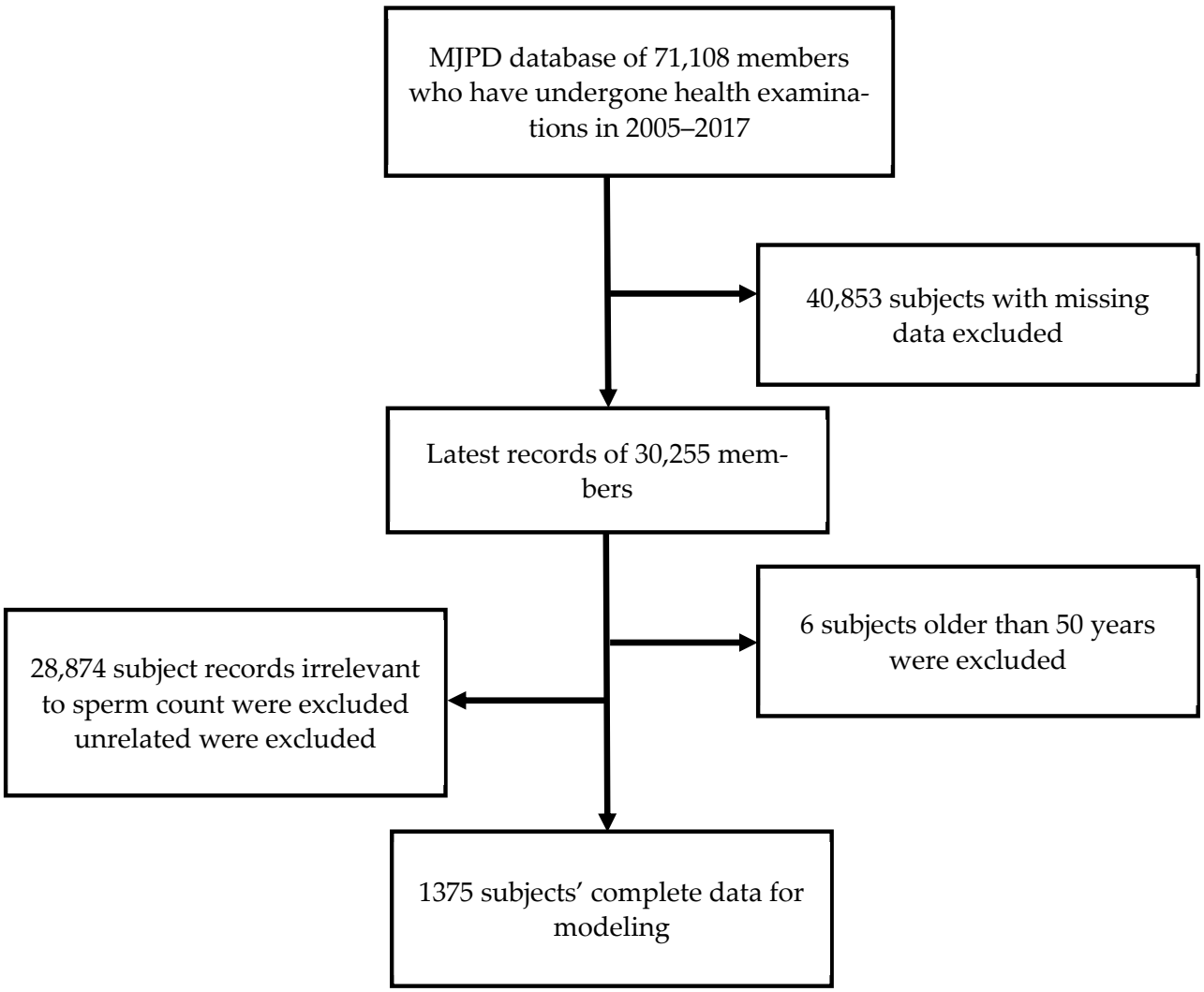

Figure 2. Subject identification process.

Table 1. The independent variables and the dependent variable analyzed in this study.

Independent Variable	N = 1375; n (%)	Independent Variable	N = 1375; n (%)
CS: Current smokers		ST: Sleep time (hours)	
(1) Never	911 (66.25%)	(1) <4	7 (0.51%)
(2) Passive smoking	56 (4.07%)	(2) 4–6	265 (19.27%)
(3) Quit	114 (8.29%)	(3) 6–7	811 (58.98%)
(4) Occasional	58 (4.22%)	(4) 7–8	248 (18.04%)
(5) Addicted	236 (17.16%)	(5) 8–9	44 (3.20%)
AD: Alcohol drinker		(6) >9	NA
(1) Never	1143 (83.13%)	MetS	
(2) Quit	17 (1.24%)	(1) No	1241 (90.25%)
(3) 1–2 times a week	169 (12.29%)	(2) Yes	134 (9.75%)
(4) 3–4 times a week	39 (2.84%)	**Independent Variable**	**Mean ± SD**
(5) 5–6 times a week	NA	Age	33.22 ± 4.36
(6) Addicted	7 (0.51%)	BMI (body mass index, kg/m^2)	24.27 ± 3.37
Vitamin C supplementation		BF (body fat, %)	24.36 ± 5.57
(1) No	1156 (84.07%)	WC (waist circumference, cm)	82.26 ± 8.34
(2) Yes	219 (15.93%)	WHR (waist–hip ratio, %)	0.84 ± 0.05
Vitamin E supplementation		SBP (systolic blood pressure, mmHg)	118.22 ± 12.60
(1) No	1289 (93.75%)	DBP (diastolic blood pressure, mmHg)	72.99 ± 9.62
(2) Yes	86 (6.25%)	Hb (hemoglobin, g/dL)	15.22 ± 0.99
Consumption of Omega-3 rich food		FPG (fasting plasma glucose, mg/dL)	98.61 ± 10.60
(1) No	1283 (93.31%)	SGOT (serum glutamic oxaloacetic transaminase, U/L)	25.78 ± 20.02
(2) Yes	92 (6.69%)	SGPT (serum glutamic pyruvic transaminase, U/L)	36.97 ± 36.02
Consumption of sugar-containing beverages		BUN (blood urea nitrogen, mg/dL)	± 3.01
(1) No or less than 1 cup per week	356 (25.89%)	e-GFR (estimated glomerular filtration rate, ml/min/1.73m$^{2)}$	± 11.23
(2) 1 to 3 cups per week	460 (33.45%)	UA (uric acid, mg/dL)	6.68 ± 1.27
(3) 4 to 6 cups per week	266 (19.35%)	TG (triglyceride, mg/dL)	118.3 ± 68.94
(4) 1 cup per day	198 (14.40%)	T-Cho (total cholesterol, mg/dl)	193.42 ± 32.54
(5) 2 or more than 2 cups per day	95 (6.91%)	HDL-C (high-density lipoprotein cholesterol, mg/dL)	52.36 ± 11.67
Daily physical activity		LDL-C (low-density lipoprotein cholesterol, mg/dL)	119.55 ± 30.63
(1) Sedentary most of the time	928 (67.49%)	C/H (T-Cho/HDL-C)	3.85 ± 0.96
(2) Frequent repeated sitting and ambulation	311 (22.62%)	AFP (alpha–fetoprotein, ng/mL)	2.74 ± 1.33
(3) Standing or ambulation most of the time	111 (8.07%)	**Dependent Variable**	**Mean ± SD**
(4) Requires whole body muscle usage most of the time	25 (1.82%)	S-C (sperm count)	53.3 ± 42.24

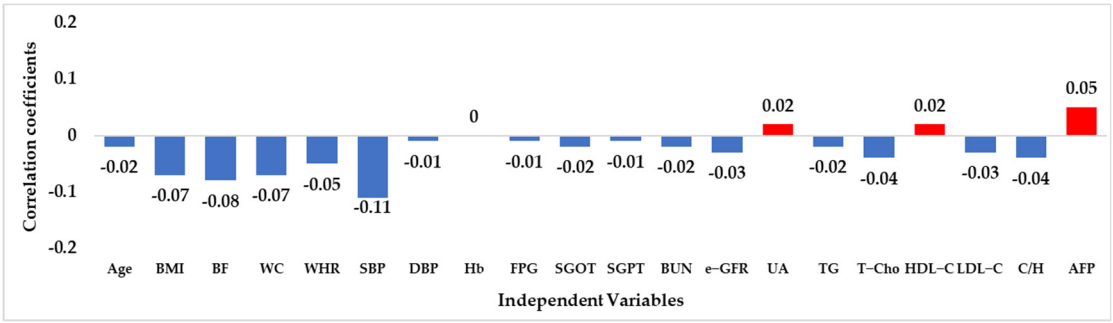

Figure 3. Correlation coefficients between numerical independent variables and sperm count. Note: BMI: body mass index; BF: body fat; WC: waist circumference; WHR: waist–hip ratio; SBP: systolic blood pressure; DBP: diastolic blood pressure; Hb: hemoglobin; FPG: fasting plasma glucose; SGOT: serum glutamic oxaloacetic transaminase; SGPT; serum glutamic pyruvic transaminase; BUN: blood urea nitrogen; e-GFR: estimated glomerular filtration rate; UA: uric acid; TG: triglyceride; T-Cho: total cholesterol; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; C/H: T-Cho/HDL-C; AFP: alpha–fetoprotein.

2.2. Proposed Framework

In this study, a framework was constructed using the five ML prediction models for the identification of important risk factors (independent variables) affecting sperm count, integration, and deliberation. The proposed ML prediction model-based risk factor evaluation framework is shown in Figure 4.

In the proposed framework, the first step involved selecting subjects from the MJPD for the analysis. In the second step, candidate risk variables were chosen and target variables were defined. Twenty-nine risk factors were used as predictor (independent) variables and sperm count was the target (dependent) variable. In the third step, the sperm count of each subject was identified. After the data were organized, the fourth step involved construction of the prediction model for sperm count using the five ML techniques: RF; SGB; Lasso; Ridge; and XGBoost.

RF is a technique that integrates decision tree methods [40]. It randomly generates multiple different and unpruned decision trees, each of which determines the growth of the tree based on the Gini index, and integrates all the trees generated into a forest. It then averages or votes for the trees in the forest to produce a stable ensemble model, thereby reducing correlation between trees and generalization error. Eventually, a stable ensemble model is generated. SGB implements a combination of bagging and boosting [41,42] to generate numerous additive regression trees by multiple iterations. Each tree is trained according to the residuals of the previous iteration [42]. The final number of additive regression trees is determined by satisfying the maximum number of iterations or the convergence condition. Finally, the cumulative result of multiple trees is obtained by weighted summation to determine the final stable model.

Lasso is an extension of the conventional regression method and is based on the principle of using the least absolute shrinkage and selection operator (L1 regularization) to reduce the overfitting problem by forcing the coefficients that contribute less variance to the model to exactly zero, thereby obtaining a lower variance [43,44]. Ridge has the same basic concept as Lasso, with the main difference being that Ridge uses L2 regularization to reduce the coefficients in the model. Ridge adds an appropriate L2 penalty to the model to reduce all coefficients to non-zero values or values close to zero, and then minimizes the sum of squared errors to further control the trade-off between bias and variance to reduce overfitting [45].

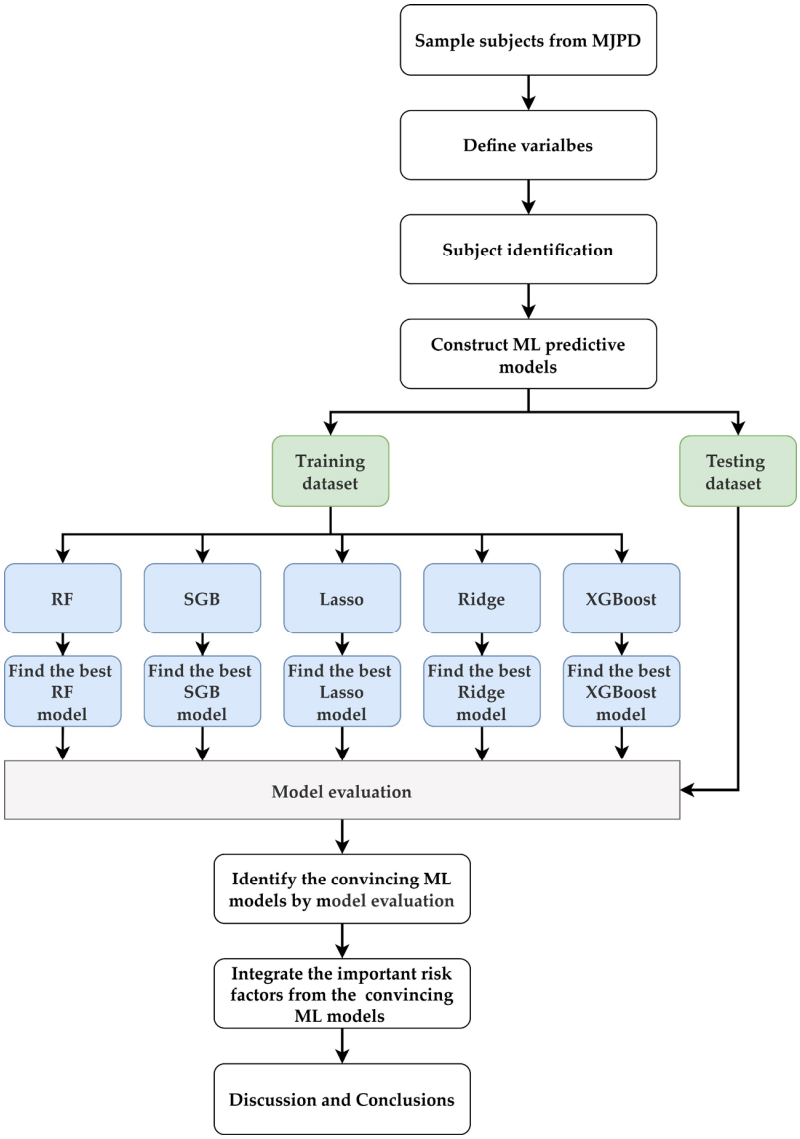

Figure 4. Proposed framework based on machine learning prediction models. Note: MJPD: MJ health-check-up-based population database; ML: machine learning; RF: random forest; SGB: stochastic gradient boosting; Lasso: least absolute shrinkage and selection operator regression; Ridge: ridge regression; XGBoost: extreme gradient boosting.

XGBoost is an optimized gradient-boosting decision tree method. The concept is to generate multiple decision tree models in a sequential manner, with each model generated to fit the residuals of the previous model and a regularization term used to control the complexity of each model, eventually combining all the decision trees generated to improve the accuracy of the prediction [46].

When constructing each ML model, the data were randomly divided into a training data set with 80% of the data and a test data set with 20% of the data. The training data set was used to perform hyperparameter tuning and validation of the model using a 10-fold cross-validation method. Then, the model with the best hyperparameter was selected as the final model, and information on the importance of the corresponding variable was obtained. Finally, the best model predictive performance of each ML method was evaluated with the test data set. To verify the accuracy of the models generated, the performance of each model was measured using four key evaluation metrics—symmetric mean absolute percentage error (SMAPE), relative absolute error (RAE), root relative squared error (RRSE), and root mean squared error (RMSE) (Table 2).

Table 2. Equations for calculating performance metrics.

Metric	Description	Calculation
SMAPE	Symmetric mean absolute percentage error	$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i-\hat{y}_i\rvert}{(\lvert y_i\rvert+\lvert\hat{y}_i\rvert)/2}\times 100$
RAE	Relative absolute error	$\text{RAE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i)^2}}$
RRSE	Root relative squared error	$\text{RRSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}}$
RMSE	Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

$\hat{y}_i$ and y_i represent predicted and actual values, respectively; n stands for the number of instances.

After constructing valid RF, SGB, Lasso, Ridge, and XGBoost predictive models, the fifth step involved obtaining the relative importance values generated by each method for each predictor variable/risk factor according to the converging ML model. The importance of the most and least important risk factors were 100 and 0, respectively.

In the sixth step, each ML method generated different importance values for each predictor variable since the different methods had individual characteristics. In order to integrate the advantages of these methods and obtain more stable results, the average importance value was used to integrate and compare the predictor variables that were more important overall in the set of importance rankings, thus, improving stability and completeness. In the seventh step, a final analysis was performed and the results discussed to obtain the final conclusion.

In order to construct accurate predictive semen count models, all predictive models were built with R version 3.6.2 and RStudio version 1.1.453 (http://www.R-project.org; accessed on 25 May 2022; https://www.rstudio.com/products/rstudio/; accessed on 25 May 2022). Each model was constructed using the associated software packages of R. RF, SGB, Lasso, Ridge, and XGBoost are available in the "ran-domForest" package version 4.7-1.1 [47], "gbm" package version 2.1.8 [48], "glmnet" package version 4.1-1 [49], and "XGBoost" package version 1.6.0.1 [50]. Finally, version 6.0-93 of the "caret" package was used to find the optimal hyperparameters for all models [51].

3. Results

We mainly targeted the younger health screening group for our study sample; therefore, the average age of the sample is relatively low (33.22 ± 4.36 years) and the descriptive statistics show that the study group consists of relatively young healthy and subhealth groups (Table 1). Although the study was a one-time semen analysis, through different ML algorithms, we were able to identify risk factors that may affect semen quality, which could contribute to the prevention of poor sperm quality in unmarried men. We used five ML techniques, RF, SGB, Lasso, Ridge, and XGBoost, to construct predictive models for sperm count. Each method was evaluated based on four performance indicators (SMAPE, RAE, RRSE, and RMSE); we found that the smaller the indicator, the better the predictive performance of the model. Table 3 provides the results of comparison of the predictive performance of the five models. Ridge shows the best performance for SMAPE (0.530) and RAE (0.964) and Lasso shows the best performance for RRSE (1.005) and RMSE (52.608).

Overall, although the predictive performance of the ML algorithms is slightly different, that of the five models is similar and excellent. The five ML methods use different concepts to obtain the variable importance of each risk factor. Therefore, we average the importance values generated by the five methods for the same risk factor and rank each risk factor in descending order of its average variable importance in order to integrate the variable importance information generated by the methods to obtain more robust results and to find the top 10 important risk factors for predicting sperm count.

Table 3. Model performance in predicting sperm count.

Methods	SMAPE	RAE	RRSE	RMSE
RF	0.537	0.984	1.014	53.060
SGB	0.536	0.977	1.017	53.218
Lasso	0.534	0.972	1.005	52.608
Ridge	0.530	0.964	1.006	52.674
XGBoost	0.532	0.968	1.011	52.913

Note: RF: random forest; SGB: stochastic gradient boosting; Lasso: least absolute shrinkage and selection operator regression; Ridge: ridge regression; XGBoost: extreme gradient boosting.

Figure 5 illustrates the average top 10 risk factors in the five ML methods. The top ranked and most important risk factor is sleep time (ST) with an average importance of 66.6 (avg. 66.6). As mentioned before, the five ML methods generate different variable importance values for ST; RF generates a variable importance value of 37.8, SGB of 54.4, Lasso of 100, Ridge of 85.8, and XGBoost of 55.3. The average importance (the average of these five importance values) is 66.6. The second most important risk factor is AFP with an average importance of 61.9 (avg. 61.9). Similarly, the third to tenth important variables are BF, SBP, BUN, BMI, C/H, UA, T-Cho, and WHR, in that order.

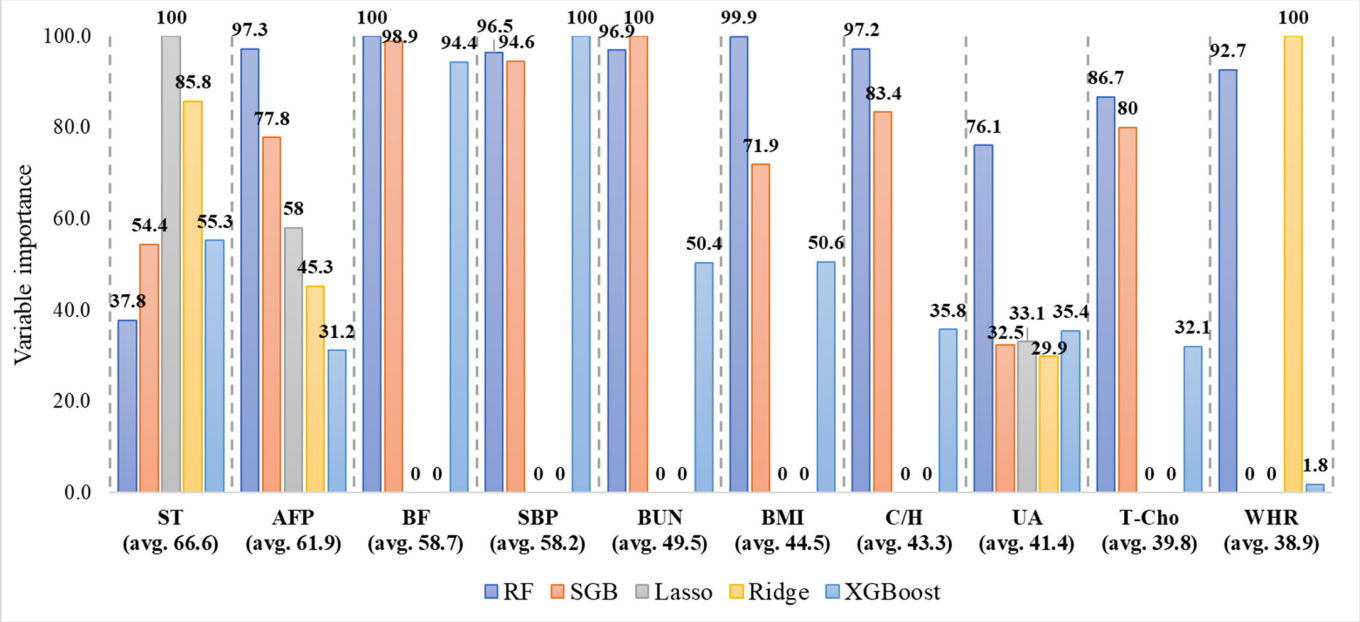

Figure 5. The variable importance to generated by the five algorithms for each risk factor. Note: ST: sleep time; AFP: alpha–fetoprotein; BF: body fat; SBP: systolic blood pressure; BUN: blood urea nitrogen; BMI: body mass index; C/H: T-Cho/HDL-C; UA: uric acid; T-Cho: total cholesterol; WHR: waist–hip ratio; RF: random forest; SGB: stochastic gradient boosting; Lasso: least absolute shrinkage and selection operator regression; Ridge: ridge regression; XGBoost: extreme gradient boosting.

To investigate the variables with greater clinical relevance, we focus on the top five important risk factors identified in this study, namely, ST, AFP, BF, SBP, and BUN.

4. Discussion

Both too-short and too-long sleep durations result in poor-quality semen [52]. Sleep disturbance is also associated with parameters indicating poor semen quality; men suffering from disturbed sleep show lower total sperm count, percentage of total and progressive motility, and percentage of morphologically normal spermatozoa compared to men enjoying high-quality sleep [53]. Sleep deprivation in rats increases stressful stimuli, which

leads to the activation of the hypothalamus–pituitary–adrenal axis and causes elevated serum corticosteroid levels and decreased testosterone levels [54]. However, no difference in sperm count or sperm motility was found in this sleep-deprived animal model compared to the control groups. Therefore, whether sleep duration affects sperm quality through changing reproductive hormone levels or through different pathways affecting gene expression patterns related to spermatogenesis remains inconclusive.

Our study indicates that a shorter sleep duration has adverse effects on sperm count. It is possible that with a shorter sleep duration, reproductive hormone levels might be changed to a level that causes lower spermatogenesis. Further investigations into the link between sleep duration and sperm count are needed.

Alpha-fetoprotein is another risk factor identified by our established model. Few studies highlight this link between AFP and semen quality. In experiments with cryptorchid mice, AFP is specifically expressed in spermatocytes and secreted into the circulation [55]. Injection of AFP into the seminiferous tubules of normal mice could block spermiogenesis, the final step of spermatogenesis. A recent study found high serum AFP in male patients with aberrant sperm counts [56].

However, some of these studies were based on injecting AFP into the semen of animals, and the resulting concentration of AFP should be much higher than that found in healthy male patients. In our current study, we find a positive relationship between AFP and male sperm count. We suspect that there may be a U-shaped relationship between AFP and sperm count, meaning that both too low and too high levels have negative impacts on sperm count. However, it is still required for maintaining normal sperm count, and more studies are needed to illustrate its relationship with male sperm count.

BF, SBP, and other factors in our top 10 list of risk factors (BMI, C/H, T-Cho, and WHR) are related to metabolic syndrome, which has become a global epidemic. Metabolic syndrome has been linked to male infertility and poor semen quality [57], and many studies show that reproductive hormones are altered in males with the syndrome [58–60]. Our results support the view that more severe metabolic syndrome has an adverse effect on sperm count.

In the case of BUN, the fifth risk factor in our ranking, no investigations to date have been performed to find its direct link with male fertility or semen quality. However, chronic kidney disease (CKD) has been found to be associated with poor semen quality by affecting spermatogenesis and sperm motility [61]. The link between CKD and semen quality could be multifactorial. Most of these studies were based on the analysis of advanced CKD or patients under hemodialysis. However, in relatively healthy male patients, higher BUN levels seem to have a negative effect on sperm count. Therefore, the link between elevated BUN and sperm count in the healthy population or prior to the development of CKD requires further detailed study.

In summary, the established ML model successfully reproduces the findings of previous studies that sleep duration, BF, SBP, and BUN negatively affect sperm count. AFP is a lesser-known risk factor and more studies are needed to identify its relationship with male sperm count.

5. Limitations

This study was a cross-sectional study investigating the links between health examination data and sperm count of middle-aged males in Taiwan. The participants included 686 (49.89%) married and 619 (45.02%) unmarried males. Our study used five ML methods to analyze the risk factors affecting sperm count in healthy males. We listed these risk factors according to their importance in affecting sperm counts. Our study was based on a single analysis of semen; therefore, it does not truly reflect the participants' fertility, which needs multiple analyses of semen at different time points. With enough participants, a cross-sectional study could more comprehensively identify risk factors linked to sperm count changes. In addition, ML enables the analysis of nonlinear relationships and complex interactions between multiple predictor variables in this study. However, the top five risk

factors, except AFP, all have a negative impact on male sperm count. AFP shows a positive influence on male sperm count; however, there may be a U-shape relationship between AFP and sperm count. It is necessary for maintaining sperm count; however, both too much or too little can have adverse effects on sperm production. To support this hypothesis, more sophisticated algorithms are needed to identify these U-shaped relationships with sperm count.

6. Conclusions

From Taiwan's health screening data of 1375 male patients, the established ML model predicts many risk factors affecting male semen qualities. Some of our predicted risk factors are consistent with previous results and thoroughly studied. Specially, ST is recognized in different algorithms and is the highest-ranking risk factor after sorting. After becoming a developed country, late marriage and low birth rate are important problems that need to be dealt with. Based on our studies and previous research, regular lifestyle and enough sleep duration are strongly suggested to improve semen quality and decrease the risk of male infertility indirectly.

The different algorithms in this study found sleep time to be the most important variable for predicting semen quality after joint ranking. Most residents of cities in developed countries, with a similar demographic and economic environment to that of Taiwan, tend to marry late and have fewer children. In view of the preliminary results of this study and its corroboration of findings of previous investigations, we suggest that the relevant government departments or health authorities in Taiwan should promote appropriate health information to the male population of reproductive age and advocate normal workloads and sufficient sleep and rest. This may help to avoid the risk of decreased sperm count or an indirect negative impact on male fertility.

Author Contributions: S.-J.H. and M.-S.C. were the equivalent contribution to the first author H.-H.H. and provided the study idea and resources; C.-J.L. and M.-J.J. were the correspondence authors who conceived and designed the experiments; S.-J.H. and M.-S.C. performed the experiments; M.-J.J. analyzed the data; C.-J.L. and M.-S.C. were supervisors; H.-H.H., S.-J.H., and M.-S.C. performed the project administration; although all authors including T.-C.L., H.-L.S., C.-T.Y., C.-C.H., and Y.-Y.Y. contributed to writing and finalizing this article. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Far Eastern Memorial Hospital (NSC-RD-110-1-10-503) and by the National Science and Technology Council, Taiwan (NSTC 111-2221-E-030-009; NSTC-110-2221-E-161-003). This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: The study was approved by the ClinicalTrials.gov (ID: NCT05225454) and by the Research Ethics Review Committee of Far Eastern Memorial Hospital (No: IRB-110027-E/Approved date; 15 February 2022). It was conducted according to the guidelines of the Declaration of Helsinki.

Informed Consent Statement: Informed consent was not required.

Data Availability Statement: All of the datasets collected from the MJ Health Research Foundation, the data need to apply and authorize the use, and the application procedures are accessed via this link. http://www.mjhrf.org/main/page/release1/en/#release01 (accessed on 1 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Central Intelligence Agency. The World Factbook—Central Intelligence Agency: East and Southeast Asia: Taiwan. Available online: https://www.cia.gov/the-world-factbook/countries/taiwan/ (accessed on 14 July 2022).
2. Plachot, M.; Belaisch-Allart, J.; Mayenga, J.M.; Chouraqui, A.; Tesquier, L.; Serkine, A.M. Outcome of conventional IVF and ICSI on sibling oocytes in mild male fator infertility. *Hum. Reprod.* **2002**, *17*, 362–369. [CrossRef]
3. Carlsen, E.; Giwercman, A.; Keiding, N.; Skakkebaek, N.E. Evidence for decreasing quality of semen during past 50 years. *BMJ* **1992**, *305*, 609–613. [CrossRef] [PubMed]

4. Tiegs, A.W.; Landis, J.; Garrido, N.; Scott, R.T., Jr.; Hotaling, J.M. Total Motile Sperm Count Trend Over Time: Evaluation of Semen Analyses From 119,972 Men From Subfertile Couples. *Urology* **2019**, *132*, 109–116. [CrossRef] [PubMed]
5. Krausz, C.; Cioppi, F.; Riera-Escamilla, A. Testing for genetic contributions to infertility: Potential clinical impact. *Expert Rev. Mol. Diagn.* **2018**, *18*, 331–346. [CrossRef]
6. Sharpe, R.M.; Fisher, J.S.; Millar, M.M.; Jobling, S.; Sumpter, J.P. Gestational and lactational exposure of rats to xenoestrogens results in reduced testicular size and sperm production. *Environ. Health Perspect.* **1995**, *103*, 1136–1143. [CrossRef] [PubMed]
7. Jurewicz, J.; Hanke, W.; Radwan, M.; Bonde, J.P. Environmental factors and semen quality. *Int. J. Occup. Med. Environ. Health* **2009**, *22*, 305–329. [CrossRef]
8. Martins, A.D.; Majzoub, A.; Agawal, A. Metabolic Syndrome and Male Fertility. *World J. Men's Health* **2019**, *37*, 113–127. [CrossRef]
9. Palmer, N.O.; Bakos, H.W.; Fullston, T.; Lane, M. Impact of obesity on male fertility, sperm function and molecular composition. *Spermatogenesis* **2012**, *2*, 253–263. [CrossRef] [PubMed]
10. Shi, X.; Chan, C.; Waters, T.; Chi, L.; Chan, D.; Li, T.C. Lifestyle and demographic factors associated with human semen quality and sperm function. *Syst. Biol. Reprod. Med.* **2018**, *64*, 358–367. [CrossRef]
11. Choy, J.T.; Amory, J.K. Nonsurgical Management of Oligozoospermia. *J. Clin. Endocrinol. Metab.* **2020**, *105*, e4194–e4207. [CrossRef]
12. Chiu, Y.-L.; Jhou, M.-J.; Lee, T.-S.; Lu, C.-J.; Chen, M.-S. Health Data-Driven Machine Learning Algorithms Applied to Risk Indicators Assessment for Chronic Kidney Disease. *Risk Manag. Healthc. Policy* **2021**, *14*, 4401–4412. [CrossRef]
13. Belladelli, F.; Boeri, L.; Pozzi, E.; Fallara, G.; Corsini, C.; Candela, L.; Cazzaniga, W.; Cignoli, D.; Pagliardini, L.; D'Arma, A.; et al. Triglycerides/Glucose Index Is Associated with Sperm Parameters and Sperm DNA Fragmentation in Primary Infertile Men: A Cross-Sectional Study. *Metabolites* **2022**, *12*, 143. [CrossRef]
14. Arafa, M.; Agarwal, A.; Majzoub, A.; Panner Selvam, M.K.; Baskaran, S.; Henkel, R.; Elbardisi, H. Efficacy of Antioxidant Supplementation on Conventional and Advanced Sperm Function Tests in Patients with Idiopathic Male Infertility. *Antioxidants* **2020**, *9*, 219. [CrossRef]
15. Akhter, M.S.; Hamali, H.A.; Iqbal, J.; Mobarki, A.A.; Rashid, H.; Dobie, G.; Madkhali, A.M.; Arishi, B.Y.H.; Ageeli, E.O.O.; Laghbi, O.S.H. Iron Deficiency Anemia as a Factor in Male Infertility: Awareness in Health College Students in the Jazan Region of Saudi Arabia. *Int. J. Environ. Res. Public Health* **2021**, *18*, 12866. [CrossRef]
16. Marill, K.A. Advanced statistics: Linear regression, part II: Multiple linear regression. *Acad Emerg Med.* **2004**, *11*, 94–102. [CrossRef]
17. Niazian, M.; Sadat-Noori, S.A.; Abdipour, M. Artificial neural network and multiple regression analysis models to predict essential oil content of ajowan (*Carum copticum* L.). *J. Appl. Res. Med. Aromat. Plants* **2018**, *9*, 124–131. [CrossRef]
18. Tenekedjiev, K.; Abdussamie, N.; An, H.; Nikolova, N. Regression Diagnostics with Predicted Residuals of Linear Model with Improved Singular Value Classification Applied to Forecast the Hydrodynamic Efficiency of Wave Energy Converters. *Appl. Sci.* **2021**, *11*, 2990. [CrossRef]
19. Miller, D.D.; Brown, E.W. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am. J. Med.* **2018**, *131*, 129–133. [CrossRef]
20. Liu, Y.; Chen, P.-H.C.; Krause, J.; Peng, L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA* **2019**, *322*, 1806–1816. [CrossRef] [PubMed]
21. Triantafyllidis, A.K.; Tsanas, A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. *J. Med. Internet Res.* **2019**, *21*, e12286. [CrossRef]
22. Peiffer-Smadja, N.; Rawson, T.M.; Ahmad, R.; Buchard, A.; Georgiou, P.; Lescure, F.X.; Birgand, G.; Holmes, A.H. Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. *Clin. Microbiol. Infect.* **2020**, *26*, 584–595. [CrossRef]
23. Song, Q.; Zheng, Y.-J.; Yang, J. Effects of Food Contamination on Gastrointestinal Morbidity: Comparison of Different Machine-Learning Methods. *Int. J. Environ. Res. Public Health* **2019**, *16*, 838. [CrossRef]
24. Wu, T.-E.; Chen, H.-A.; Jhou, M.-J.; Chen, Y.-N.; Chang, T.-J.; Lu, C.-J. Evaluating the Effect of Topical Atropine Use for Myopia Control on Intraocular Pressure by Using Machine Learning. *J. Clin. Med.* **2021**, *10*, 111. [CrossRef]
25. Huang, L.-Y.; Chen, F.-Y.; Jhou, M.-J.; Kuo, C.-H.; Wu, C.-Z.; Lu, C.-H.; Chen, Y.-L.; Pei, D.; Cheng, Y.-F.; Lu, C.-J. Comparing Multiple Linear Regression and Machine Learning in Predicting Diabetic Urine Albumin–Creatinine Ratio in a 4-Year Follow-Up Study. *J. Clin. Med.* **2022**, *11*, 3661. [CrossRef]
26. Shah, S.H.; Angel, Y.; Houborg, R.; Ali, S.; McCabe, M.F. A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. *Remote Sens.* **2019**, *11*, 920. [CrossRef]
27. Wang, H.; Xu, Q.; Zhou, L. Seminal Quality Prediction Using Clustering-Based Decision Forests. *Algorithms* **2014**, *7*, 405–417. [CrossRef]
28. Iqbal, I.; Mustafa, G.; Ma, J. Deep Learning-Based Morphological Classification of Human Sperm Heads. *Diagnostics* **2020**, *10*, 325. [CrossRef] [PubMed]
29. Liu, K.; Zhang, Y.; Martin, C.; Ma, X.; Shen, B. Translational Bioinformatics for Human Reproductive Biology Research: Examples, Opportunities and Challenges for a Future Reproductive Medicine. *Int. J. Mol. Sci.* **2023**, *24*, 4. [CrossRef]
30. Tseng, C.-J.; Lu, C.-J.; Chang, C.-C.; Chen, G.-D.; Cheewakriangkrai, C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artif Intell Med.* **2017**, *78*, 47–54. [CrossRef]

31. Ting, W.-C.; Chang, H.-R.; Chang, C.-C.; Lu, C.-J. Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Colorectal Cancer Survivors. *Appl. Sci.* **2020**, *10*, 1355. [CrossRef]
32. Lee, T.-S.; Chen, I.-F.; Chang, T.-J.; Lu, C.-J. Forecasting Weekly Influenza Outpatient Visits Using a Two-Dimensional Hierarchical Decision Tree Scheme. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4743. [CrossRef]
33. Chang, C.-C.; Yeh, J.-H.; Chen, Y.-M.; Jhou, M.-J.; Lu, C.-J. Clinical Predictors of Prolonged Hospital Stay in Patients with Myasthenia Gravis: A Study Using Machine Learning Algorithms. *J. Clin. Med.* **2021**, *10*, 4393. [CrossRef]
34. Chang, C.-C.; Huang, T.-H.; Shueng, P.-W.; Chen, S.-H.; Chen, C.-C.; Lu, C.-J.; Tseng, Y.-J. Developing a Stacked Ensemble-Based Classification Scheme to Predict Second Primary Cancers in Head and Neck Cancer Survivors. *Int. J. Environ. Res. Public Health* **2021**, *18*, 12499. [CrossRef]
35. Wu, C.-W.; Shen, H.-L.; Lu, C.-J.; Chen, S.-H.; Chen, H.-Y. Comparison of Different Machine Learning Classifiers for Glaucoma Diagnosis Based on Spectralis OCT. *Diagnostics* **2021**, *11*, 1718. [CrossRef]
36. Huang, Y.-C.; Cheng, Y.-C.; Jhou, M.-J.; Chen, M.; Lu, C.-J. Important Risk Factors in Patients with Nonvalvular Atrial Fibrillation Taking Dabigatran Using Integrated Machine Learning Scheme—A Post Hoc Analysis. *J. Pers. Med.* **2022**, *12*, 756. [CrossRef] [PubMed]
37. Jhou, M.-J.; Chen, M.-S.; Lee, T.-S.; Yang, C.-T.; Chiu, Y.-L.; Lu, C.-J. A Hybrid Risk Factor Evaluation Scheme for Metabolic Syndrome and Stage 3 Chronic Kidney Disease Based on Multiple Machine Learning Techniques. *Healthcare* **2022**, *10*, 2496. [CrossRef]
38. Sun, C.-K.; Tang, Y.-X.; Liu, T.-C.; Lu, C.-J. An Integrated Machine Learning Scheme for Predicting Mammographic Anomalies in High-Risk Individuals Using Questionnaire-Based Predictors. *Int. J. Environ. Res. Public Health* **2022**, *19*, 9756. [CrossRef]
39. Liao, P.-C.; Chen, M.-S.; Jhou, M.-J.; Chen, T.-C.; Yang, C.-T.; Lu, C.-J. Integrating Health Data-Driven Machine Learning Algorithms to Evaluate Risk Factors of Early Stage Hypertension at Different Levels of HDL and LDL Cholesterol. *Diagnostics* **2022**, *12*, 1965. [CrossRef] [PubMed]
40. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
41. Friedman, J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. Available online: http://www.jstor.org/stable/2699986 (accessed on 25 May 2022). [CrossRef]
42. Guindo, M.L.; Kabir, M.H.; Chen, R.; Liu, F. Particle Swarm Optimization and Multiple Stacked Generalizations to Detect Nitrogen and Organic-Matter in Organic-Fertilizer Using Vis-NIR. *Sensors* **2021**, *21*, 4882. [CrossRef]
43. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; CRC Press: Boca Raton, FL, USA, 2015. [CrossRef]
44. Kwon, S.; Lee, S.; Na, O. Tuning parameter selection for the adaptive Lasso in the autoregressive model. *J. Korean Stat. Soc.* **2017**, *46*, 285–297. [CrossRef]
45. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
46. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
47. Breiman, L.; Cutler, A. RandomForest: Breiman and Cutler's Random Forests for Classification and Regression. 2022. R Package Version, 4.7-1.1. Available online: https://CRAN.R-project.org/package=randomForest (accessed on 25 May 2022).
48. Greenwell, B.; Boehmke, B.; Cunningham, J. Gbm: Generalized Boosted Regression Models. 2020. R Package Version, 2.1.8. Available online: https://CRAN.R-project.org/package=gbm (accessed on 25 May 2022).
49. Friedman, J.; Hastie, T.; Tibshirani, R.; Narasimhan, B.; Tay, K.; Simon, N.; Qian, J.; Yang, J. Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. 2022. R Package Version, 4.1-4. Available online: https://CRAN.R-project.org/package=glmnet (accessed on 25 May 2022).
50. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: Extreme Gradient Boosting. 2022. R Package Version, 1.5.0.2. Available online: https://CRAN.R-project.org/package=xgboost (accessed on 1 January 2022).
51. Kuhn, M. Caret: Classification and Regression Training. 2022. R Package Version, 6.0-92. Available online: https://CRAN.R-project.org/package=caret (accessed on 25 May 2022).
52. Chen, Q.; Yang, H.; Zhou, N.; Sun, L.; Bao, H.; Tan, L.; Chen, H.; Ling, X.; Zhang, G.; Huang, L.; et al. Inverse U-shaped Association between Sleep Duration and Semen Quality: Longitudinal Observational Study (MARHCS) in Chongqing, China. *Sleep* **2016**, *39*, 79–86. [CrossRef] [PubMed]
53. Jensen, T.K.; Andersson, A.M.; Skakkebæk, N.E.; Joensen, U.N.; Blomberg Jensen, M.; Lassen, T.H.; Nordkap, L.; Olesen, I.A.; Hansen, Å.M.; Rod, N.H.; et al. Association of sleep disturbances with reduced semen quality: A cross-sectional study among 953 healthy young Danish men. *Am. J. Epidemiol.* **2013**, *177*, 1027–1037. [CrossRef]
54. Choi, J.H.; Lee, S.H.; Bae, J.H.; Shim, J.S.; Park, H.S.; Kim, Y.S.; Shin, C. Effect of sleep deprivation on the male reproductive system in rats. *J. Korean Med. Sci.* **2016**, *31*, 1624–1630. [CrossRef]
55. Yazama, F.; Tai, A. Unexpected role of α-fetoprotein in spermatogenesis. *PLoS ONE* **2011**, *6*, e19387. [CrossRef]
56. Corsini, C.; Fallara, G.; Candela, L.; Raffo, M.; Pozzi, E.; Belladelli, F.; Capogrosso, P.; Boeri, L.; Costa, A.; Schifano, N.; et al. High serum alpha-fetoprotein levels in primary infertile men. *Andrology* **2023**, *11*, 86–92. [CrossRef]

57. Jensen, T.K.; Andersson, A.M.; Jorgensen, N.; Andersen, A.G.; Carlsen, E.; Petersen, J.H.; Skakkebaek, N.E. Body mass index in relation to semen quality and reproductive hormones among 1,558 danish men. *Fertil. Steril.* **2004**, *82*, 863–870. [CrossRef] [PubMed]
58. Ergün, A.; Köse, S.K.; Aydos, K.; Ata, A.; Avci, A. Correlation of seminal parameters with serum lipid profile and sex hormones. *Arch. Androl.* **2007**, *53*, 21–23. [CrossRef] [PubMed]
59. Fogari, R.; Zoppi, A.; Preti, P.; Rinaldi, A.; Marasi, G.; Vanasia, A.; Mugellini, A. Sexual activity and plasma testosterone levels in hypertensive males. *Am. J. Hypertens.* **2002**, *15*, 217–221. [CrossRef] [PubMed]
60. Macdonald, A.; Stewart, A.; Farquhar, C. Body mass index in relation to semen quality and reproductive hormones in New Zealand men: A cross-sectional study in fertility clinics. *Hum. Reprod.* **2013**, *28*, 3178–3187. [CrossRef] [PubMed]
61. Edey, M.M. Male Sexual Dysfunction and Chronic Kidney Disease. *Front. Med.* **2017**, *4*, 32. [CrossRef] [PubMed]