Received: 20 November 2021 | Revised: 10 December 2021 | Accepted: 10 December 2021

DOI: 10.1111/andr.13141

ANDROLOGY WILEY

ORIGINAL ARTICLE

Machine learning based prediction models in male reproductive health: Development of a proof-of-concept model for Klinefelter Syndrome in azoospermic patients

Henrike Krenz[1] | Andrea Sansone[2,3] | Michael Fujarski[1] | Claudia Krallmann[2] | Michael Zitzmann[2] | Martin Dugas[5] | Sabine Kliesch[2] | Julian Varghese[1] | Frank Tüttelmann[4] | Jörg Gromoll[2]

[1] Institute of Medical Informatics, University of Münster, Münster, Germany

[2] Centre of Reproductive Medicine and Andrology, University Hospital Münster, Münster, Germany

[3] Chair of Endocrinology and Sexual Medicine (ENDOSEX), Department of Systems Medicine, University of Rome Tor Vergata, Rome, Italy

[4] Institute of Reproductive Genetics, University of Münster, Münster, Germany

[5] Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany

Correspondence
Jörg Gromoll, Centre of Reproductive Medicine, Albert-Schweitzer-Campus D11, 48149 Münster, Germany.
Email: joerg.gromoll@ukmuenster.de

Henrike Krenz, Andrea Sansone contributed equally.

Frank Tüttelmann, Jörg Gromoll share last authorship.

Funding information
German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) Clinical Research Unit "Male Germ Cells: from Genes to Function", Grant/Award Number: CRU326

Abstract

Background: Due to the highly variable clinical phenotype, Klinefelter Syndrome is underdiagnosed.

Objective: Assessment of supervised machine learning based prediction models for identification of Klinefelter Syndrome among azoospermic patients, and comparison to expert clinical evaluation.

Materials and methods: Retrospective patient data (karyotype, age, height, weight, testis volume, follicle-stimulating hormone, luteinizing hormone, testosterone, estradiol, prolactin, semen pH and semen volume) collected between January 2005 and June 2019 were retrieved from a patient data bank of a University Centre. Models were trained, validated and benchmarked based on different supervised machine learning algorithms. Models were then tested on an independent, prospectively acquired set of patient data (between July 2019 and July 2020). Benchmarking against physicians was performed in addition.

Results: Based on average performance, support vector machines and CatBoost were particularly well-suited models, with 100% sensitivity and >93% specificity on the test dataset. Compared to a group of 18 expert clinicians, the machine learning models had significantly better median sensitivity (100% vs. 87.5%, $p = 0.0455$) and fared comparably with regards to specificity (90% vs. 89.9%, $p = 0.4795$), thereby possibly improving diagnosis rate. A Klinefelter Syndrome Score Calculator based on the prediction models is available on http://klinefelter-score-calculator.uni-muenster.de.

Discussion: Differentiating Klinefelter Syndrome patients from azoospermic patients with normal karyotype (46,XY) is a problem that can be solved with supervised machine learning techniques, improving patient care.

Conclusions: Machine learning could improve the diagnostic rate of Klinefelter Syndrome among azoospermic patients, even more for less-experienced physicians.

KEYWORDS
azoospermia, Klinefelter Syndrome, machine learning, prediction models, reproductive genetics, reproductive health

1 | INTRODUCTION

By combining the benefits of statistics, data processing and computer science, supervised machine learning (sML) bears the potential to offer powerful support systems for medical decision-making. Common ways sML is currently used in medicine include identifying thyroid or lung nodules in ultrasound or computed tomography (CT) images or for developing risk assessment models, such as the Framingham Risk Score for heart disease.[1] Another promising way sML can be used is for proposing diagnoses to patients based on clinical data. Overall, sML systems can standardise, simplify and speed up decision processes, thereby especially supporting physicians with limited experience and patients with rare diseases.

Regardless of the specific research question, availability of large sets of patient data is crucial for sML algorithms to detect the underlying connections between observable input (e.g., CT images) and physician-concluded output (e.g., annotation of a cluster of pixels as a nodule). Due to the successful digitalisation of patient data in hospitals and private practices, more and more such datasets are becoming accessible, comprising parameters like hormone levels, image data, genomic variants and, most importantly, diagnoses physicians made based on this information.

Recently, efforts to apply sML in the field of reproductive medicine and male infertility have increased,[2,3] which is an important advancement because almost 15% of all couples trying to conceive are affected by infertility. In approximately half of these cases, male infertility is the sole or a contributing factor.[4] Several authors have used sML techniques to perform sperm assessments and streamline and standardise the results.[5–7] Others, for example, Santi et al.,[8] Zeadna et al.[9] and Akinsal et al.,[10] have attempted to classify patients according to their semen quality, predict the success of testicular sperm extraction (TESE) and detect chromosomal abnormalities, respectively. As diagnosing and identifying the causes of infertility are clinically challenging because of the heterogeneity of underlying pathologies, such sML-based prediction models could present a valuable tool for improving diagnostic precision and, therefore, patient treatment.

One severe form of male infertility in which no sperm can be identified in the ejaculate is azoospermia. The major contributors to this condition are genetic factors, including gene mutations and chromosomal abnormalities.[11–13] However, genetic causes of male infertility may be underdiagnosed during routine clinical evaluation. Even for prevalent syndromes with typical clinical features, such as Klinefelter Syndrome (KS) (1:400 newborn males), only a minority is diagnosed correctly[14] resulting in direct consequences for the affected patients because of inadequate treatment.[15] In fact, it is currently estimated that only 26% of patients will be diagnosed in life,[16] possibly because of the poor awareness of andrological health or to the lack of any facial distinguishing features.[17] The presence of a highly variable clinical phenotype could also contribute to diagnostic delay: as paucisymptomatic patients are rarely seeking consultation, the diagnosis often occurs during assessment of couple infertility.[18] Such delay in diagnosing KS also results in worse clinical outcomes, such as higher prevalence of metabolic syndrome,[19] poorer cardiovascular[20] and bone health,[21] delays in speech acquisition[16] and declining success rates for sperm retrieval.[22]

KS patients have an exclusive karyotype of 47,XXY and a typical but non-exclusive phenotype of being tall, with long legs and very small, firm testes.[14] Testicular atrophy is a hallmark sign of KS, resulting from degeneration and hyalinisation of tubules: azoospermia is a consequence of this phenomenon, although in some men focal areas of preserved spermatogenesis can be identified. Follicle-stimulating hormone (FSH) and luteinizing hormone (LH) levels reach supraphysiological levels, while testosterone is most often in the subnormal range. Further, relative hyperestrogenism often occurs in KS patients[23]: this is possibly because of increased activity and expression of the aromatase enzyme, which is boosted by elevated serum LH levels, as well as increased peripheral conversion of testosterone because of increased visceral adiposity of KS patients, finally also resulting in gynecomastia. Elevated levels of FSH and LH, small testis and gynecomastia, distinctive characteristics first described by Klinefelter et al.,[24] occur with a prevalence of 10%–12% among azoospermic patients,[14] making KS a relevant condition for a proof-of-concept study of using sML in the field of male infertility.

In order to evaluate the potential of sML prediction models to automatically differentiate azoospermic KS patients with karyotype 47,XXY from azoospermic patients with a normal 46,XY karyotype (non-KS), we trained, validated, tested and benchmarked multiple models based on different sML algorithms. Additionally, to assess whether the models could contribute to increasing the share of diagnosed KS patients, we compared their performance to the manual evaluations performed by 18 physicians from urological practices or specialised clinics.

2 | MATERIAL AND METHODS

2.1 | Ethical approval

The study was carried out in accordance with the protocols approved by the Ethics Committee of the Medical Faculty and the state medical board (Az. 4 I Nie).

2.2 | Study population

Data were retrieved retrospectively from Androbase, the in-house developed database of the Centre of Reproductive Medicine and Andrology (CeRA), University Hospital Münster, Germany.[25] Since its implementation in 2004, CeRA has collected data from its male infertility patients during systematic diagnostic work-ups. The data include information on the history of the patients, anthropometric measurements, laboratory test results, genetic testing and clinical data, such as ultrasonography of the testis.[26] With over 42,000 patients documented (status as of 01.11.2021), Androbase is likely among the world's largest electronic databases for sexual and reproductive medicine. For patients presenting with azoospermia at the CeRA without any obvious reason for this condition, for example, previous cancer treatment or vasectomy, the karyotype is usually assessed. Thus, KS diagnoses in Androbase are not solely based on physicians' assessments but also on the result of an independent test, which is the gold standard in KS diagnosis.

For the sML project, Androbase was queried retrospectively for all adult patients with primary azoospermia whose first visit at the CeRA was between January 2005 and June 2019, and who had no missing data in any of the selected features and no obvious reason for azoospermia. This resulted in a set of 345 KS patients and 994 non-KS patients for developing the prediction models. For all inclusion and exclusion criteria and number of patients for each filter step, see Figure 1. Additionally, data from 32 KS and 105 non-KS patients attending the CeRA during the model development phase (between July 2019 and July 2020) was collected prospectively in order to assess the quality of the final models with completely new data. To compare the model performance with physicians' assessments, we used as a benchmark the manual evaluations for the 137 prospective patients of 18 physicians from both urological practices and specialised clinics. Also, to evaluate the models under conditions of noisy data, we used data from 57 patients with cryptozoospermia and karyotype 46,XY as well as one azoospermic patient with 46,XX, one with a ring chromosome Y and four with translocations.

2.3 | Statistical analysis and machine learning methods and algorithms

The sML approach evaluated in this study focused on developing prediction models that assign specific labels to entities based on a set of observable or measurable input features.[27] For an overview on how sML prediction models are created and evaluated, see the extended methods in Supporting Information. Here, the labels to be assigned are the karyotypes 46,XY and 47,XXY, and the entities to be evaluated are azoospermic patients.

For this task, we selected five characteristic parameters in andrological diagnostics as input features: height (cm), FSH (mIU/ml), LH (mIU/ml), total testosterone (nmol/L) and total testis volume (ml). Six additional input features were chosen that might be relevant for differentiating between patient groups and that were available for most patients in Androbase[25]: (i) age (years), (ii) body mass index (BMI) (kg/m^2), (iii) semen pH, as a surrogate marker of obstructive azoospermia, (iv) estradiol (pmol/L) accounting for relative hyperestrogenism in KS patients,[23] (v) prolactin (mIU/L), which has an inhibitory effect on FSH and LH and possible direct detrimental effects on spermatogenesis and (vi) ejaculate volume (ml).[28]

The computational part of this project was conducted with Python (version 3.8.8). The script is accessible on Github (https://github.com/Klinefelter-Score/DataAnalysis). Additionally, the main steps of the workflow are described in the following. More details are presented in Figures S1 and S2.

Seven different sML algorithms were optimised and benchmarked with regard to their performance in separating KS and non-KS patients using various Python modules. From scikit-learn (1.0.0)[29] we used adaptive boosting (AdaBoost),[30] gaussian process (GP),[31] k-nearest neighbours (kNN), multilayer perceptron (MLP—SKLearn) and support vector machine (SVM).[32] Additionally we used gradient boosting on decision trees (CatBoost)[33] and MLP implemented with tensorflow (MLP—Tensorflow).[34] All estimators were evaluated with a nested crossvalidation on the retrospective azoospermic (AP) dataset (see Figure S2). The inner crossvalidation was used to optimise the estimator's hyperparameters. The outer crossvalidation was used for testing the model. The best performing models were refit on the whole retrospective dataset and then tested on the prospective dataset of non-obstructive azoospermia patients and the prospective cryptozoospermia dataset.

First, the general capability of the features to differentiate KS and non-KS patients was checked by creating a descriptive statistic for the retrospective data. It comprises the computation of (i) median and ranges for each feature, (ii) significance of featurewise difference in distribution of values for KS and non-KS patients based on Mann–Whitney *U*-test and two-sample Kolmogorov–Smirnov test, (iii) effect size of the aforementioned difference in distribution, (iv) pairwise Spearman's correlation coefficient of features and (v) first two components of the uniform manifold approximation and projection (UMAP) embedding, in order to visualise the distances between KS and non-KS patients in the two-dimensional space.

Afterwards, the features that were identified as significant were used to evaluate the predictive capabilities of the sML algorithms. A nested crossvalidation with five stratified folds each was used to first optimise the hyperparameters of the estimator and afterwards determine the metrics of the best performing model. The metrics were averaged over the folds and the variance was included. For a full hyperparameter list for each estimator see Table S1. For a 1:1 distribution of KS and non-KS patients in the training sets, KS patients were oversampled using the Python module imbalanced-learn (version 0.8.0).

In order to assess the fit of the sML algorithms to the problem structure, mean receiver operating characteristics (ROC) curves and area under curve (AUC) were calculated for the models based on their respective validation sets as well as mean sensitivity, specificity and balanced accuracy. AUC is a combined measure of sensitivity and specificity that is (i) independent of case/control ratio and (ii) independent of (arbitrarily chosen) thresholds.[35] Therefore, the best

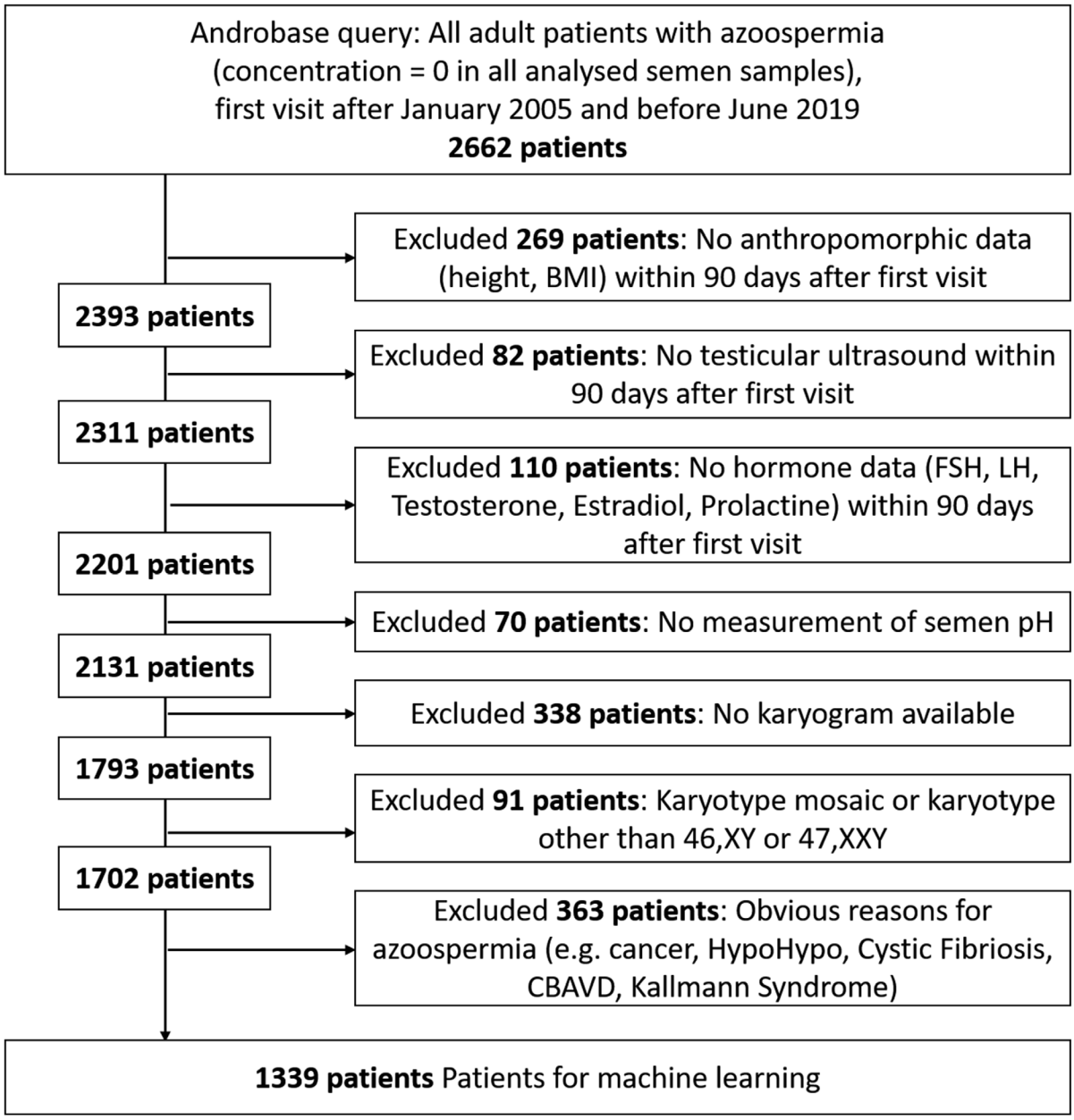

FIGURE 1 Inclusion and exclusion criteria for study population. All adult patients with azoospermia that had their first visit between January 2005 and June 2019 were included into the study population. From these, patients with missing data in either anthropomorphic data, testicular ultrasound, hormone levels, semen pH or karyogram as well as patients with any karyotype other than 46,XY and 47,XXY and patients with obvious reasons for azoospermia were excluded

fit sML algorithm was chosen by comparing mean AUC values, and its best model was identified as the final model also by the best AUC. An appropriate decision threshold for each sML algorithm was set such that, on average, 95% of KS patients were correctly identified in the inner validation sets.

Chosen model and threshold were then evaluated based on retrospective test data of the outer crossvalidation as well as on an external validation set consisting of prospective data of KS and non-KS patients, prospective data of 57 cryptozoospermic patients with karyotype 46,XY and six azoospermic patients with chromosomal abnormalities other than KS. Additionally, the performance on the prospective external validation set was compared with the manual assessments of 18 physicians. For their assessments, the physicians were provided with the exact same information as the models, that is, 10 feature values per patient. The results of the patients' karyogram was not shown to the physicians. Comparison of the models' and physicians' median sensitivity and specificity was performed using the McNemar's test. Statistical significance was set at $p < 0.05$.

For a more comprehensive analysis of the best fit sML algorithm, a feature importance analysis was conducted, revealing the relevance of features for decision-making. The feature importance was quantified by the mean SHAP[36] (SHapley Additive exPlanations) values of each

TABLE 1 Descriptive statistics and statistical analysis of parameters in azoospermic patients with Klinefelter Syndrome (KS) and without KS (non-KS) in the retrospective dataset ($n = 1339$)

Feature	All patients ($n = 1339$)	Non-KS patients ($n = 994$)	KS patients ($n = 345$)	Mann–Whitney *U*-test	Two samples KS test	Cohen's *d*	Point-biserial correlation coefficient
Age[a]	33 (18–62)	34 (18–62)	30 (18–60)	<0.0001	<0.0001	−0.6496	−0.2495
BMI	26.5 (15.2–59.1)	26.6 (15.2–59.1)	25.7 (15.2–53.8)	0.0006	0.0006	−0.2177	−0.0860
Height[a]	182 (141–209)	180 (141–206)	185 (160–209)	<0.0001	<0.0001	0.5230	0.2031
Weight	90.4 (46–216)	90.2 (46–194)	91.3 (46–216)	0.7069	0.0774	0.0627	0.0248
Total testis volume[a]	16 (0.3–95)	21 (0.3–95)	3.8 (1.5–16)	<0.0001	<0.0001	−1.4687	−0.5059
Ejaculate volume[a]	3.1 (0.1–20)	3.3 (0.1–20)	2.3 (0.1–10.5)	<0.0001	<0.0001	−0.5800	−0.2250
pH	7.9 (6–10)	7.9 (6–10)	7.9 (6.5–9.5)	0.0021	0.0807	0.2393	0.0948
LH[a]	7 (0.1–48.4)	5.5 (0.1–33.9)	14.1 (0.1–48.4)	<0.0001	<0.0001	1.4104	0.4884
FSH[a]	18.6 (0.1–115)	15.35 (0.1–115)	28.1 (0.1–95.2)	<0.0001	<0.0001	0.8803	0.3298
Prolactin	167 (22–1640)	165 (22–1580)	171 (56–1640)	0.7886	0.8529	−0.0729	−0.0289
Total testosterone	13.5 (0.5–145.6)	14.4 (0.6–55.3)	11.4 (0.5–145.6)	<0.0001	<0.0001	−0.3043	−0.1199
Estradiol	80 (0.1–355)	79 (0.1–355)	81 (0.1–333)	0.2067	0.3530	0.1092	0.0433

Note: Values are median (range). Mann–Whitney *U*-test and two-sample Kolmogorov–Smirnov test were used to determine significant differences of means. Cohen's *d* and point-biserial correlation coefficient were used to determine the effect size.
Abbreviations: BMI, body mass index; FSH, follicle-stimulating hormone; LH, luteinizing hormone.
[a]Features with an absolute effect size greater than 0.4 and 0.2 for both tests, respectively, were considered for the supervised machine learning (sML) algorithms.

feature after applying a repeated stratified *k*-fold on the best performing model parameters. The absolute SHAP values were averaged and normalised in order to be interpreted as a percentage of information gain.

3 | RESULTS

3.1 | Differences between KS and non-KS patients in the study data

For differentiating between the two groups, sML algorithms detect structural differences between feature values. In the data of this study, six out of 12 features differed significantly between KS and non-KS patients in the retrospective dataset (Mann–Whitney *U*-test *p*-values < 0.001; see Table 1), indicating a well-chosen feature set and, thus, good applicability of sML algorithms. These features are age, height, LH, FSH, total testis volume and ejaculate volume (Figure 2). Medians and ranges of all features are presented in Table 1 and violin plots for both populations in Figure 2 for the retrospective dataset. Table S2 shows the medians and ranges for the prospective datasets. Additionally, the two groups can be differentiated visually in a plot of the first two components of the UMAP embedding[37]: an approximation of distances between retrospective patients in the two-dimensional space based on all feature values is shown in Figure 3. The two groups of patients overlap but are clearly shifted. Approximately 86% of KS patients (8% of non-KS patients) have a first component of at most (at least) two.

Three of the features have mutual correlations in the retrospective data: FSH and LH are positively correlated (Spearman correlation coefficient of 0.78). FSH and total testis volume as well as LH and total testis volume are negatively correlated (Spearman correlation coefficient -0.54 and -0.62, respectively). Also, there is a weak correlation between total testosterone and estradiol (Spearman correlation coefficient of 0.28). The presence of correlated features in the data is relevant for feature importance analyses, because correlated features can partially compensate for each other. Also, information that is present in multiple features increases the overall weight of that information for the whole model.

3.2 | Identification of best fit sML algorithm

In terms of AUC most sML algorithms performed comparably well on the retrospective AP dataset. On the retrospective dataset all estimators achieved an average AUC on all outer folds of at least 0.95. The best performing models even scored >0.97 with a variance of less than 0.01. The corresponding ROC curves of all outer folds combined are depicted in Figure S3.

Since all chosen classifiers worked comparably well on the retrospective dataset, they were refit on the whole dataset and further evaluated on the two separate test sets.

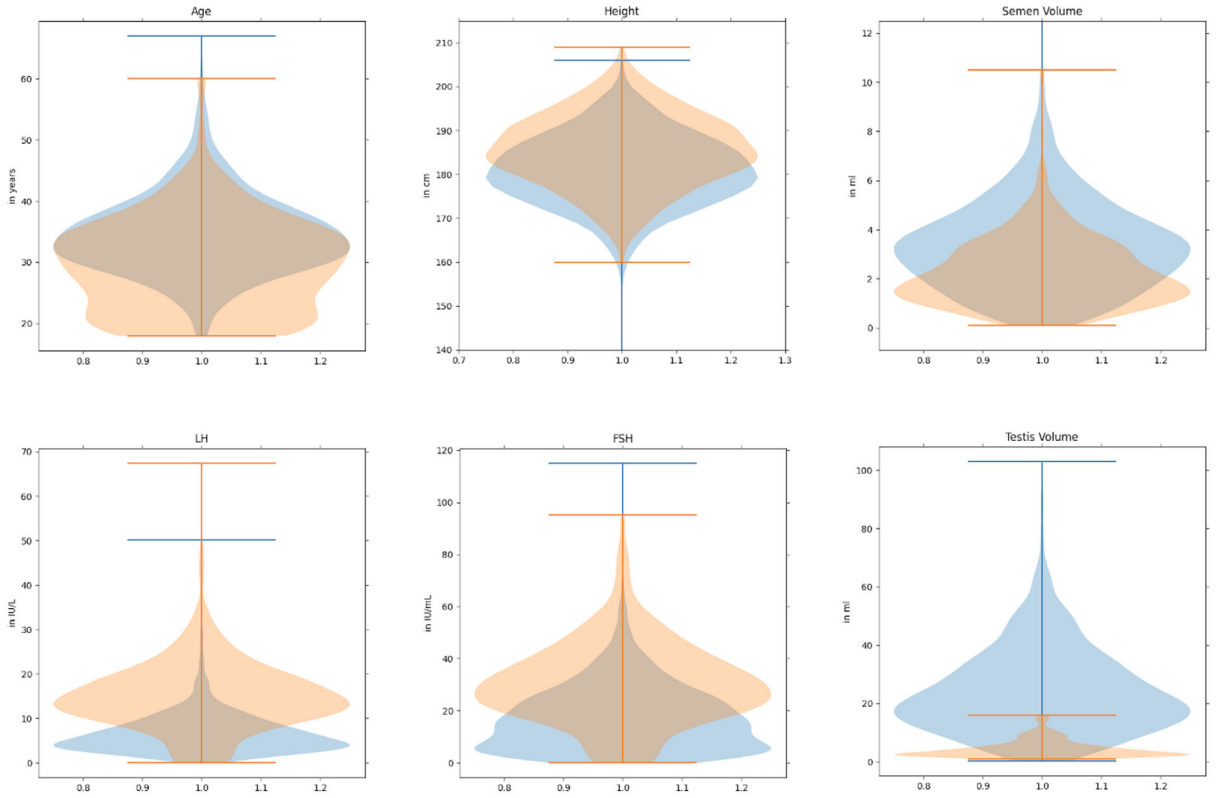

FIGURE 2 Violin plots for the six selected features. The distribution is presented for Klinefelter (orange) and non-Klinefelter (blue) azoospermic patients from the retrospective dataset

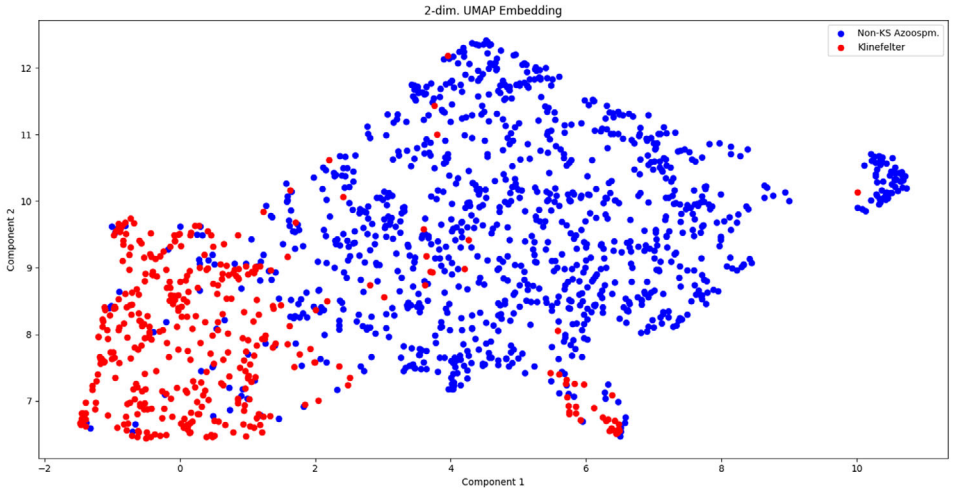

FIGURE 3 Approximated distances between patients from the retrospective data in the two-dimensional space. Dimension reduction based on uniform manifold approximation and projection (UMAP). Klinefelter: azoospermic patients with Klinefelter Syndrome (KS). Non-KS control patients: azoospermic patients with karyotype 46,XY

3.3 | Feature importance analyses

Feature importance analyses were conducted on the three different classifiers. CatBoost, MLP and SVM were chosen because of their high performance and high dissimilarity of the underlying algorithms. The feature importance is estimated using the mean absolute SHAP values for all samples in the current test set. The value is unity based normalised in order to provide a percentage of impact on the model's decision and averaged on all outer folds. For all three resulting models, the total testis volume has the highest importance (49.1%). LH has the second highest importance (18.7%). Although both, LH and FSH, have a significant difference of median and a high effect size, FSH has only an

TABLE 2 Quality measures for final models on retrospective test data of azoospermic patients (AP) and prospective external validation data of AP and of cryptozoospermic patients (CP)

Dataset	Classifier	Threshold	AUC	Sensitivity	Specificity	Balanced accuracy
Retrospective AP	AdaBoost	0.467 (0.017)	0.968 (0.008)	0.963 (0.026)	0.855 (0.042)	0.909 (0.015)
	CatBoost	0.277 (0.011)	0.975 (0.006)	0.931 (0.036)	0.919 (0.028)	0.925 (0.017)
	Gaussian process	0.430 (0.017)	0.975 (0.008)	0.971 (0.015)	0.883 (0.028)	0.927 (0.015)
	k-Nearest neighbours	0.336 (0.032)	0.957 (0.008)	0.952 (0.014)	0.879 (0.027)	0.916 (0.013)
	MLP—SKLearn	0.376 (0.015)	0.977 (0.006)	0.952 (0.027)	0.908 (0.030)	0.930 (0.014)
	MLP—Tensorflow	0.382 (0.017)	0.977 (0.006)	0.958 (0.030)	0.908 (0.033)	0.933 (0.011)
	SVM (RBF)	0.365 (0.014)	0.976 (0.007)	0.958 (0.023)	0.914 (0.03)	0.936 (0.009)
	For the following datasets the classifiers have a fixed classification threshold that was determined on the retrospective AP dataset					
Prospective AP	AdaBoost	0.467	0.989	1.000	0.899	0.950
	CatBoost	0.277	0.994	1.000	0.962	0.981
	Gaussian process	0.430	0.990	1.000	0.881	0.940
	k-Nearest neighbours	0.336	0.984	1.000	0.908	0.954
	MLP—SKLearn	0.376	0.987	1.000	0.899	0.950
	MLP—Tensorflow	0.382	0.988	1.000	0.890	0.945
	SVM (RBF)	0.365	0.996	1.000	0.936	0.968
Prospective CP	AdaBoost	0.467	–	–	0.947	–
	CatBoost	0.277	–	–	0.982	–
	Gaussian process	0.430	–	–	0.965	–
	k-Nearest neighbours	0.336	–	–	0.982	–
	MLP—SKLearn	0.376	–	–	0.982	–
	MLP—Tensorflow	0.382	–	–	0.965	–
	SVM (RBF)	0.365	–	–	1.000	–

Note: Values rounded to third decimal place. Thresholds are calculated on the validation set of the inner fold. Performances are calculated on the test set of the outer fold. Variances between the outer folds are shown in brackets.
Abbreviations: AUC, area under curve; MLP, multilayer perceptron; RBF, radial basis function; SVM, support vector machine.

importance of 5.3%. For more details on the feature importance, see Table S3 and S4. As FSH and LH were closely correlated, as expected from physiopathology, we considered exclusion of FSH from the models for the sake of simplicity; however, as the models were not affected by this, and FSH and LH are routinely measured in cases of male infertility, we decided to leave both FSH and LH values included in the final models.

3.4 | Evaluation of final models

The performance of the final models on the retrospective and prospective data is shown in Table 2. All models performed better on the prospective AP data than on the retrospective data. For the retrospective test set, a sensitivity of >93% (range: 93.1%–97.1%), specificity of >85% (range: 85.5%–91.9%) and balanced accuracy of 90.9%–93.6% were reached. In case of the prospective AP data, all models classified all KS patients correctly (sensitivity = 100%), while the non-KS patients were classified with a specificity of at least 88.1% (range: 88.1%–96.2%). The balanced accuracy was between 94% and 98.1%. For the prospective data of cryptozoospermic patients, sensitivity could not be calculated because there were no KS patients in the set. However, specificity was >94% (range: 94.7%–100%) for all models; the balanced accuracy equals the specificity because the dataset consists of only non-KS patients. For predicting the patients that had neither a 47,XXY nor a 46,XY karyotype, all models behaved the same: patients with ring Y or translocation on the Y chromosome were classified as non-KS, and the XX patient was classified as KS.

Overall, the seven models all performed well on the different datasets. The CatBoost and the SVM models performed consistently better on all three datasets with highest AUC or highest balanced accuracy. Therefore, we made these two models accessible as a *KS Score Calculator* via a web application: https://klinefelter-score-calculator.uni-muenster.de. For research purposes, researchers and clinicians can enter values of the six features for their patients into a web form and calculate their KS scores.

3.5 | Comparing final models and manual evaluations by physicians

In addition to comparing each of the model performances on the prospective test data with each other, their performances were also compared with manual evaluations made by 18 physicians

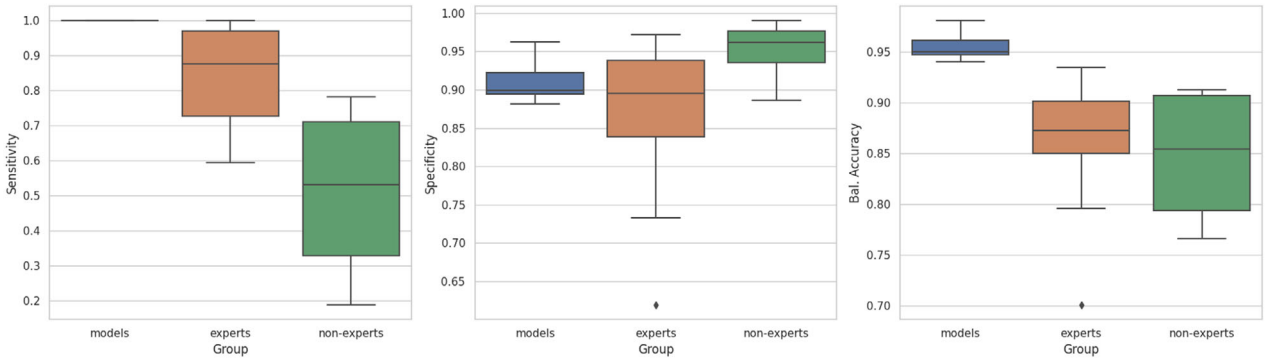

FIGURE 4 Comparison of prediction model performances and manual evaluation by physicians. Boxplots for sensitivity, specificity and classification errors of best fit prediction models and physicians (experts, $n = 14$ and non-experts, $n = 4$)

(andrologists or urologists). Some of these physicians regularly diagnose and treat KS patients, while others only rarely have contact with this condition. Accordingly, their answers were grouped as "KS-experts" ($n = 14$) and "non-KS-experts" ($n = 4$).

In this dataset, the models recognised all KS patients correctly, but on median they also assigned the KS label to ~10.1% of the non-KS patients. By comparison, the KS-expert group recognised fewer KS patients correctly (median 87.5%) and equally many non-KS patients (median ~89.52%), suggesting that the models were significantly more reliable in terms of median sensitivity ($p = 0.0455$) while performing comparably in regards to specificity ($p = 0.4795$). The non-KS-expert group labelled almost all non-KS patients correctly (median ~96.19%) but missed approximately half of the KS patients (median ~53.13), with a significantly higher classification error. In addition, physicians' results had a wider variability than the models (see Figure 4). Figure S4 also shows the ROC curve of the three best models and the physicians' predictions as a voting ensemble divided into KS-experts and non-KS-experts.

Further analysis of the patients that were misclassified by any of the models revealed that all final models mostly failed for the same individuals. Out of 156 patients from the retrospective data that were misclassified by any of the models, 136 were misclassified by two different models, and 97 were misclassified by at least three different models. Similarly, five of 11 misclassified patients from the prospective AP data were misclassified by at least three models. All five patients from the prospective data were recognised correctly by at least two of the physicians and, on median, by 12 physicians. Misclassified non-KS patients (false positives) had a significantly lower testis volume, higher FSH level and higher LH level than other non-KS patients (Mann–Whitney *U*-test *p*-values < 0.00005), while misclassified KS patients (false negatives) had a slightly higher age, lower FSH level and lower LH level than other KS patients (Mann–Whitney *U*-test *p*-values < 0.05) (see Table S5).

4 | DISCUSSION

The presented results show that differentiating KS patients from azoospermic patients with normal karyotype (46,XY) is a problem that can be addressed by sML techniques. In our study three sML algorithms performed equally well on all three datasets, and their decisions were mainly based on the parameters LH and total testis volume. This is unsurprising, because the nearly complete absence of germ cells and hypergonadotropic hypogonadism are hallmark clinical features of KS. Though both values are correlated with FSH, the feature importance of FSH drops to 5% on average. Indeed, elevated FSH levels are commonly found in all forms of non-obstructive azoospermia, whereas LH levels are usually different between KS and non-KS patients, as also occurring in our study population. While in clinical practice it would make little sense to measure only one of the two gonadotropins given a suspicion of KS, the machine learning (ML) models suggest that in fact LH and testicular volume would be more "important" than FSH to predict KS in an azoospermic patient.

4.1 | Clinical relevance of the findings

Despite its prevalence, KS is vastly underdiagnosed with the consequence that only about one out of four KS men seem to be detected throughout their lifetime.[38] Albeit the condition is frequent, respective experience in diagnosis and treatment is clustered in expert centres and could be improved elsewhere. Thus, an increment of general knowledge as well as establishment of standard diagnostic tools in multidisciplinary networks is mandatory. In agreement with a current guideline on KS,[15] physicians should be given a tool to facilitate detection of KS patients. A higher detection rate of KS is likely to promote the patients' self-esteem, assure quality of life and improve social adaption by early access to professional care. Finally, preservation of the fertility potential will be optimised and early detection of the onset of hypogonadism will lead to improved treatment options. Non-invasive predictors for TESE outcome for KS patients are lacking,[39] and given the lower chances (approximately 40%–50%) for sperm retrieval among KS patients,[40] automated tools to promote earlier access to treatment would improve the patients' chances of fatherhood. Thus, prevention of the medical complications/comorbidities associated with KS should be standardised as far as possible and early diagnosis is important because it is associated with better outcomes in terms of fertility as well as for

quality of life, and improved diagnosis rates may in turn influence lifetime morbidity and mortality rates.[15,22] The ML tool can contribute to such a standardisation.

Even if experienced physicians with specialised training in andrology or urology can likely identify most patients with KS based on clinical expertise, less-experienced physicians may have more difficulties in recognising the signs and symptoms of KS. This hypothesis is supported by the results of our physicians' assessments of azoospermic patients. In real-world settings, fewer KS patients would likely have been identified because physicians would not have been biased towards suspecting it in each patient. Yet, formulating a diagnosis based only on clinical values is, of course, much more difficult than doing so in a consultation. Interestingly, although physicians had a much higher variance in their answers than the models and did not reach an equal sensitivity level, they correctly recognised patients for which the models failed, indicating that sML algorithms can complement but not replace physicians for this type of decision.

4.2 | Strengths and limitations

The present study is among the first to use sML in the context of male reproductive health. Its main strengths are the good quality of the data and comparably large amount of data. Further, essential for this study is the fact that a KS diagnosis is available as the result of an independent test (karyogram), as this enables the models to learn the real relation between the features and KS rather than learning to replicate the (biased) assessments of physicians. The models were tested with three different sets of test data, indicating good generalizability. However, training and validation of the models were restricted to azoospermic patients, possibly limiting their power for, for example, oligozoospermic patients. The main limitation of the study is that all the data originated from the same clinic; thus, truly independent test data has not been evaluated. To address this issue, the CatBoost, MLP and SVM models are accessible through a public webpage (https://klinefelter-score-calculator.uni-muenster.de), such that researchers can enter feature values and karyotypes of patients and check whether the labels are correctly predicted. This tool is currently intended for research use only. Additionally, while the age range of patients included in the different sets was broad, including men from 18 to 62 years of age (Tables 1 and S2), the ML models used in the present study were not devised for patients below 18 years of age, and, therefore, we cannot draw any conclusions regarding reliability of the tested models in a paediatric population.

5 | CONCLUSION

This first proof-of-concept study on azoospermic patients indicates that supervised machine learning methods can be used to increase the diagnostic rate of Klinefelter Syndrome among azoospermic patients. If used as part of an automated tool in an electronic medical record or domain-specific database, the supervised machine learning methods will likely lead to earlier diagnoses, which, in consequence, should improve overall patient care and possibly even result in better chances of sperm retrieval by testicular sperm extraction and, thus, potentially fatherhood.[22] This highlights the importance of integrating novel technologies such as machine learning into the field of reproductive health as a way to further improve patient care.

ACKNOWLEDGEMENTS

We wish to thank all clinicians and technicians working at the CeRA for their daily updating of the Androbase with data from new patients. We also wish to thank Dr. Biagio Cangiano, Dr. Settimio d'Andrea, Dr. Riccardo Giovannone, Dr. Francesco Pallotti, Prof. Francesco Romanelli, Dr. Chiara Maria Trovato, Dr. Walter Vena, Jonas Maliske, Simone Bier, Yousif Rassam and all other physicians who reviewed the prospective test data.

FUNDING INFORMATION

This work was carried out within the frame of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) Clinical Research Unit 'Male Germ Cells: from Genes to Function' (CRU326).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Henrike Krenz and Michael Fujarski developed and analysed the sML models. Andrea Sansone contributed to the data analysis. Andrea Sansone and Frank Tüttelmann were responsible for the collection of data. Michael Fujarski and Julian Varghese optimised the sML models and conducted feature importance analysis. Henrike Krenz, Andrea Sansone and Michael Fujarski drafted the manuscript. Claudia Krallmann, Michael Zitzmann, Martin Dugas, J. Varghese, Sabine Kliesch, Frank Tüttelmann and Jörg Gromoll participated in the interpretation of the study data. Frank Tüttelmann and Jörg Gromoll designed and supervised the whole study. All authors critically revised the manuscript and approved the final version.

DATA AVAILABILITY STATEMENT

The raw data underlying this article cannot be shared publicly because of the privacy of individuals that participated in the study. The data will be shared on reasonable request to the corresponding author.

ORCID

Andrea Sansone https://orcid.org/0000-0002-1210-2843
Michael Zitzmann https://orcid.org/0000-0003-3629-7160
Frank Tüttelmann https://orcid.org/0000-0003-2745-9965

REFERENCES

1. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

2. Curchoe CL, Bormann CL. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. *J Assist Reprod Genet*. 2019;36(4):591-600. https://doi.org/10.1007/s10815-019-01408-x
3. Chu KY, Nassau DE, Arora H, Lokeshwar SD, Madhusoodanan V, Ramasamy R. Artificial intelligence in reproductive urology. *Curr Urol Rep*. 2019;20(9):52. https://doi.org/10.1007/s11934-019-0914-4
4. Isidori AM, Sansone A, Gianfrilli D. Hormonal treatment of male infertility: gonadotropins and beyond. In: Simoni M, Huhtaniemi I, eds. *Endocrinology of the Testis and Male Reproduction*. Chapter 36. Springer; 2017:1071-1090.
5. Sahoo AJ, Kumar Y. Seminal quality prediction using data mining methods. *Technol Health Care*. 2014;22(4):531-545. https://doi.org/10.3233/THC-140816
6. Chang V, Garcia A, Hitschfeld N, Hartel S. Gold-standard for computer-assisted morphological sperm analysis. *Comput Biol Med*. 2017;83:143-150. https://doi.org/10.1016/j.compbiomed.2017.03.004
7. Riordon J, McCallum C, Sinton D. Deep learning for the classification of human sperm. *Comput Biol Med*. 2019;111:103342. https://doi.org/10.1016/j.compbiomed.2019.103342
8. Santi D, Spaggiari G, Casonati A, et al. Multilevel approach to male fertility by machine learning highlights a hidden link between haematological and spermatogenetic cells. *Andrology*. 2020;8(5):1021-1029. https://doi.org/10.1111/andr.12826
9. Zeadna A, Khateeb N, Rokach L, et al. Prediction of sperm extraction in non-obstructive azoospermia patients: a machine-learning perspective. *Hum Reprod*. 2020;35(7):1505-1514. https://doi.org/10.1093/humrep/deaa109
10. Akinsal EC, Haznedar B, Baydilli N, Kalinli A, Ozturk A, Ekmekcioglu O. Artificial neural network for the prediction of chromosomal abnormalities in azoospermic males. *Urol J*. 2018;15(3):122-125. https://doi.org/10.22037/uj.v0i0.4029
11. Lee JY, Dada R, Sabanegh E, Carpi A, Agarwal A. Role of genetics in azoospermia. *Urology*. 2011;77(3):598-601. https://doi.org/10.1016/j.urology.2010.10.001
12. Tournaye H, Krausz C, Oates RD. Concepts in diagnosis and therapy for male reproductive impairment. *Lancet Diab Endocrinol*. 2017;5(7):554-564.
13. Tuttelmann F, Ruckert C, Ropke A. Disorders of spermatogenesis: perspectives for novel genetic diagnostics after 20 years of unchanged routine. *Med Genet*. 2018;30(1):12-20. https://doi.org/10.1007/s11825-018-0181-7
14. Kanakis GA, Nieschlag E. Klinefelter syndrome: more than hypogonadism. *Metabolism*. 2018;86:135-144. https://doi.org/10.1016/j.metabol.2017.09.017
15. Zitzmann M, Aksglaede L, Corona G, et al. European academy of andrology guidelines on Klinefelter Syndrome Endorsing Organization: European Society of Endocrinology. *Andrology*. 2021;9(1):145-167. https://doi.org/10.1111/andr.12909
16. Zitzmann M, Rohayem J. Gonadal dysfunction and beyond: clinical challenges in children, adolescents, and adults with 47,XXY Klinefelter syndrome. *Am J Med Genet C Semin Med Genet*. 2020;184(2):302-312. https://doi.org/10.1002/ajmg.c.31786
17. Bonomi M, Rochira V, Pasquali D, et al. Klinefelter syndrome (KS): genetics, clinical phenotype and hypogonadism. *J Endocrinol Invest*. 2017;40(2):123-134. https://doi.org/10.1007/s40618-016-0541-6
18. Lanfranco F, Kamischke A, Zitzmann M, Nieschlag E. Klinefelter's syndrome. *Lancet*. 2004;364(9430):273-283. https://doi.org/10.1016/S0140-6736(04)16678-6
19. Bojesen A, Kristensen K, Birkebaek NH, et al. The metabolic syndrome is frequent in Klinefelter's syndrome and is associated with abdominal obesity and hypogonadism. *Diab Care*. 2006;29(7):1591-1598. https://doi.org/10.2337/dc06-0145
20. Calogero AE, Giagulli VA, Mongioi LM, et al. Klinefelter syndrome: cardiovascular abnormalities and metabolic disorders. *J Endocrinol Invest*. 2017;40(7):705-712. https://doi.org/10.1007/s40618-017-0619-9
21. Herlihy AS, Halliday JL, Cock ML, McLachlan RI. The prevalence and diagnosis rates of Klinefelter syndrome: an Australian comparison. *Med J Aust*. 2011;194(1):24-28. https://doi.org/10.5694/j.1326-5377.2011.tb04141.x
22. Rohayem J, Fricke R, Czeloth K, et al. Age and markers of Leydig cell function, but not of Sertoli cell function predict the success of sperm retrieval in adolescents and adults with Klinefelter's syndrome. *Andrology*. 2015;3(5):868-875. https://doi.org/10.1111/andr.12067
23. Santi D, De Vincentis S, Scaltriti S, Rochira V. Relative hyperestrogenism in Klinefelter Syndrome: results from a meta-analysis. *Endocrine*. 2019;64(2):209-219. https://doi.org/10.1007/s12020-019-01850-y
24. Klinefelter HF, Reifenstein EC, Albright F. Syndrome characterized by gynecomastia, aspermatogenesis without a-leydigism, and increased excretion of follicle-stimulating hormone1. *J Clin Endocrinol Metab*. 1942;2(11):615-627. https://doi.org/10.1210/jcem-2-11-615
25. Tuttelmann F, Luetjens CM, Nieschlag E. Optimising workflow in andrology: a new electronic patient record and database. *Asian J Androl*. 2006;8(2):235-241. https://doi.org/10.1111/j.1745-7262.2006.00131.x
26. Kliesch S. Diagnosis of male infertility: diagnostic work-up of the infertile man. *Eur Urol Suppl*. 2014;13(4):73-82. https://doi.org/10.1016/j.eursup.2014.08.002
27. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. New York, NY: Springer; 2013.
28. Hair WM, Gubbay O, Jabbour HN, Lincoln GA. Prolactin receptor expression in human testis and accessory tissues: localization and function. *Mol Hum Reprod*. 2002;8(7):606-611. https://doi.org/10.1093/molehr/8.7.606
29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
30. Schapire RE. Explaining AdaBoost. In: *Empirical Inference*. Chapter 5. Springer; 2013:37-52.
31. Rasmussen CE. Gaussian processes in machine learning. In: *Advanced Lectures on Machine Learning*. Chapter 4. Springer; 2004:63-71.
32. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. https://doi.org/10.1007/bf00994018
33. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV & Gulin A. CatBoost: unbiased boosting with categorical features. In: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. ACM; 2018:6639-6649. arXiv preprint arXiv:170609516.
34. Abadi M, Barham P, Chen J, et al. Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA. USENIX Association; 2016:265-283.
35. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*. 2013;4(2):627-635.
36. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM; 2017:4768-4777.
37. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.
38. Bojesen A, Juul S, Gravholt CH. Prenatal and postnatal prevalence of Klinefelter syndrome: a national registry study. *J Clin Endocrinol Metab*. 2003;88(2):622-626. https://doi.org/10.1210/jc.2002-021491
39. Busch AS, Tuttelmann F, Cremers JF, et al. FSHB -211 G>T polymorphism as predictor for TESE success in patients with unexplained azoospermia. *J Clin Endocrinol Metab*. 2019;104(6):2315-2324. https://doi.org/10.1210/jc.2018-02249

40. Corona G, Minhas S, Giwercman A, et al. Sperm recovery and ICSI outcomes in men with non-obstructive azoospermia: a systematic review and meta-analysis. *Hum Reprod Update*. 2019;25(6):733-757. https://doi.org/10.1093/humupd/dmz028

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Krenz H, Sansone A, Fujarski M, et al. Machine learning based prediction models in male reproductive health: Development of a proof-of-concept model for Klinefelter Syndrome in azoospermic patients. *Andrology*. 2022;10:534–544. https://doi.org/10.1111/andr.13141