Original Article

Assessment and Establishment of Correlation between Reactive Oxidation Species, Citric Acid, and Fructose Level in Infertile Male Individuals: A Machine-Learning Approach

Golnaz Shemshaki, Ashitha S. Niranjana Murthy[1], Suttur S. Malini

Departments of Studies in Zoology and [1]Genetics and Genomics, University of Mysore, Mysore, Karnataka, India

ABSTRACT

Background: Biochemical complexity of seminal plasma and obesity has an important role in male infertility (MI); so far, it has not been possible to provide evidence of clinical significance for all of them. **Aims:** Our goal here is to evaluate the correlation between biochemical markers with semen parameters, which might play a role in MI. **Study Setting and Design:** We enlisted 100 infertile men as patients and 50 fertile men as controls to evaluate the sperm parameters and biochemical markers in ascertaining MI. **Materials and Methods:** Semen analyses, seminal fructose, citric acid, and reactive oxidation species (ROS) were measured in 100 patients and 50 controls. **Statistical Analysis:** Descriptive statistics, an independent *t*-test, Pearson correlation, and machine-learning approaches were used to integrate the various biochemical and seminal parameters measured to quantify the inter-relatedness between these measurements. **Results:** Pearson correlation results showed a significant positive correlation between body mass index (BMI) and fructose levels. Citric acid had a positive correlation with sperm count, morphology, motility, and volume but displayed a negative correlation with BMI and basal metabolic rate (BMR). However, BMI and BMR had a positive correlation with ROS. Sperm count, morphology, and motility were negative correlations with ROS. The machine-learning approach detected that pH was the most critical parameter with an inverse effect on citric acid, and BMI and motility were the most critical parameter for ROS. **Conclusion:** We recommend that evaluation of biochemical markers of seminal fluid may benefit in understanding the etiology of MI based on the functionality of accessory glands and ROS levels.

Keywords: *Citric acid, fructose, linear regression, machine-learning, reactive oxygen species, support vector machine*

Introduction

Infertility is the third most serious disease of the 21st century. About 60–80 million couples suffer from infertility.[1] In India, the incidence of overall infertility is reported between 3.9% and 16.8%.[2] Approximately one-third of these cases include male infertility (MI) caused mainly due to sperm failure[3] and 30%–50% of infertility cases are idiopathic[4] triggered by molecular irregularities and obesity.

One of the significant causes of idiopathic infertility is oxidative stress (OS) which is defined as an imbalance between reactive oxidation species (ROS) and antioxidants.[5,6] Around 30%–40% of infertile men have increased ROS levels in seminal plasma.[7] Seminal fluid contains several components besides spermatozoa; semen

Received: 06-02-2021 **Revised:** 14-02-2021
Accepted: 23-03-2021 **Published:** 28-06-2021

Access this article online
Quick Response Code:
Website:
www.jhrsonline.org
DOI:
10.4103/jhrs.jhrs_26_21

Address for correspondence: Prof. Suttur S. Malini,
Department of Studies in Zoology, University of Mysore,
Mysore - 570 006, Karnataka, India.
E-mail: sutturmalini@yahoo.com

How to cite this article: Shemshaki G, Murthy AS, Malini SS. Assessment and establishment of correlation between reactive oxidation species, citric acid, and fructose level in infertile male individuals: A machine-learning approach. J Hum Reprod Sci 2021;14:129-36.

contains citric acid, fructose, etc.[8-10] Consequently, these biochemical secretions serve as markers of their respective glands.[11]

Fructose is an important energy source for sperm motility and is a marker for seminal vesicle function.[12,13] The lower level of fructose oxidation intensity in gamete mitochondria leads to lactate accumulation and dehydrogenase activity inhibition.[14] Citric acid is an essential biochemical component of seminal plasma associated with semen coagulation and motility[15,16] and a diagnostic tool to evaluate secretory dysfunction of the prostate.[17] Accordingly, fructose is used as a marker for seminal vesicle, and citric acid is used as a marker for prostate and both have a critical role in MI.

Machine learning (ML) has been used in different fields,[18-22] including genetics and genomics,[23-26] limited studies used ML methods to predict infertility based on experimental data.

In the previous researches, ML was added to MI to integrate sperm, blood, and environmental parameters to determine their association[27] and to estimate OS levels from a database of biochemical analyses of OS biomarkers in blood, plasma, and urine.[28,29] Dubey *et al*., conducted a quantitative analysis of human spermatozoa under OS conditions using ML.[5] These studies used ML methods to a limited extent leaving a gap in the utilization of its full strength in determining important parameters in MI. Furthermore, to date, most biochemical markers have been assessed uni-dimensionally against individual parameters such as sperm count and body mass index (BMI), but their combined effects have not been examined.

To address this limitation, we used ML approaches to integrate the various biochemical and seminal parameters measured to estimate the levels of ROS, citric acid, and fructose in human subjects. This approach would further determine the degree of inter-relatedness; quantify the interdependency between these variables; and aid in the prioritization of the variable producing the most severe effect on infertility.

Material and Methods

Semen collection and examination

A case–control study was conducted among 100 infertile men and 50 control groups at the University of Mysore, India, between September 2018 and October 2020. Based on previous research, we assumed that the approximate prevalence of MI is around 5.6%,[2,3] and the following simple formula will be used to calculate the acceptable sample size in a prevalence study: Where n is the sample size, Z is the statistic corresponding to the level of confidence, P is expected prevalence, and d is precision (corresponding to impact size).[30] According to the formula, our minimum sample size with a 95% confidence level should be 82 individuals, but we opted to increase the sample size to 100 to get more accurate results.

$$n = Z^2 P\ (1 - P)/d^2 \qquad (1)$$

Semen samples were collected after 3–7 days of ejaculatory abstinence. Semen analysis was performed according to the WHO criteria. Men having known clinical factors including genetic factors which affect fertility status were excluded. The controls consisted of 50 fertile men who had at least one child and had normal semen analysis. All participants gave written informed consent to participate in this study which was approved by the Institutional Human Ethics Committee (IHEC-UOM No. 152/Ph. D/2017-18), University of Mysore.

Estimation of seminal fructose

Estimation of seminal fructose was done by Karvonen and Malm (1955) method. 20 µl of seminal plasma was diluted with 220 µl to distilled water and mixed by adding 50 µl of ZnSO4 and 50 µl NaOH. This mixture was incubated for 15 min at room temperature and centrifuged at 2500 RPM for 15 min. 200 µl of the cleared supernatant was mixed with indole reagent followed by 32% hydrochloric acid. Reading was taken at 470 nm after cooling of 10 min.[31]

Estimation of seminal citric acid

One hundred microliter of seminal plasma was added to 100 µl of 50% trichloroacetic subjected to centrifugation at 7000 rpm for 15 min and supernatant was obtained. Eight hundred microliter of anhydrous acetic anhydride was added and incubated in a water bath at 60°C for 10 min in the supernatant. One milliliter of dry reagent grade pyridine was added and incubated at 60°C for 40 min and absorbance was measured at 400 nm.[32]

Estimation of reactive oxidation species by chemiluminescence method

Liquified semen was centrifuged at 3000 rpm for 7 min, and the seminal plasma was separated. The pellet was washed with phosphate buffer saline (PBS) and washed in wash media at a concentration of 20 × 106 sperm/ml. Four hundred microliter of suspension aliquots was used. Ten microliter of luminol prepared as a 5 mm stock of dimethylsulfoxide was added to the mixture used as a probe. As a negative control, 10 µl of 5 mM of luminol with 400 µl of PBS was used. Samples were loaded in 96-welled noncoated microplates, and readings were taken in the Thermo scientific multimode plate reader.

Statistical analysis

The data collected were statistically analyzed and expressed as a mean value with deviations (± standard deviation). An independent *t*-test and Pearson correlation were used to assess if there is a significant mean difference between patients and controls ($P < 0.05$).

Machine-learning methods

Four different ML methods such as linear regression (LR), artificial neural network (ANN), support vector machine (SVM), and random Forrest (RF) were used and compared to predict the density of citric acid and ROS.

LR models generate a linear relationship between two sets of variables, dependent and independent [Supplementary Material Equation 1].[33,34] ANN is a tool to model the neurons in the human brain[35] and consists of three layers [Supplementary Material Equations 2 and 3].[35-37] SVM is a supervised learning method used for classification, regression, and density estimation problems.[36] RF method can handle high-dimensional data and use a large number of trees in the ensemble.[38-40]

Although visual exploration of graphs helps identify the performance of the models, scientific error indices will help determine the effectiveness of the models numerically. For this purpose, to find out the difference between the mean of observed and predicted values, the bias index was computed [Supplementary Material Equations 4-7].

Results

The mean age of patients and controls was 34.6 ± 6.5 and 32.5 ± 3.2 years, respectively. Infertile men diagnosed based on sperm count, motility, and morphology were classified into nine subgroups [Table 1]. All MI groups showed a significant mean difference in sperm count and motility compared to controls and all subgroups except idiopathic demonstrated significant differences for mean fructose concentration compared to controls. Fructose concentration was highest in oligoteratozoospermia (164 ± 15.2) and lowest in asthenozoospermia (52.2 ± 11.2) compared to other MI subgroups. Interestingly, the idiopathic subgroup displayed the highest citric acid concentration compared to controls.

All infertile subgroups except azoospermia showed increased ROS level with statistical significance between groups. Teratozoospermia showed increased ROS, followed by oligoasthenozoospermia and OAT. However, when all infertile subgroups were combined and compared to controls using oneway ANOVA, there was a significant difference in sperm count, motility, citric acid level, fructose level, and ROS compared to controls [Table 2]. Pearson correlation results exhibited a significant positive correlation among BMI ($r = 0.295$), basal metabolic rate (BMR) ($r = 0.279$), and fructose levels.

Seminal citric acid concentration had a positive correlation with sperm count (r = 0.471), morphology ($r = 0.519$), motility (0.294), and volume (0.236), whereas it had a negative correlation with BMI (−0.576) and BMR (−0.383). While BMI ($r = 0.637$) and BMR ($r = 0.371$) showed a significant positive correlation with ROS, sperm count (−0.361), morphology (−0.506), and motility (-0.398) showed a significant negative correlation with ROS [Table 3].

Manual estimation of biochemical parameters is time consuming and involves intensive preparation. In addition, due to restricted reproducibility and high interpersonal variation, the validity of manual biochemical assay has been challenged. All presented ML methods provided predictions within 5 min, including time-consuming data preparation. This is much easier than manual biochemical assessments, requires less time, and no repetitions which save reagents making it cost-effective.

Four different ML methods were applied on 70% of both citric acid and ROS data to train the models, and the trained models were used for the prediction of the remaining 30% of data. In Figure 1a, the condensed data around the bisector indicated that all models showed high accuracy for citric acid as the predicted data were close to each other, and ANN outperformed the rest. Contrastingly, visual inspection of the ROS graph [Figure 1b] showed that the RF model's predicted values outperform the rest since the predictions from this model are more condense around the bisector.

In Table 4, all four models displayed similar

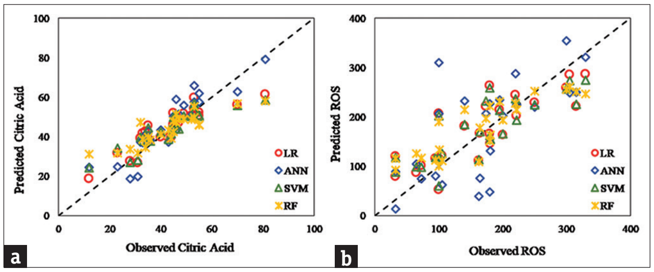

Figure 1: Predicted values vs. observed values of (a) citric acid, and (b) reactive oxidation species using linear regression, artificial neural network, support vector machine, and random Forrest machine-learning methods

Table 1: Seminal parameters and biochemical marker concentration of individual infertile groups and controls

Patients	*n*	Age (years)	BMI	pH	Sperm volume	Sperm count (mil/ml)	Sperm motility (%)	Sperm morphology	Fructose (≥13 µmole/ ejaculate)	Citric acid (≥13 µmole/ ejaculate)	ROS
AS	22	34.8±4.9	25.2±3.9	7.5±0.5	1.8±0.9	39.6±3.5*	20.5±6.3*	30.6±3.5	52.2±11.2*	38.6±6.9*	159±61.7*
AZ	17	34.2±6.3	26.3±3.5	7.6±0.5	1.7±0.7	0.00±0.0*	0.00±0.0*	0.00±0.0	129±5.4	32.7±3.5*	0.00±0.0*
ID	21	36.3±5.2	25.7±4.3	7.5±1.7	2.8±1.5	59.9±23.4*	57.7±15.4	46.1±9.01	126.9±6.9*	53.28±10.2	114±45.7
OL	12	34.3±6.5	26±3.4	7.6±0.5	2.1±1.3	9.4±3.9*	57.4±16	42.6±4.8	111±8.5*	38.7±6.4*	160±66.5*
OA	14	34±5.6	27.1±3.5	7.6±0.4	2.6±1.1	8.5±2.5*	13.5±10.2*	41±6.6	108±6.5*	41.2±8.1	248±67.3*
AT	2	37.5±0.7	28.4±6.0	8±0.7	3.7±3.1	45±21.2*	25±0.00*	18±1.4	115±7.7	41±11.3	216±94*
OAT	4	31.2±3.3	24±1.4	7.6±0.7	2.4±1.7	5.5±3.3*	13.5±7.6*	12.2±4	129±14.8	44±13.9	245±98*
OT	5	31.8±3.8	26.8±3.1	7.8±0.3	1.6±1.4	9.4±1.9*	65±7.07*	17.2±3.9	164±15.2*	22.8±8.1*	212±67*
T	3	35.3±11	25.8±1.2	7.5±0.8	2.4±1.9	50±20*	46.6±11.5*	15±3.6	137±17*	45±6.02	265±50*
Control	50	32.5±3.2	25±2.1	7.5±0.1	2.3±0.2	68.7±6.3*	54.2±1.9	30.1±2.1	126.4±4.6	49.4±1.2	71.4±4.0

*$P<0.05$ defines the level of significance. All values are presented as mean±SD. SD=Standard deviation, AS=Asthenozoospermia, AZ=Azoospermia, ID=Idiopathic, OA=Oligoasthenozoospermia, AT=Astenoteratozoospermia, OL=Oligozoospermia, T=Teratozoospermia, ROS=Reactive oxidation species, BMI=Body mass index

Table 2: Descriptive value of Fertility status and seminal biochemistry in fertile and infertile individuals

Patients (*n*)	Sperm count (mil/ml)	Sperm motility (%)	Fructose (≥13 µmole/ ejaculate)	Citric acid (≥13 µmole/ ejaculate)	ROS
Infertile	26.7±27.03	31.1±25.5	105±0.31.07	40.7±10.9	145±97.4
Control	68.7±6.3	54.2±1.9	126.4±4.6	49.4±1.2	71.4±4.05
Significant level 0.05	<0.05*	<0.05*	<0.05*	<0.05*	<0.05*

Sig level (P): *$P<0.05$ defines the level of significance. All values are presented as mean±SD. SD=Standard deviation, ROS=Reactive oxidation species

Table 3: Pearson correlation coefficient among the study variables

	ROS	Citric acid	Fructose
Age	−0.043	0.064	0.069
BMI	0.637*	−0.576*	0.295*
Count	−0.361*	0.471*	−0.139
Morphology	−0.506*	0.519	−0.168
Motility	−0.398*	0.294*	0.136
Volume	−0.082	0.236**	0.117
BMR	0.371*	−0.383*	0.279*

*$P<0.05$ defines the level of significance, **$P<0.01$ defines the level of significance. ROS=Reactive oxidation species, BMI=Body mass index, BMR=Basal metabolic rate

Table 4: Error indices for predicted values of citric acid

Model	Bias	*r*	RMSE	SI
LR	0.50	0.89	6.81	0.16
ANN	2.29	0.90	6.59	0.15
SVM	0.27	0.89	7.22	0.17
RF	0.25	0.87	8.16	0.19

RMSE=Root means square error, SI=Scatter index, LR=Linear regression, ANN=Artificial neural network, SVM=Support vector machine, RF=Random forrest

performances. Smaller root means square error (RMSE) values from the ANN model also confirmed that the spread of predicted values was smaller than other models. Finally, the smallest value of the scatter index (SI) index for the ANN model compared to the other models showed that the ANN model predicted the citric Acid values better than the other three models. These indices confirmed the visual findings from Figure 1a where ANN and RF models resulted in slightly denser and more spread values around the bisector, respectively. Compared to other models, the most extensive bias index of the ANN model could be meaningful if the other three indices follow the same trend.

According to coefficient index (*r*) values, the RF model predicts the ROS values 5%, 28%, and 5% better than SVM, ANN, and LR, respectively. The SI index for LR, SVM, and RF has a similar value of 0.28, which concerning *r* and RMSE depict that the RF model performs better in predicting ROS values among all models [Table 5]. These indices confirm the visual findings from Figure 1b.

Out of four investigated ML methods, LR and SVM can present a relationship between the dependent and independent variables. First, these relationships show the critical parameters, and second, show the importance of each independent parameter in predicting the dependent parameter. The previous results showed that both these models have very similar performances in predicting citric acid and ROS.

Table 5: Error indices for predicted values of reactive oxidation species

Model	Bias	*r*	RMSE	SI
LR	5.10	0.83	47.60	0.28
ANN	4.93	0.68	73.56	0.44
SVM	5.31	0.83	48.12	0.28
RF	9.28	0.87	46.52	0.28

RMSE=Root means square error, SI=Scatter index, LR=Linear regression, ANN=Artificial neural network, SVM=Support vector machine, RF=Random forrest

The LR model presented the eq. 2 to predict citric acid concentration based on the independent

parameters as:

$$\text{Citric Acid}_{\text{LR}} = -9.460 \times \text{pH} + 0.153 \times \text{Count} - 0.148 \times \text{Motility} + 0.198 \times \text{Morphology} + 0.275 \times \text{Fat Percentage} - 1.651 \times \text{BMI} + 0.008 \times \text{BMR} + 0.162 \times \text{Fructose} + 112.214 \quad (2)$$

In addition, the SVM model presented the eq. 3 for the weight of each independent parameter in

predicting acid citric as:

$$\widehat{\text{Citric Acid}_{\text{SVM}}} = 0.021 \times \widehat{\text{Age}} - 0.035 \times \widehat{\text{Volume}} - 0.260 \times \widehat{\text{pH}} + 0.177 \times \widehat{\text{Count}} - 0.175 \times \widehat{\text{Motility}} + 0.160 \times \widehat{\text{Morphology}} + 0.161 \times \widehat{\text{Fat Percentage}} - 0.348 \times \widehat{\text{BMI}} + 0.001 \times \widehat{\text{BMR}} + 0.242 \times \widehat{\text{Fructose}} + 0.452 \quad (3)$$

ˆ Denotes the normalized parameter.

Comparison of the two models presented for the citric acid shows that the LR model detected a smaller number of parameters as important ones; however, the SVM considered all the parameters and allocated weights to each of them based on their importance.

The LR model detected the pH as the most critical parameter (coefficient of −9.46) with an inverse effect, followed by BMI (coefficient of −1.651) with inverse effect, fat percentage (coefficient of 0.275) with direct effect, and morphology (coefficient of 0.198) with direct effect. The LR model indicated parameters such as count, fructose, and motility which were at the almost same level of importance (coefficients equal to 0.153, 0.162, and −0.148) with direct, direct, and inverse effects on citric acid, respectively. The BMR with direct effect was detected as the least essential parameter.

On the other side, the SVM model allocated a small weight (compared to other parameters) of 0.021 and –0.035 to age and volume, respectively. Other than these two parameters, the SVM model detected the BMI as the most critical parameter (weight of −0.348) with an inverse effect, followed by pH (weight of −0.260) with an inverse effect, and fructose (weight of 0.242) with a direct effect. This model detected the motility, count, and fat percentage with almost equal importance (weights of −0.175, 0.177, and 0.161) with inverse, direct, and direct effects, respectively. The SVM also detected the BMR as the least essential parameter with direct effect.

Similar to the prediction of the citric acid, the LR model presented the Equation 4 to predict ROS based on the independent parameters as:

$$\text{ROS}_{\text{LR}} = -1.457 \times \text{Motility} - 1.773 \times \text{Morphology} + 6.469 \times \text{BMI} - 0.067 \times \text{BMR} + 1.109 \times \text{Fructose} - 2.421 \times \text{Citric Acid} + 230.236 \quad (4)$$

In addition, the SVM model presented the eq. 5 for the weight of each independent parameter in

predicting ROS as:

$$\widehat{\text{ROS}_{\text{SVM}}} = -0.0036 \times \widehat{\text{Age}} - 0.0061 \times \widehat{\text{Volume}} + 0.165 \times \widehat{\text{pH}} - 0.061 \times \widehat{\text{Count}} - 0.392 \times \widehat{\text{Motility}} - 0.295 \times \widehat{\text{Morphology}} - 0.023 \times \widehat{\text{Fat Percentage}} + 0.259 \times \widehat{\text{BMI}} - 0.138 \times \widehat{\text{BMR}} + 0.376 \times \widehat{\text{Fructose}} - 0.362 \times \widehat{\text{Citric Acid}} + 0.581 \quad (5)$$

Comparing the standard parameters between the two models shows that both the models have similar detection of each parameter's effect, either having a direct or inverse effect. The LR model detected the BMI (coefficient of 6.469) as the most critical parameter in predicting ROS followed by citric acid (coefficient of −2.421), morphology (coefficient of −1.773), motility (coefficient of-1.457), fructose (coefficient of 1.109), and BMR (coefficient of −0.067). On the other hand, the SVM model detected motility (weight of −0.392) as the most critical parameter with an inverse effect in predicting ROS. After that, fructose and citric acid (weights of 0.376 and −0.362) have almost similar effect weight with opposite signs (direct and inverse, respectively). Morphology, BMI, pH, and BMR have similar effects (weights of −0.295, 0.259, 0.165, and −0.138) on predicting ROS with inverse, direct, direct, and inverse effects, respectively. Similar to LR models, the SVM model detected the count, volume, and age as the least essential effects with little inverse effects.

Discussion

The present study aimed to ascertain the correlation between seminal biochemical markers and sperm parameters in infertile patients and controls. Our result showed that asthenozoospermia and associated conditions showed decreased seminal fructose levels compared to controls. The motility of spermatozoa is closely connected with fructose breakdown.[41,42] Another possible reason could be explained by partial or complete obstruction of the seminal ducts of the accessory glands that secrete fructose.[43] In this study, oligoteratozoospermia condition showed an increase in seminal fructose than normal. This could be either because of abridged sperm count, abnormal sperm morphology, and low utilization by spermatozoa with morphological defects and decreased sperm activity leading to the accumulation of fructose in the semen which gets detected in our fructose test.[44,45]

The primary role of citric acid is maintaining semen pH.[42,46] It is a reliable biomarker of prostatic function and it plays a crucial role in balancing the osmotic equilibrium of semen that influences the membrane function and sperm morphology.[47-49] In this study, decreased citric acid levels were observed in azoospermia, oligoteratozoospermia, and asthenozoospermia groups. The reduced citric acid in semen could be due to inflammation, acute or chronic prostatitis, any partial or complete obstruction of the ejaculatory ducts, and prostate cancer.[45]

ROS plays a dual role in male reproduction, supporting and activating the physiological roles of sperm at basic levels while causing significant detrimental effects on male fertility at elevated concentrations.[50] In previous studies, semen parameters are reportedly affected by the oxidation of cellular components and activation of the apoptotic pathway.[51,52] This was caused by excessive ROS generation/failure of the antioxidant system. ROS can induce apoptosis in germ cells, leading to decreased sperm counts. In this study, increased ROS was observed in oligoasthenoteratozoospermia, oligoasthenozoospermia, oligoteratozoospermia, and teratozoospermia groups and also showed increased ROS in asthenoteratozoospermia and oligoasthenozoospermia groups with decreased motility, which could be attributed to sperm carrying dysfunctional mitochondria contributing to decreased motility, in turn, increased ROS production.

ML models have an exceptional inherent accuracy in predicting *in vivo* outcomes than existing *in vitro* assays, making it a powerful tool for linking the dependent variable (ROS) to multiple independent variables. All four ML models resulted in the accurate prediction of citric acid where all models' accuracy was close to each other, although the ANN model slightly outperformed others. On the other hand, RF, in the prediction of ROS, thoroughly performed better than the other three models.

The LR model detected pH as the most critical parameter, followed by BMI, fat percentage, and morphology. On the other side, SVM model detected BMI as the most critical parameter, followed by pH and fructose.

Relationships presented for ROS using LR and SVM models showed that The LR model detected the BMI as the most critical parameter in predicting ROS followed by citric acid, morphology, motility, fructose, and BMR. On the other hand, the SVM model detected motility as the most critical parameter with an inverse effect in predicting ROS. After that, fructose and citric acid have an almost similar level of effect.

We, therefore, tried to figure out which parameters affect more ROS and citric acid in simultaneous conditions and as a result, BMI and pH are the most powerful parameters for ROS and citric acid, respectively.

The present study is the first research that used ML approaches to integrate the various biochemical and seminal parameters measured to estimate the levels of OS. In this study, one of the limitations was the lack of knowledge about the patient's lifestyle (cigarette smoke, pollutants, and heavy metals) and assessment of hormone imbalance. Moreover, our results must be verified in other larger populations with the use of different techniques as a further objective.

Conclusion

Our data reveal a complex relationship between fructose, citric acid, and ROS with sperm parameters and can be used for the recognition of biological attributes of semen. More attention should be paid to the function of seminal vesicles, and therefore, evaluation of certain biochemical markers of seminal fluid may benefit in understanding the functionality of accessory glands, which subsidizes significantly to the seminal volume. We demonstrated that BMI has a significant effect on ROS compared to other parameters, and ROS significantly affects sperm quality leading to a decline in IUI and IVF success. Thereby, increased attention to obesity treatment may help in improving the success rates of MI treatment strategies. Through the ML method, we found a hidden link between BMI and various biochemical and semen parameters.

Acknowledgments

We acknowledge all participants of our study for their valuable samples. We thank Dr. N. B. Ramachandra for his help and support during the research work. We thank the Department of Studies in Genetics and Genomics, University of Mysore, Manasagangotri, Mysuru, India.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Rutstein SO, Shah IH. Infecundity, infertility, and childlessness in developing countries. Calverton, Maryland, USA ORC Macro, MEASURE DHS; 2004.
2. Mehrotra, A., Katiyar, D.K., Agarwal, A., Das, V. and Pant, K.K., 2013. Role of total antioxidant capacity and lipid peroxidation in fertile and infertile men.
3. Krausz C. Male infertility: Pathogenesis and clinical diagnosis. Best Pract Res Clin Endocrinol Metab 2011;25:271-85.
4. Jungwirth A, Giwercman A, Tournaye H, Diemer T, Kopa Z, Dohle G, *et al*. European Association of Urology Guidelines on Male Infertility: The 2012 update. Eur Urol 2012;62:324-32.
5. Sies H. Oxidative stress: A concept in redox biology and medicine. Redox Biol 2015;4:180-3.
6. Dubey V, Popova D, Ahmad A, Acharya G, Basnet P, Mehta DS, *et al*. Partially spatially coherent digital holographic microscopy and machine learning for quantitative analysis of human spermatozoa under oxidative stress condition. Sci Rep 2019;9:3564.
7. Lanzafame FM, La Vignera S, Vicari E, Calogero AE. Oxidative stress and medical antioxidant treatment in male infertility. Reprod Biomed Online 2009;19:638-59.
8. Orth JM. Proliferation of Sertoli cells in fetal and postnatal rats: A quantitative autoradiographic study. Anat Rec 1982;203:485-92.
9. Wein AJ, Kavoussi LR, Novick AC, Partin AW, Peters CA. Campbell-Walsh urology: expert consult premium edition: enhanced online features and print, 4-volume set. Philadelphia, PA.Elsevier Health Sciences 2011.
10. Plant TM, Zeleznik AJ, editors. Knobil and Neill's physiology of reproduction. Pittsburgh, PA, USA .Academic Press; 2014.
11. Aumüller G, Riva A. Morphology and functions of the human seminal vesicle. Andrologia 1992;24:183-96.
12. Gonzales GF, Villena A. True corrected seminal fructose level: A better marker of the function of seminal vesicles in infertile men. Int J Androl 2001;24:255-60.
13. Videla E, Blanco AM, Galli ME, Fernández-Collazo E. Human seminal biochemistry: Fructose, ascorbic acid, citric acid, acid phosphatase and their relationship with sperm count. Andrologia 1981;13:212-4.
14. Artifeksov SB. The biochemical characteristics of the sperm in patients with varicocele. Urol Nefrol (Mosk) 1991:50-2.Russian.
15. Toragall MM, Satapathy SK, Kadadevaru GG, Hiremath MB. Evaluation of seminal fructose and citric acid levels in men with fertility problem. J Hum Reprod Sci 2019;12:199-203.
16. Marberger H, Marberger E, Mann T, Lutwak-Mann C. Citric acid in human prostatic secretion and metastasizing cancer of prostate gland. Br Med J 1962;1:835-6.
17. Said L, Galeraud-Denis I, Carreau S, Saâd A. Relationship between semen quality and seminal plasma components: alpha-glucosidase, fructose and citrate in infertile men compared with a normospermic population of Tunisian men. Andrologia 2009;41:150-6.
18. Chen H. Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. J Am Soc Inform Sci 1995;46:194-216.
19. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001;34:28-36.
20. Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor 2015;18:1153-76.
21. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5:13087.
22. Xin Y, Kong L, Liu Z, Chen Y, Li Y, Zhu H, *et al*. Machine learning and deep learning methods for cybersecurity. IEEE Access 2018;6:35365-81.
23. Chang SW, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinformatics 2013;14:170.
24. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321-32.
25. Li B, Zhang N, Wang YG, George AW, Reverter A, Li Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 2018;9:237.
26. González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S and Crossa J. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. The plant genome2018;11: p.170104.
27. 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.
28. de la Villehuchet AM, Brack M, Dreyfus G, Oussar Y, Bonnefont-Rousselot D, Chapman MJ, *et al*. A machine-learning approach to the prediction of oxidative stress in chronic inflammatory disease. Redox Rep 2009;14:23-33.
29. Zhang W, Rhodes JS, Garg A, Takemoto JY, Qi X, Harihar S, *et al*. Label-free discrimination and quantitative analysis of oxidative stress induced cytotoxicity and potential protection of antioxidants using Raman micro-spectroscopy and machine learning. Anal Chim Acta 2020;1128:221-30.
30. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterol Hepatol Bed Bench 2013;6:14-7.
31. Karvonen MJ, Malm M. Colorimetric determination of fructose with indol. Scand J Clin Lab Invest 1955;7:305-7.
32. Polakoski, K. L. and Zaneveld, L. J. Biochemical examination of the human ejaculate. In Techniques of Human Andrology (ed. E. S. E. Hafez), 1977. pp. 265-286. Amsterdam: Elsevier/ North-Holland PressSu Y, Gao X, Li X, Tao D. Multivariate multilinear regression. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2012;42:1560-73.
33. Su Y, Gao X, Li X, Tao D. Multivariate multilinear regression. IEEE Trans Syst Man Cybern B Cybern 2012;42:1560-73.
34. Forkuor G, Hounkpatin OK, Welp G, Thiel M. High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: A comparison of machine

learning and multiple linear regression models. PLoS One 2017;12:e0170478.

35. Kamranzad B, Etemad-Shahidi A, Kazeminezhad MH. Wave height forecasting in Dayyer, the Persian Gulf. Ocean Eng 2011;38:248-55.
36. Mafi S, Amirinia G. Forecasting hurricane wave height in Gulf of Mexico using soft computing methods. Ocean Eng 2017;146:352-62.
37. Jain P, Deo MC. Real-time wave forecasts off the western Indian coast. Appl Ocean Res 2007;29:72-9.
38. Breiman L. Random forests. Mach Learn 2001;45:5-32.
39. Gislason PO, Benediktsson JA, Sveinsson JR. Random forests for land cover classification. Pattern Recognit Lett 2006;27:294-300.
40. Van Beijma S, Comber A, Lamb A. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. Remote Sens Environ 2014;149:118-29.
41. Amidu N, Owiredu WK, Bekoe MA, Quaye L. The impact of seminal zinc and fructose concentration on human sperm characteristic. J Med Biomed Sci 2012;1:14-20.
42. Ahmed Z, Khan MS, Khan MA, Ul Haq A, Ur Rahman J. Seminal fructose in various classes of infertile patients. Pak J Physiol 2010;6:36-8.
43. Buckett WM, Lewis-Jones DI. Fructose concentrations in seminal plasma from men with nonobstructive azoospermia. Arch Androl 2002;48:23-7.
44. Rajalakshmi M, Sharma RS, David GF, Kapur MM. Seminal fructose in normal and infertile men. Contraception 1989;39:299-306.
45. Abdella AM, Omer AF, Al-Aabed BH. Biochemical markers in semen and their correlation with fertility hormones and semen quality among Sudanese infertile patients. Afr J Biochem Res 2010;4:255-60.
46. Obidoa O, Ezeanyika LU, Okoli AH. Effect of scopoletin on male guinea pig reproductive organs. I. Levels of citric acid and fructose. Nutr Res 1999;19:443-8.
47. Gavella M. A simple automated method for determination of citric acid levels in semen. Int J Androl 1983;6:585-91.
48. Das S, Parveen S, Kundra CP, Pereira BM. Reproduction in male rats is vulnerable to treatment with the flavonoid-rich seed extracts of Vitex negundo. Phytother Res 2004;18:8-13.
49. Said L, Galeraud-Denis I, Carreau S, Saad A. Relationship between semen quality and seminal plasma components: Alpha-glucosidase, fructose and citrate in infertile men compared with a normospermic population of Tunisian men. Andrologia 2009;41:150-6.
50. Alahmar AT. Role of oxidative stress in male infertility: An updated review. J Hum Reprod Sci 2019;12:4-18.
51. Bardaweel SK, Gul M, Alzweiri M, Ishaqat A, ALSalamat HA, Bashatwah RM. Reactive oxygen species: The dual role in physiological and pathological conditions of the human body. Eurasian J Med 2018;50:193-201.
52. Vessey W, Saifi S, Sharma A, McDonald C, Almeida P, Figueiredo M, *et al*. Baseline levels of seminal reactive oxygen species predict improvements in sperm function following antioxidant therapy in men with infertility. Clin Endocrinol (Oxf) 2021;94:102-10.

Supplementary Material

Appendix

Equation 1 shows the general form of LR models:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \in \tag{1}$$

Where, y is the dependent variable, x_i represents the ith independent variable, β_0 is the constant value(y-intercept), β_i is the coefficient for ith independent variable, and ϵ is the model error.

Equations 2 and 3 presents the relation between input and output parameters as:

$$P = 1/[1 + e^s] \tag{2}$$

$$s = \sum_{i=1}^{n} \alpha_i w_i + B \tag{3}$$

Where P is the output of each node, a_i is the input value, w_i is the weight, and B is the bias. By suitable training the ANN model and by adjusting the weights, the overall error between the outputs and actual observations should be reduced.

$$Bias = \overline{y} - \overline{x} \tag{4}$$

Where $\overline{x}$ and $\overline{y}$ represent the mean observed and predicted values, respectively. To explore the correlation of the observed and predicted values, correlation coefficient index, r was used as:

$$r = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^N (x_i - \overline{x})^2 \sum_i^N (y_i - \overline{y})^2}} \tag{5}$$

Where x_i and y_i represent the observed and predicted values at ith data respectively, and N is the number of observations. To measure the spread of predicted values, the root mean square error, *RMSE* was used as:

$$RMSE = \sqrt{\frac{\sum_i^N (x_i - y_i)^2}{N}} \tag{6}$$

In addition, to measure the percentage of *RMSE* difference with respect to mean observation, scatter index, *SI* was computed as:

$$SI = \frac{RMSE}{\overline{x}} \tag{7}$$