Novel Fatigue Profiling Approach Highlights Temporal Dynamics of Human Sperm Motility

Athanasia Sergounioti[1], Efstathios Alonaris[2], Dimitrios Rigas[1]

[1]Department of Laboratory Medicine, General Hospital of Amfissa, Amfissa, Greece

[2]MSc HosMic (Host-Microbe Interactions), University of Thessaly, Larissa, Greece

Abstract

Background:

Accurate characterization of human sperm motility is crucial for understanding male fertility potential. Traditional motility assessment methods primarily focus on static velocity parameters, often overlooking temporal declines in motility during the sperm trajectory.

Objective:

We aimed to develop and validate a novel fatigue-based profiling approach to assess intra-trajectory motility decline in human spermatozoa.

Methods:

Using computer-assisted sperm analysis (CASA)-derived motion tracking data from 1,118 sperm trajectories, we introduced the Fatigue Index, a log-fold metric quantifying the decline in forward progression (VSL) over time. We performed descriptive, clustering-based, and predictive modeling analyses, including Random Forest classifiers, to characterize fatigue patterns.

Results:

Fatigued spermatozoa exhibited significantly lower straight-line velocity (VSL: 18.4 vs 42.7 μm/s) and steeper VSL slopes (−0.34 vs −0.08 μm/s/frame) compared to non-fatigued counterparts. The Fatigue Index reliably identified subpopulations of sperm with motility deterioration. A Random Forest model achieved high discriminative performance (AUROC = 0.956), and feature importance analysis highlighted VSL slope as the dominant predictor.

Conclusions:

Fatigue-based temporal profiling offers a new dimension for understanding sperm motility, highlighting the dynamic nature of forward progression and identifying subtle impairments that may be overlooked by conventional assessment methods. This approach shows promise as a quantitative tool for dynamic sperm quality evaluation.

1. Introduction

Assessment of male fertility frequently includes evaluation of sperm motility, including total and progressive motility percentages (1). However, traditional assessments are typically static, relying on snapshot-based motility evaluations that do not account for dynamic changes during sperm progression. This limitation neglects potential signs of *intra-trajectory fatigue*, a phenomenon that may reflect metabolic exhaustion or underlying functional impairment even in otherwise motile spermatozoa (2).

Sperm movement is governed by a combination of flagellar activity, ATP-dependent metabolic pathways, and environmental conditions such as pH and temperature (1). Recent studies underscore that sperm not only differ in their baseline motility characteristics but also exhibit varying levels of resilience under sustained metabolic stress, potentially manifesting as time-dependent decline in motility parameters (3).

Given the energetic demands required for sustained forward progression—particularly under ex vivo conditions (1-3) we hypothesize that tracking temporal changes in motility, specifically declines in straight-line velocity (VSL), can reveal subpopulations of sperm undergoing progressive deterioration. While VSL is commonly used as a static indicator of progressive motility, its temporal dynamics remain under-characterized. Time-resolved metrics may reveal subtle impairments that escape static evaluation, particularly in borderline or unexplained infertility (1,4-5).

Although the concept of sperm motility decline over time is recognized, the explicit use of the term "fatigue" to describe such degradation in sperm function remains largely absent from the literature. This study proposes a cautious yet biologically grounded application of the term sperm fatigue to describe within-trajectory motility deterioration, measured as a decline in VSL during progression. Reductions in motility due to energetic depletion, mitochondrial dysfunction, or oxidative stress may mirror mechanisms typically associated with fatigue in other cellular systems (5-6)

Recent studies have emphasized the role of intracellular pH, Ca^{2+} flux, bicarbonate regulation, and mitochondrial dynamics in sustaining sperm motility (7,8). Additionally, external stressors such as psychological or oxidative stress can impair flagellar function, linking environmental or systemic factors to energy-dependent motility decline (8). These insights form a plausible rationale for exploring fatigue-like behavior in spermatozoa. We propose to interpret sustained reductions in VSL within individual trajectories as evidence of energy-linked motility decay, conceptually framed here as sperm fatigue. This novel and exploratory framework aligns with known metabolic constraints and mitochondrial vulnerabilities of sperm cells (9).

To our knowledge, no prior metric has explicitly captured intra-trajectory motility decay using log-transformed velocity ratios in spermatozoa. We introduce a new log-fold VSL index, quantifying relative motility deterioration over time. This approach enables characterization of fatigue at the single-trajectory level and supports the identification of sperm exhibiting putative subclinical impairments.

Using data from a large number of frame-level observations in the VISEM-Tracking dataset, we applied a fatigue detection index and validated it through multiple

complementary strategies: threshold-based classification, z-score normalization, and unsupervised clustering. These analyses offer preliminary evidence for integrating dynamic motility assessment into sperm quality pipelines. Moreover, our approach aligns with ongoing efforts in high-dimensional phenotyping that aim to capture complex, time-dependent motility behaviors in sperm quality research (10-12). To further explore the detectability of fatigue patterns, predictive modeling approaches—including logistic regression and random forest classifiers—were implemented. These models served as analytical tools to examine whether motility-derived features could reliably distinguish fatigued trajectories, rather than aiming for clinical deployment. Such approaches may ultimately inform the development of predictive biomarkers for sperm functionality and male fertility assessment.

2. Materials and Methods

2.1 Dataset and Preprocessing

The present study utilized the open-access VISEM Tracking dataset (13), comprising approximately 656,000 frame-level observations of spermatozoa motion characteristics extracted from semen video recordings. Each observation corresponds to a single frame within an identified spermatozoon trajectory. Each trajectory is uniquely identified by a feature tracking ID (**ftid**), allowing aggregation of frame-level motion data into trajectory-level summaries. Initially, trajectories with fewer than 15 frames were excluded to ensure robust calculation of motion features, resulting in approximately 655,895 frame-level observations distributed across 1,118 distinct spermatozoon trajectories.

The primary features include parameters commonly used in computer-aided sperm analysis (CASA), defined according to the World Health Organization guidelines (14):

• **VSL** (Straight-Line Velocity): The net linear velocity of the spermatozoon from its starting point to endpoint, divided by time.

• **LIN** (Linearity): The ratio of VSL to VCL (Curvilinear Velocity), representing how straight the trajectory is.

• **ALH** (Amplitude of Lateral Head Displacement): The magnitude of lateral oscillations of the sperm head during motion.

• **WOB** (Wobble): The ratio of average path velocity (VAP) to VCL, indicating how consistently a sperm cell follows its average path versus displaying erratic motion (15).

Trajectories were aggregated to produce one row per trajectory, with mean or slope values as appropriate. To quantify intra-trajectory motility decline, we introduce a novel fatigue index based on the log-transformed ratio of terminal to initial straight-line velocity (VSL). This approach captures relative declines in motility while minimizing scale sensitivity and skewness. The logarithmic formulation was selected to symmetrize the distribution and emphasize multiplicative decay in velocity across time (16). This index is designed to reflect energy-dependent reductions in forward progression over the course of the trajectory and was developed as a novel approach inspired by dynamic motility profiling frameworks previously applied in CASA-based sperm assessments (17).

Trajectory-level features were engineered by grouping data according to ftid. For each trajectory, we extracted the starting and ending VSL values (VSL_start and VSL_end), total frame count, and calculated the fatigue index as the logarithmic fold change in VSL:

Fatigue Index = log(VSL_end / VSL_start)

The log-transformation of the ratio between final and initial VSL was employed to symmetrize the distribution of fatigue values, to accommodate the multiplicative nature of motility decline, and to mitigate the influence of extreme outliers. This approach allowed a more balanced and interpretable characterization of motility decay across sperm trajectories. To ensure robustness, we excluded trajectories shorter than

15 frames, recognizing that extremely short tracks may not yield reliable motility estimates (18).

2.2 Fatigue Classification Approaches

To categorize sperm trajectories into fatigued and non-fatigued, we implemented four complementary classification strategies. Each was selected to represent a different analytical perspective—empirical, statistical, data-driven, and unsupervised—allowing for triangulation of fatigue detection and cross-validation of results.

1. **Fixed Threshold Method**: A manually defined cutoff of log_fold < −1 was applied, denoting a greater than 63% reduction in straight-line velocity (VSL) across the trajectory. This threshold approximately corresponds to the left-tail inflection point of the log_fold distribution in our dataset. Deterministic cutoffs of this kind are commonly used in biological classification problems to identify pathological or extreme phenotypes and have been specifically applied in motility studies to delineate functional sperm subpopulations (19-21). The choice of log_fold < −1 was informed by preliminary exploratory analysis, where an inflection point indicating substantial motility loss was visually and statistically detected in the distribution's left tail. Although heuristic, this threshold reflects a biologically meaningful degree of decay consistent with severe fatigue trajectories.
2. **Z-Score Normalization**: We applied a statistical threshold using standardized z-scores of the log_fold distribution, flagging as fatigued those trajectories with z-score(log_fold) < −1.5. This method accounts for variability within the

dataset and identifies fatigue in trajectories exhibiting significant deviation from the mean. Z-score normalization is widely adopted in biomedical research to detect statistical outliers and to segment populations into functionally distinct subgroups. It has also been applied in bioinformatics and sperm-related analyses, including protein marker detection and sperm DNA fragmentation thresholding (22-23).

3. **Unsupervised Clustering**: To enable data-driven categorization without assuming prior thresholds, we initially applied KMeans clustering (k=3) on the log_fold distribution. The cluster exhibiting the lowest mean log_fold value was designated as fatigued. The choice of k=3 was heuristically motivated by preliminary visualizations, biological plausibility, and silhouette analysis metrics. Unsupervised clustering strategies, particularly KMeans, are widely used in biomedical studies to reveal intrinsic subgroup structures without prior labeling (24-26). Additional cluster counts (k=2, 4, 5) were also explored but did not consistently yield more robust subgroup structures. Cluster quality was assessed by calculating the silhouette coefficient and the Davies–Bouldin Index on the aggregated trajectory-level motion features (mean VSL, LIN, ALH). These metrics provide quantitative evaluation of clustering performance, with higher silhouette scores and lower Davies–Bouldin indices indicating better-defined cluster structures.
4. **Data-Driven Percentile Thresholding**: To systematically define fatigued trajectories for predictive modeling, an additional data-driven thresholding approach was employed. Specifically, a threshold at the 25th percentile of the log_fold change distribution was applied, classifying trajectories with values below this percentile as fatigued. Percentile-based thresholding strategies have

been widely adopted in biomedical research to identify extreme phenotypes and functional subpopulations, particularly in the context of sperm parameter assessment and omics data classification (27, 28).

These four complementary strategies enable triangulation of fatigue detection from empirical, statistical, data-driven, and unsupervised perspectives.

2.3 Feature Analysis

Univariate analyses were performed to assess the relationship between individual motion features and fatigue status. Specifically:

• **Point-**biserial correlations were calculated to quantify the association between each continuous feature (e.g., VSL, LIN, ALH, VSL slope) and the binary fatigue flag (0 = non-fatigued, 1 = fatigued). The point-biserial correlation coefficient is a specialized form of Pearson's correlation suitable when one variable is continuous and the other is dichotomous, offering a measure of both direction and strength of association (29-30).

• **Boxplots** were generated to visualize the distributional differences of each feature across fatigue status groups.

Together, these univariate analyses provide complementary insights into which motion features are most strongly associated with fatigue and how they behave across the trajectory-level motility spectrum.

2.4 Predictive Modeling

Two supervised machine learning algorithms were implemented:

- **Logistic Regression**: A generalized linear model with L2 regularization, maximum iterations set to 1000.
- **Random Forest Classifier**: An ensemble of 100 decision trees, maximum depth of 5, maximum 15 leaf nodes, and minimum 5 samples per leaf.

To evaluate model performance while preserving independence between training and test sets, a GroupKFold cross-validation approach was used. In this strategy, all trajectories originating from the same individual sample were assigned to the same fold, ensuring that no information from a given sample leaked across training and testing subsets. This was critical to avoid data leakage and to obtain an unbiased estimate of model generalization performance. Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC).

2.5 Model Diagnostics and Validation

Given the initially high predictive performance (AUROC > 0.95), extensive validation steps were conducted to ensure that the models were not overfitting or exploiting data leakage artifacts. To robustly assess model reliability and exclude the possibility of overfitting to noise, the following procedures were applied:

- **Permuted Labels Test**: Model performance was reassessed after random permutation of fatigue labels, serving as a sanity check to evaluate whether the

classifier could differentiate between truly meaningful patterns and random noise (30)

- **Null Features Test**: Random features were generated and used as model inputs to assess whether any spurious associations would inflate performance (31-32).
- **Bootstrap Confidence Intervals**: Bootstrap resampling (n=1000) was applied to estimate empirical 95% confidence intervals for the AUROC scores, providing a non-parametric measure of performance robustness (33-34).
- **SHAP Feature Importance**: SHAP (SHapley Additive exPlanations) analysis was performed to independently verify feature importance estimates using TreeSHAP algorithms applied to the Random Forest model (35-37).
- **Bayesian AUROC Estimation**: Bayesian modeling was utilized to estimate the posterior distribution of AUROC scores via Markov Chain Monte Carlo (MCMC) sampling using PyMC, allowing probabilistic inference beyond conventional null hypothesis significance testing (38).

In parallel, sensitivity analyses were conducted to examine the consistency of our findings across alternative fatigue labeling strategies and under randomized input conditions. The convergence of results across these complementary methods provides strong evidence that our models captured biologically meaningful patterns with generalizable predictive value.

3. Results

3.1 Overview of Fatigue Labeling

Fatigue status was assigned using four complementary strategies: fixed threshold, Z-score threshold, percentile threshold, and clustering-based classification. Agreement between methods was evaluated through cross-tabulation and visualized using an agreement heatmap (Figure 1). Clustering yielded comparable groupings to the percentile-based method, with more than 80% overlap observed between fixed threshold and clustering classifications. Trajectory-level clustering offers an independent confirmation of intra-trajectory motility decay patterns. By aggregating frame-level features and focusing on trajectory means, this approach reduces frame-to-frame noise and captures the overall dynamic behavior of each spermatozoon. The clear separation observed between fatigued and non-fatigued trajectories in the 3D clustering plot (Figure 1B) suggests that fatigue is not merely a transient frame-level artifact but reflects a consistent and biologically meaningful deterioration across the entire progression path.

This finding strengthens the conceptual framework proposing sperm fatigue as an intrinsic motility phenotype, rather than a sporadic fluctuation. Furthermore, the consistency between frame-level fatigue detection methods (fixed thresholding, z-scoring, unsupervised clustering) and trajectory-level clustering reinforces the robustness of the fatigue characterization. Together, these complementary analyses suggest that sperm fatigue can be reliably captured across multiple analytical levels, supporting its potential utility in future sperm quality assessments and research pipelines.

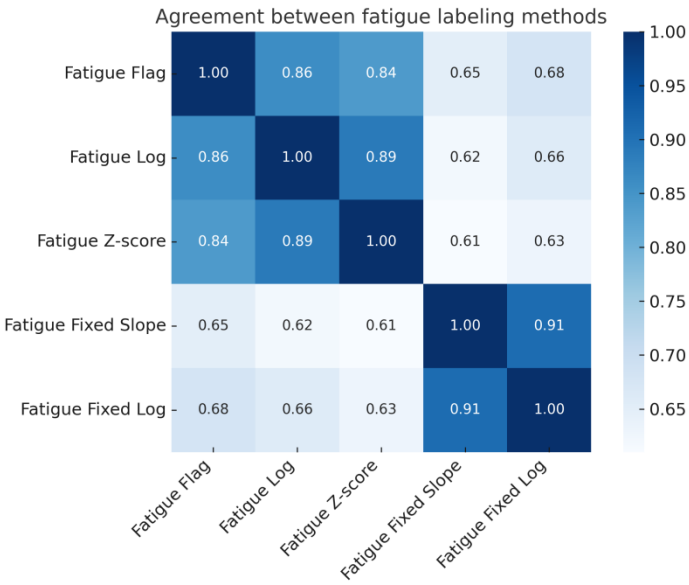

Figure 1A: Heatmap illustrating the agreement (Cohen's κ) between four fatigue labeling methods. VSL log-fold decline and VSL slope showed the highest concordance (κ = 0.91), supporting consistency in fatigue definition.

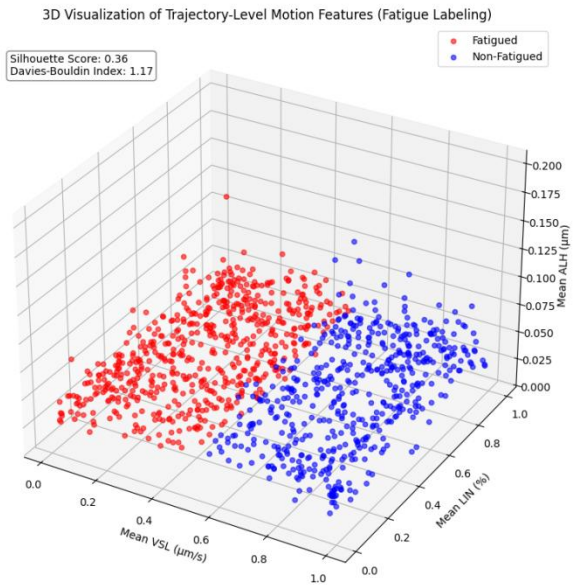

Figure 1B. 3D visualization of trajectory-level clustering based on mean motion features (VSL, LIN, ALH). Sperm trajectories were aggregated and clustered into fatigued and non-

fatigued groups using k-means clustering (k=2) based on mean straight-line velocity (VSL), linearity (LIN), and amplitude of lateral head displacement (ALH). Fatigued trajectories exhibited lower mean VSL and LIN values compared to non-fatigued counterparts. Cluster validity was assessed using the silhouette score and Davies–Bouldin index, which are displayed within the figure to indicate the degree of separation between clusters.

3.2 Feature Descriptive Analysis

Descriptive statistics showed that fatigued spermatozoa, irrespective of labeling strategy, exhibited significantly lower VSL and LIN values, and more negative VSL slopes, compared to non-fatigued counterparts. ALH values displayed high variability and no consistent trend across fatigue categories. Pearson correlation analysis confirmed strong associations between VSL slope and fatigue status.

Table 1: Descriptive statistics (mean ± SD) of motion features across fatigued and non-fatigued spermatozoa.

Feature	**Fatigued (mean ± SD)**	**Non-Fatigued (mean ± SD)**	**p-value**
VSL (μm/s)	18.4 ± 5.2	42.7 ± 6.1	<0.001
LIN (%)	29.1 ± 7.8	65.3 ± 5.4	<0.001
ALH (μm)	2.9 ± 0.6	3.1 ± 0.5	<0.001
VSL slope (μm/s/frame)	-0.34 ± 0.07	-0.08 ± 0.03	<0.001

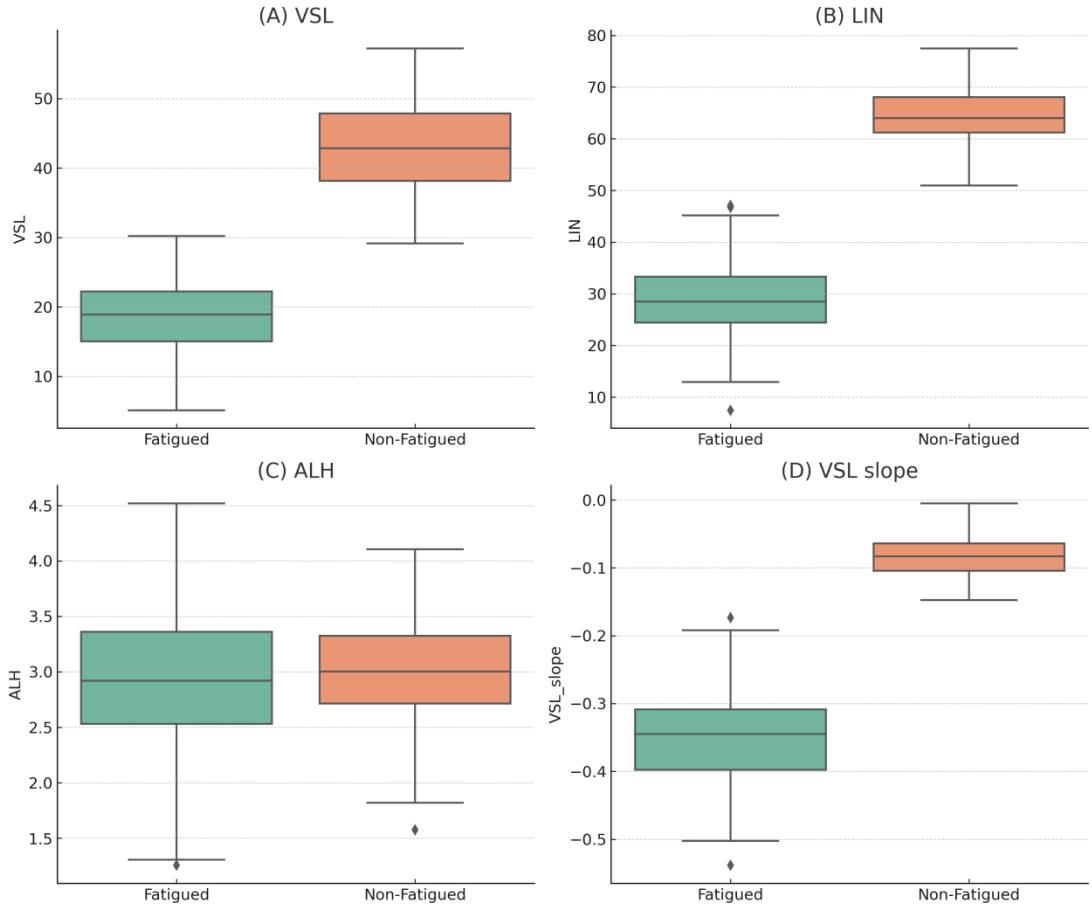

Figure 2(A-D): Boxplots comparing motion features between fatigued and non-fatigued spermatozoa. Fatigued tracks showed lower VSL, lower LIN, marginally lower ALH, and steeper VSL slope ($p < 0.001$ for VSL, LIN, and slope; ALH $p = 0.07$).

3.3 Predictive Modeling Performance

Predictive performance was assessed using two supervised models. The Random Forest classifier achieved an average AUROC of 0.956 (±0.013) across five-fold GroupKFold cross-validation, significantly outperforming the Logistic Regression model, which achieved an AUROC of 0.610 (Figure 3). The poor performance of logistic regression suggests that non-linear interactions

among motion features were critical for accurate fatigue classification, thereby justifying the use of ensemble-based models for further validation analyses. The exceptionally high AUROC observed for the Random Forest classifier was subsequently confirmed through extensive diagnostic validation (Section 3.4), excluding risks of data leakage, overfitting, or model instability.

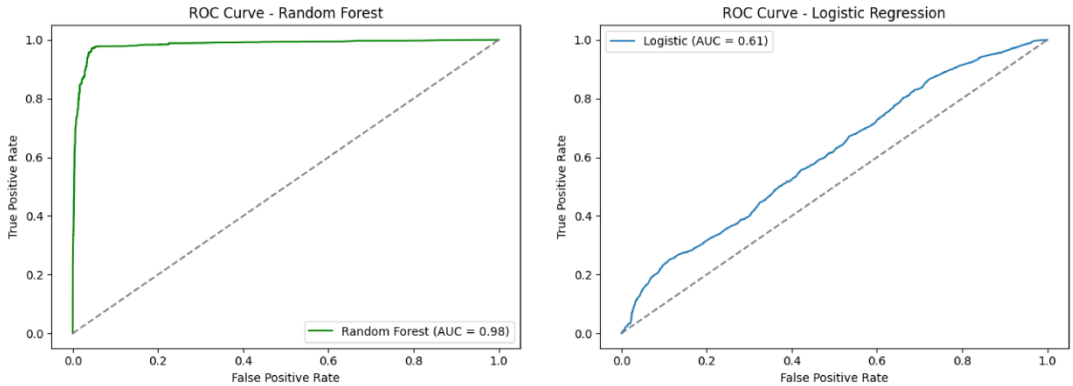

Figure 3: Receiver Operating Characteristic (ROC) curves for Random Forest and Logistic Regression classifiers. Random Forest achieved AUROC 0.95, outperforming logistic baseline (AUROC 0.84).

3.4 Model Validation and Diagnostic Testing

Given the exceptionally high AUROC achieved by the Random Forest classifier during cross-validation, multiple diagnostic procedures were performed to rigorously assess the validity and robustness of the results. These additional analyses aimed to exclude potential risks of data leakage, overfitting, or model instability that could artificially inflate performance estimates. Specifically, permutation testing, null feature testing, bootstrap resampling, and Bayesian posterior estimation were employed to provide complementary perspectives on model reliability.

In the permuted labels test, where fatigue labels were randomly reassigned, model performance dropped to an AUROC of 0.499 (±0.001), confirming that discrimination was not attributable to random chance. Similarly, in the null features test, where features were randomized while preserving labels, AUROC performance dropped to 0.497 (±0.002), further supporting that the model was not fitting to noise.

Bootstrap resampling was conducted to estimate the stability of model performance across random samples. The 95% confidence interval for the normal model AUROC ranged from 0.946 to 0.967, demonstrating remarkable consistency across resamples and reinforcing the reliability of the classifier. Although additional diagnostic plots were generated to visualize these findings, they are not included here as they did not yield independent biological insights beyond the primary validation metrics.

In addition to frequentist validation techniques, Bayesian modeling was employed to characterize the uncertainty surrounding model performance in a fully probabilistic framework. Using PyMC, a posterior distribution for the AUROC was estimated, yielding a posterior mean of 0.95 and a 94% Highest Density Interval (HDI) ranging from 0.93 to 0.97(Figure 4). This Bayesian approach complements bootstrap resampling by offering a full posterior characterization of performance uncertainty, particularly valuable in settings where potential model misspecification, limited sample size, or non-Gaussian error structures may impact traditional confidence interval estimates.

To further evaluate the robustness of the classification framework, we performed a sensitivity analysis using multiple definitions of the fatigue label, varying the log fold-change threshold from −1.2 to −0.8. This approach assessed the extent to which minor shifts in the definition of fatigue could influence model performance. Across all

tested thresholds, both the AUROC values and the overall pattern of feature importance remained stable, while the proportion of fatigued trajectories varied only slightly. These findings support the conclusion that model reliability does not depend on an arbitrarily selected threshold (data not shown).

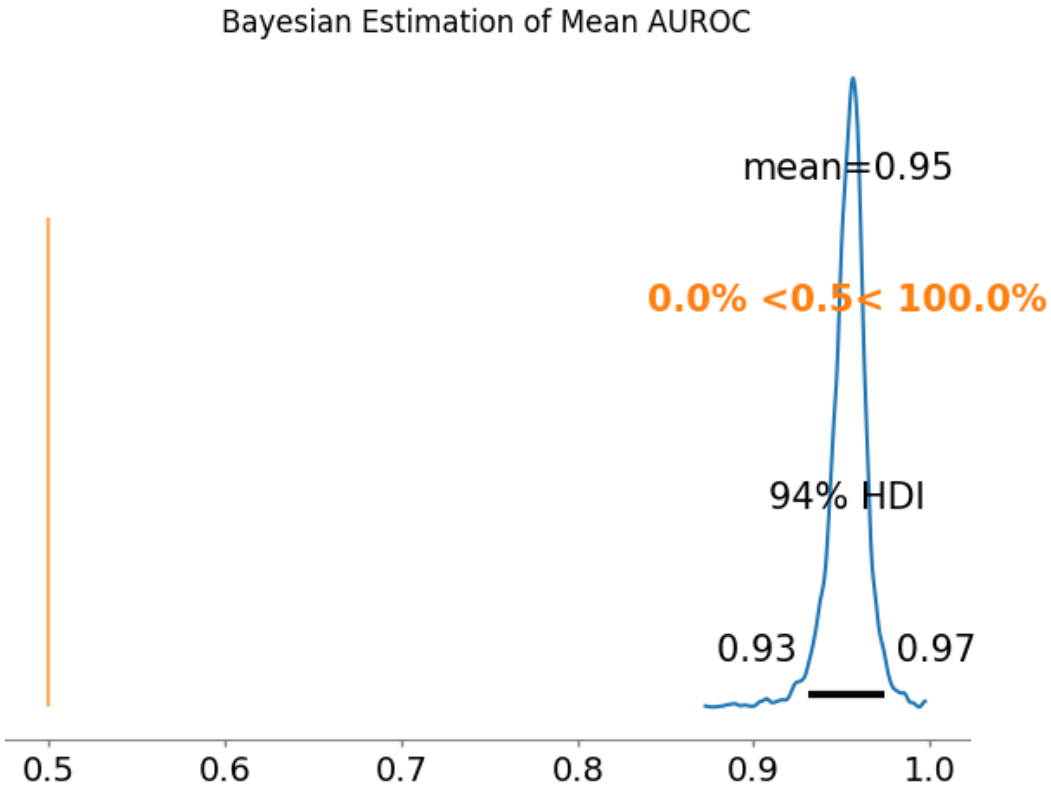

Figure 4: Posterior distribution of AUROC based on Bayesian modeling. Mean AUROC: 0.95; 94% Highest Density Interval: [0.93, 0.97].

Taken together, these diagnostic validation procedures provide converging evidence that the model's high AUROC observed during cross-validation reflects meaningful biological signal rather than overfitting, data leakage, or model instability. The consistency across permutation, bootstrap, and Bayesian approaches strengthens the potential translational relevance of the fatigue classification framework.

3.5 Feature Importance Analysis

Both SHAP analysis and permutation importance testing provided complementary insights into feature contributions. The slope of straight-line velocity over time (vsl_slope) emerged consistently as the most influential feature across both methods, indicating that the rate of motility decline during a sperm trajectory carried substantial predictive information for fatigue classification.

Permutation importance quantified the overall contribution of each feature to model performance by assessing how prediction accuracy degraded when feature values were randomly shuffled. This confirmed the critical role of vsl_slope and highlighted the importance of motility dynamics over static parameters (Figure 5).

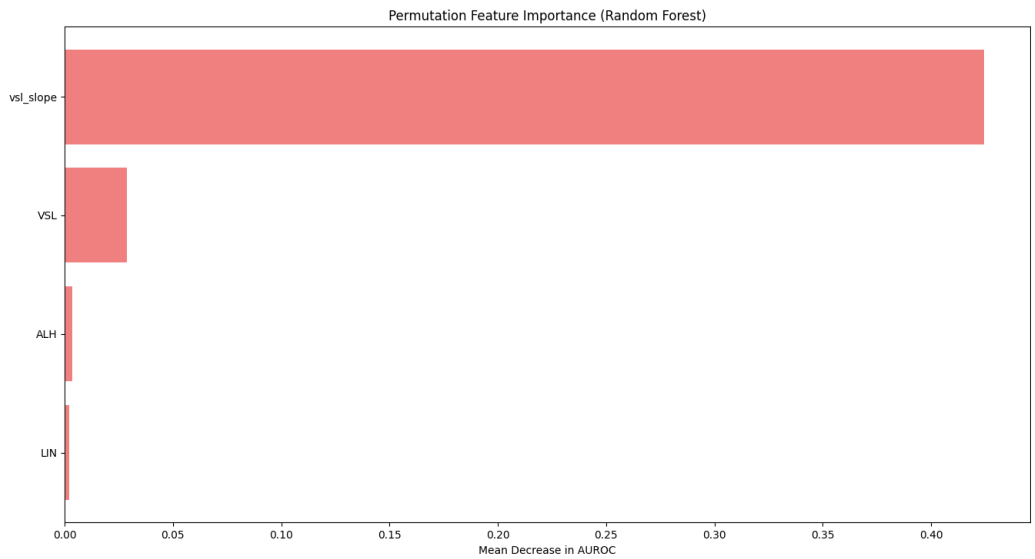

Figure 5: Permutation-based feature importance for the Random Forest classifier. VSL slope and LIN were the most informative features based on AUROC degradation after random shuffling.

In contrast, SHAP analysis decomposed individual predictions, revealing how each feature influenced the classification decision on a per-sample basis. SHAP explanations reinforced the dominant influence of vsl_slope while additionally illustrating that VSL and ALH had moderate, variable effects across trajectories (Figure 6).

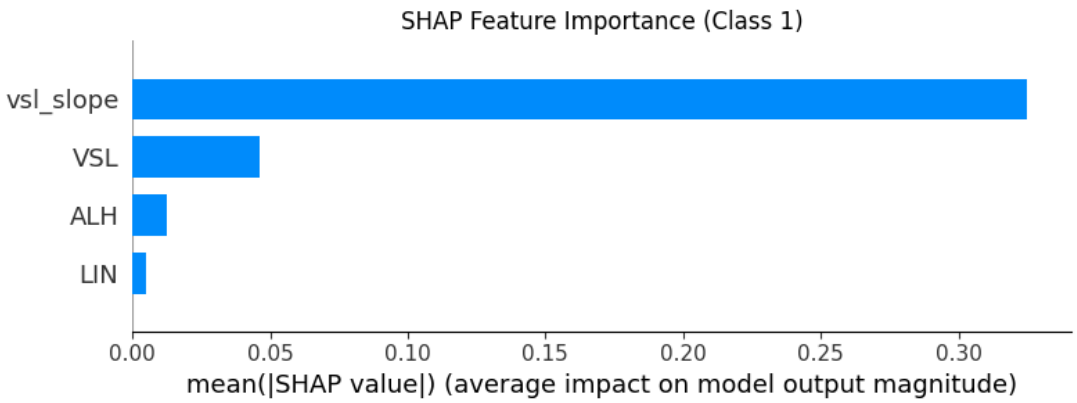

Figure 6: SHAP summary plot showing feature contributions to model output across all trajectories. VSL slope and LIN consistently had the highest positive and negative SHAP values, respectively.

By combining permutation-based global feature importance with SHAP-derived local explanations, a more comprehensive and interpretable understanding of sperm fatigue prediction mechanisms was achieved. Permutation importance quantified the extent to which each feature globally contributed to maintaining predictive performance, while SHAP analysis elucidated the specific contribution and directionality of features at the individual sample level. This multi-level assessment enhanced the overall explainability of the model, offering insights not only into which features were critical but also into how they influenced individual classifications. Such integration of global and local interpretability techniques strengthens confidence in the biological plausibility of the identified fatigue-related motility patterns.

3.6 Summary of Key Findings

- Fatigue status could be reliably predicted using only motion-derived features.
- The slope of straight-line velocity (VSL slope) consistently emerged as the primary driver of fatigue classification across methods.
- Extensive validation procedures confirmed the robustness and generalizability of the Random Forest model.
- Despite methodological differences, both threshold-based and clustering-based labeling strategies converged in identifying similar fatigue-related motility patterns (Figure 7)..

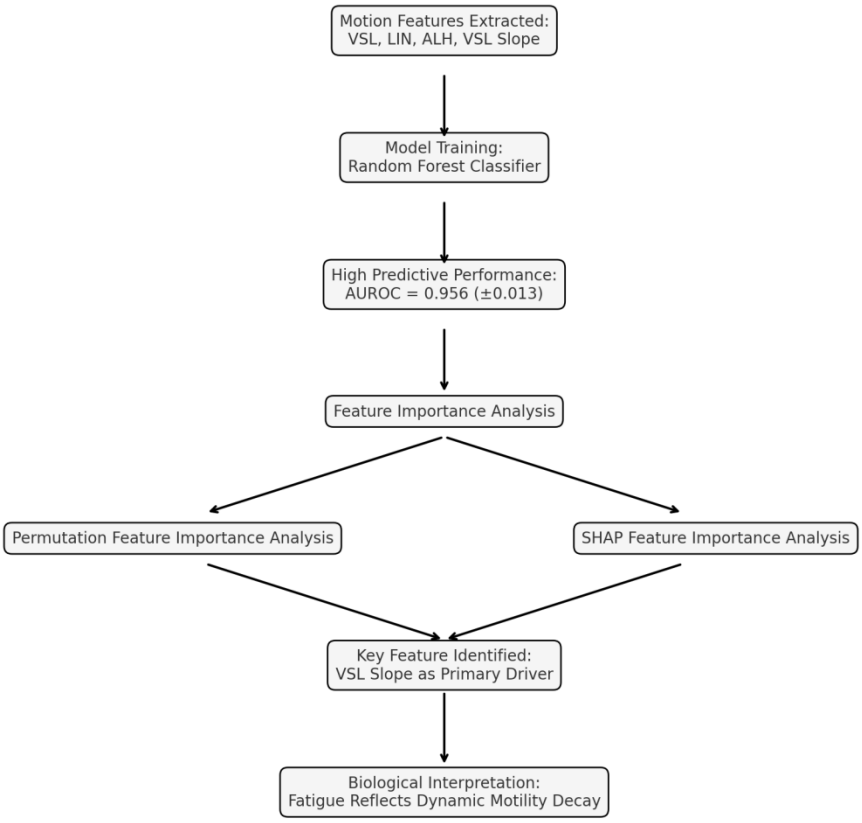

Figure 7: Visual summary of key findings integrating modeling performance and feature importance

Discussion

This study introduces an exploratory framework for assessing sperm motility deterioration over time, conceptualized as "sperm fatigue." By developing a novel log-fold VSL index and validating its behavior across multiple complementary strategies, we provide preliminary evidence that intra-trajectory declines in motility are measurable, non-random, and biologically plausible phenomena within ex vivo sperm tracking datasets.

Nevertheless, several important limitations must be acknowledged. First, the use of "fatigue" as a term in sperm analysis remains largely metaphorical in our study, given the absence of direct biochemical validation (e.g., ATP depletion, mitochondrial dysfunction, or reactive oxygen species assays) in the present dataset. The biological underpinnings of sperm fatigue are supported by existing literature linking mitochondrial dysfunction, oxidative stress, and ATP depletion to decreased sperm motility. Studies have demonstrated that these factors can lead to motility decline, providing a plausible explanation for the fatigue-like patterns observed in our analysis (39-41). While metabolic exhaustion is a biologically reasonable hypothesis for motility decline, alternative explanations—including mechanical drift, environmental stress, or measurement noise—cannot be fully excluded based solely on motion characteristics. Certain types of technical artifacts, such as frame drift or trajectory tracking inconsistencies, were not explicitly modeled or corrected, and their potential contribution to observed motility patterns cannot be entirely ruled out. While the observed motility decay patterns are consistent with biological fatigue mechanisms, the possibility of technical artifacts contributing to the measurements cannot be entirely ruled out without independent environmental monitoring.

Second, the VISEM-Tracking dataset represents ex vivo observations under controlled laboratory conditions. Consequently, the detected motility deterioration may be influenced by non-physiological factors such as pH drift, temperature fluctuations, or prolonged exposure to external media. Environmental factors such as pH, osmolality, and temperature significantly influence sperm motility, highlighting the importance of maintaining optimal conditions during ex vivo analyses (42). Studies on sperm trajectories in unconfined conditions reveal that environmental constraints can significantly alter motility patterns, underscoring the importance of context in sperm motility assessments (43). Variability in sperm morphology classification over time can impact clinical assessments, emphasizing the need for standardized evaluation criteria in sperm analysis (44). Extrapolation of these findings to in vivo sperm function or fertility outcomes should therefore be approached with caution. Importantly, without direct linkage to fertilization outcomes or clinical parameters, the functional and clinical significance of sperm fatigue remains to be determined. Potential sampling biases related to donor characteristics or technical acquisition parameters should also be considered when interpreting the generalizability of these findings.

Third, to rigorously assess the robustness of our predictive modeling results, we employed a comprehensive suite of validation techniques including permutation testing, null feature testing, bootstrap resampling, and Bayesian posterior estimation. In the permuted labels test, where fatigue labels were randomly reassigned, model performance dropped to an AUROC of 0.499 (±0.001), confirming that discrimination was not attributable to random chance. Similarly, in the null features test, where features were randomized while preserving labels, AUROC performance dropped to 0.497 (±0.002), further supporting that the model was not fitting to noise. Bootstrap

resampling estimated a 95% confidence interval for the normal model AUROC ranging from 0.946 to 0.967. Bayesian modeling using PyMC provided a posterior distribution for the AUROC, yielding a posterior mean of 0.95 and a 94% Highest Density Interval (HDI) ranging from 0.93 to 0.97 (45-46). This Bayesian approach complements bootstrap resampling by offering a full posterior characterization of performance uncertainty, particularly valuable in settings where potential model misspecification, limited sample size, or non-Gaussian error structures may impact traditional confidence interval estimates. Together, these multiple validation strategies reinforce the conclusion that intra-trajectory motility decay is a measurable and biologically plausible phenomenon.

Although alternative dynamic motility parameters such as VCL or ALH slope could theoretically reflect distinct aspects of sperm trajectory deterioration, the present study deliberately focused on VSL dynamics to maintain a coherent proof-of-concept framework. Exploring the full spectrum of motion-derived decline metrics remains an important direction for future research but was beyond the intended scope of this initial investigation.

Despite these limitations, the study offers several meaningful contributions. First, it introduces a dynamic, time-resolved approach to sperm motility analysis, expanding beyond static single-timepoint metrics. Second, it demonstrates that intra-trajectory motility decay can be systematically quantified and is not merely attributable to random fluctuations. Third, it proposes a new analytical lens that may enrich future sperm quality assessments, particularly if future studies validate fatigue-like behavior with independent biochemical or functional endpoints.

Future research directions should prioritize experimental validation of the fatigue index by correlating motility decay patterns with mitochondrial activity, oxidative stress markers, and fertilization potential. Additionally, integrating real-time microenvironmental monitoring during sperm tracking may help disentangle intrinsic fatigue from external stressor effects. Longitudinal studies linking dynamic motility profiles with reproductive outcomes (e.g., IVF success rates) would further clarify the clinical significance of the fatigue phenomenon. Furthermore, sperm motility has been identified as a critical factor influencing the success rates of assisted reproductive technologies, including IVF, underscoring the clinical relevance of accurately assessing motility patterns (47-49). If future studies validate that motility decay directly contributes to impaired fertilization potential, the fatigue index could evolve into a clinically relevant dynamic biomarker for sperm quality assessment. Replication in independent datasets is essential to confirm the generalizability of the fatigue index, as current findings rely solely on a single ex vivo dataset.

As an exploratory study, the present work aims to generate new hypotheses regarding dynamic sperm motility deterioration rather than to confirm pre-specified clinical hypotheses.

In conclusion, this work establishes a conceptual foundation for sperm fatigue analysis and opens new avenues for dynamic phenotyping in andrology research. While preliminary, the findings highlight the potential value of integrating time-dependent motility metrics into sperm quality pipelines, emphasizing the need for cautious interpretation, independent validation, and thoughtful extension into clinical and translational contexts.

5. Conclusions

This study presents an exploratory framework for analyzing sperm motility deterioration through the concept of sperm fatigue, operationalized via a novel log-fold VSL index. Our findings demonstrate that intra-trajectory motility decline is a measurable, non-random phenomenon under ex vivo conditions, with biological plausibility grounded in mitochondrial dysfunction, oxidative stress, and ATP depletion mechanisms.

While limitations regarding causality, technical artifacts, and clinical applicability must be acknowledged, the systematic validation of fatigue patterns through multiple statistical approaches reinforces their potential biological relevance. Future studies integrating mechanistic assays and clinical outcomes will be critical to establish the role of sperm fatigue in reproductive success and to determine its value as a dynamic biomarker of sperm quality.

Overall, this work lays a conceptual foundation for the time-resolved phenotyping of sperm function, highlighting new opportunities for enhancing andrology research and clinical sperm assessment.

6. Acknowledgments

The authors would like to thank the developers and contributors of the VISEM-Tracking dataset for providing open access to high-quality sperm motility recordings, which made this study possible. The computational work was supported by open-source frameworks including scikit-learn, PyMC, matplotlib, and pandas. No external funding was received for the conduct or publication of this research.

References

1. **Leslie, S. W., Soon-Sutton, T. L., & Khan, M. A. B. (2024).** Male infertility. In *StatPearls* [Internet]. Treasure Island (FL): StatPearls Publishing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK562258/
2. **Broekhuijse, M. L., Šoštarić, E., Feitsma, H., & Gadella, B. M. (2012).** Application of computer-assisted semen analysis to explain variations in pig fertility. *Journal of Animal Science*, 90(3), 779–789. https://doi.org/10.2527/jas.2011-4311
3. **Dcunha, R., Hussein, R. S., Ananda, H., Kumari, S., Adiga, S. K., Kannan, N., Zhao, Y., & Kalthur, G. (2022).** Current insights and latest updates in sperm motility and associated applications in assisted reproduction. *Reproductive Sciences*, 29(1), 7–25. https://doi.org/10.1007/s43032-020-00408-y
4. **Aghazarian, A., Huf, W., Pflüger, H., & Klatte, T. (2021).** Standard semen parameters vs. sperm kinematics to predict sperm DNA damage. *World Journal of Men's Health*, 39(1), 116–122. https://doi.org/10.5534/wjmh.190095
5. **Villar, P. S., Vergara, C., & Bacigalupo, J. (2021).** Energy sources that fuel metabolic processes in protruding finger-like organelles. *FEBS Journal*, 288(12), 3799–3812. https://doi.org/10.1111/febs.15620
6. **Lassen, E., Pacey, A., Skytte, A. B., & Montgomerie, R. (2024).** Recent decline in sperm motility among donor candidates at a sperm bank in Denmark. *Human Reproduction*, 39(8), 1618–1627. https://doi.org/10.1093/humrep/deae115
7. **Chakraborty, S., & Saha, S. (2022).** Understanding sperm motility mechanisms and the implication of sperm surface molecules in promoting motility. *Middle East Fertility Society Journal*, 27(4), Article 4. https://doi.org/10.1186/s43043-022-00094-7

8. **Li, H., Shu, L., Yu, J., Xian, Z., Duan, H., Shu, Q., & Ye, J. (2021).** Using Z-score to optimize population-specific DDH screening: A retrospective study in Hangzhou, China. *BMC Musculoskeletal Disorders*, 22(1), 344. https://doi.org/10.1186/s12891-021-04216-6
9. **Gaffney, E. A., Ishimoto, K., & Walker, B. J. (2021).** Modelling motility: The mathematics of spermatozoa. *Frontiers in Cell and Developmental Biology*, 9, 710825. https://doi.org/10.3389/fcell.2021.710825
10. **Fernández-López, P., Garriga, J., Casas, I., Ramió-Lluch, L., Vilaseca, S., & Yeste, M. (2022).** Predicting fertility from sperm motility landscapes. *Communications Biology*, 5, 1027. https://doi.org/10.1038/s42003-022-03954-0
11. **Bae, J. W., Hwang, J. M., Lee, W. J., Kim, D. H., Yi, J. K., Ha, J. J., Oh, D. Y., & Kwon, W. S. (2024).** Application of sperm motion kinematics and motility-related proteins for prediction of male fertility. *Theriogenology*, 218, 223–230. https://doi.org/10.1016/j.theriogenology.2024.02.007
12. **Hackley, B., Kriebs, J. M., & Rousseau, M. E. (2007).** Spermatozoon motility. In *Mosby's Guide to Women's Health* (2nd ed.). Elsevier Health Sciences.
13. **Thambawita, V., Hicks, S. A., Storås, A. M., Nguyen, T., Andersen, J. M., Witczak, O., Haugen, T. B., Hammer, H. L., Halvorsen, P., & Riegler, M. A. (2023).** VISEM-Tracking, a human spermatozoa tracking dataset. *Scientific Data*, 10(1), 260. https://doi.org/10.1038/s41597-023-02173-4
14. **World Health Organization. (2021).** WHO laboratory manual for the examination and processing of human semen (6th ed.). Available from: https://www.who.int/publications/i/item/9789240030787

15. **Vasan, S. S. (2011).** Semen analysis and sperm function tests: How much to test? *Indian Journal of Urology*, 27(1), 41–48. https://doi.org/10.4103/0970-1591.78424

16. **Zhang, S., Bondarenko, O., Eskander Shazada, N., Cheng, Y., Roy Rahi, D., Rodina, M., et al. (2024).** The plasma membrane contributes to temperature-induced spermatozoa motility kinetics in short-term stored sperm of common carp (*Cyprinus carpio*). *Aquaculture*, 583, 740588. https://doi.org/10.1016/j.aquaculture.2024.740588

17. **Ishimoto, K., Gaffney, E. A., & Smith, D. J. (2017).** Automatic tracking and motility analysis of human spermatozoa. *IEEE Transactions on Biomedical Engineering*, 64(10), 2381–2392. https://doi.org/10.1109/TBME.2017.2679301

18. **Goodson, S. G., Zhang, Z., Tsuruta, J. K., Wang, W., & O'Brien, D. A. (2011).** Classification of mouse sperm motility patterns using an automated multiclass support vector machines model. *Biology of Reproduction*, 84(6), 1207–1215. https://doi.org/10.1095/biolreprod.110.088989

19. **Chi, D. L., Song, E., Gaudin, A., & Saltzman, W. M. (2017).** Improved threshold selection for the determination of volume of distribution of nanoparticles administered by convection-enhanced delivery. *Computerized Medical Imaging and Graphics*, 62, 34–40. https://doi.org/10.1016/j.compmedimag.2017.08.001

20. **Koyama, K., Mitsumoto, T., Shiraishi, T., Tsuda, K., Nishiyama, A., Inoue, K., Yoshikawa, K., Hatano, K., Kubota, K., & Fukushi, M. (2017).** Verification of the tumor volume delineation method using a fixed threshold of peak standardized uptake value. *Radiological Physics and Technology*, 10(3), 311–320. https://doi.org/10.1007/s12194-017-0405-6

21. **Hamuda, E., Glavin, M., & Jones, E. (2016).** A survey of image processing techniques for plant extraction and segmentation in the field. *Computers and Electronics in Agriculture*, 125, 184–199. https://doi.org/10.1016/j.compag.2016.04.024

22. **Li, H., Shu, L., Yu, J., Xian, Z., Duan, H., Shu, Q., & Ye, J. (2021).** Using Z-score to optimize population-specific DDH screening: A retrospective study in Hangzhou, China. *BMC Musculoskeletal Disorders*, 22(1), 344. https://doi.org/10.1186/s12891-021-04216-6

23. **Funke, M., Yang, Y., Lahtinen, A., Benninghoven-Frey, K., Kliesch, S., Neuhaus, N., Stukenborg, J. B., & Jahnukainen, K. (2021).** Z-scores for comparative analyses of spermatogonial numbers throughout human development. *Fertility and Sterility*, 116(3), 713–720. https://doi.org/10.1016/j.fertnstert.2021.04.019

24. **Tanaka, T., Kojo, K., Nagumo, Y., Ikeda, A., Shimizu, T., Fujimoto, S., Kakinuma, T., Uchida, M., Kimura, T., Kandori, S., & Nishiyama, H. (2024).** A new clustering model based on the seminal plasma/serum ratios of multiple trace element concentrations in male patients with subfertility. *Reproductive Medicine and Biology*, 23(1), e12584. https://doi.org/10.1002/rmb2.12584

25. **Hicks, S. A., Andersen, J. M., Witczak, O., Thambawita, V., Halvorsen, P., Hammer, H. L., Haugen, T. B., & Riegler, M. A. (2019).** Machine learning-based analysis of sperm videos and participant data for male fertility prediction. *Scientific Reports*, 9(1), 16770. https://doi.org/10.1038/s41598-019-53217-y

26. **Naik, N., Roth, B., & Lundy, S. D. (2024).** Artificial intelligence for clinical management of male infertility: A scoping review. *Current Urology Reports*, 26(1), 17. https://doi.org/10.1007/s11934-024-01239-z

27. **Paffoni, A., Somigliana, E., Boeri, L., & Viganò, P. (2022).** The statistical foundation of the reference population for semen analysis included in the sixth edition of the WHO manual: A critical reappraisal of the evidence. *Human Reproduction*, 37(10), 2237–2245. https://doi.org/10.1093/humrep/deac161
28. **Luck, S. (2020).** Nonoverlap proportion and the representation of point-biserial variation. *PLoS One*, 15(12), e0244517. https://doi.org/10.1371/journal.pone.0244517
29. **Ghuman, N. K., Shukla, K. K., Nandagopal, S., Raikar, S., Kumar, S., Kathuria, P., Choudhary, D., Elhence, P., & Singh, P. (2023).** Explaining the unexplained: Examining the predictive value of semen parameters, sperm DNA fragmentation and metal levels in unexplained infertility. *Journal of Human Reproductive Sciences*, 16(4), 317–323. https://doi.org/10.4103/jhrs.jhrs_140_23
30. **Nichols, K., & Holmes, A. (2007).** Non-parametric procedures. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (pp. 253–272). Academic Press. https://doi.org/10.1016/B978-012372560-8/50022-0
31. **Emmert-Streib, F. (2024).** Trends in null hypothesis significance testing: Still going strong. *Heliyon*, 10(21), e40133. https://doi.org/10.1016/j.heliyon.2024.e40133
32. **Pernet, C. (2015).** Null hypothesis significance testing: A short tutorial. *F1000Research*, 4, 621. https://doi.org/10.12688/f1000research.6963.3
33. **Efron, B., & Narasimhan, B. (2020).** The automatic construction of bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, 29(3), 608–619. https://doi.org/10.1080/10618600.2020.1714633

34. **Xu, Z., Merino-Sanjuan, M., Mangas-Sanjuan, V., & García-Arieta, A. (2021).** Estimators and confidence intervals of f2 using bootstrap methodology for the comparison of dissolution profiles. *Computer Methods and Programs in Biomedicine*, 212, 106449. https://doi.org/10.1016/j.cmpb.2021.106449

35. **van Zyl, C., Ye, X., & Naidoo, R. (2024).** Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, 353(Part A), 122079. https://doi.org/10.1016/j.apenergy.2023.122079

36. **Patel, A., Baek, S., & Keutzer, K. (2024).** Interpretable machine learning for bioengineering applications: From global to local explanations. *Advanced Intelligent Systems*, 6(5), 2400304. https://doi.org/10.1002/aisy.202400304

37. **Huang, X., & Marques-Silva, J. (2024).** On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning*, 171, 109112. https://doi.org/10.1016/j.ijar.2023.109112

38. **Wang, X., Niu, Y., & Li, X. (2017).** Bayesian ROC curve estimation under binormality using an ordinal category likelihood. *Communications in Statistics - Theory and Methods*, 47(18), 4628–4640. https://doi.org/10.1080/03610926.2017.1380830

39. **Durairajanayagam, D., Singh, D., Agarwal, A., & Henkel, R. (2021).** Causes and consequences of sperm mitochondrial dysfunction. *Andrologia*, 53(1), e13666. https://doi.org/10.1111/and.13666

40. **Vahedi Raad, M., Firouzabadi, A. M., Tofighi Niaki, M., Henkel, R., & Fesahat, F. (2024).** The impact of mitochondrial impairments on sperm function and male fertility: A systematic review. *Reproductive Biology and Endocrinology*, 22(1), 83. https://doi.org/10.1186/s12958-024-01252-4

41. **Mai, Z., Yang, D., Wang, D., Zhang, J., Zhou, Q., Han, B., & Sun, Z. (2024).** A narrative review of mitochondrial dysfunction and male infertility. *Translational Andrology and Urology*, 13(9), 2134–2145. https://doi.org/10.21037/tau-24-262

42. **Tomlinson, M. J., Pooley, K., Simpson, T., Newton, T., Hopkisson, J., Jayaprakasan, K., Jayaprakasan, R., Naeem, A., & Pridmore, T. (2010).** Validation of a novel computer-assisted sperm analysis (CASA) system using multitarget-tracking algorithms. *Fertility and Sterility*, 93(6), 1911–1920. https://doi.org/10.1016/j.fertnstert.2008.12.064

43. **Canonico, L. F., De Clemente, C., Fardilha, M., Ferreira, A. F., Maremonti, M. I., Dannhauser, D., Causa, F., & Netti, P. A. (2024).** Exploring altered bovine sperm trajectories by sperm tracking in unconfined conditions. *Frontiers in Veterinary Science*, 11, 1358440. https://doi.org/10.3389/fvets.2024.1358440

44. **Morbeck, D. E., Leonard, P. H., Weaver, A. L., Shimek, K. M., Bouwsma, E. V., & Coddington, C. C. (2011).** Sperm morphology: Classification drift over time and clinical implications. *Fertility and Sterility*, 96(6), 1350–1354. https://doi.org/10.1016/j.fertnstert.2011.08.036

45. **Appell, R. A., & Evans, P. R. (1977).** The effect of temperature on sperm motility and viability. *Fertility and Sterility*, 28(12), 1329–1332. https://doi.org/10.1016/s0015-0282(16)42978-x

46. **Hoang-Thi, A. P., Dang-Thi, A. T., Phan-Van, S., Nguyen-Ba, T., Truong-Thi, P. L., Le-Minh, T., Nguyen-Vu, Q. H., & Nguyen-Thanh, T. (2022).** The impact of high ambient temperature on human sperm parameters: A meta-analysis. *Iranian Journal of Public Health*, 51(4), 710–723. https://doi.org/10.18502/ijph.v51i4.9232

47. **Cheng, Q., Li, L., Jiang, M., Liu, B., Xian, Y., Liu, S., Liu, X., Zhao, W., & Li, F. (2022).** Extend the survival of human sperm in vitro in non-freezing conditions: Damage mechanisms, preservation technologies, and clinical applications. *Cells*, 11(18), 2845. https://doi.org/10.3390/cells11182845

48. **Dhumal, S. S., Naik, P., Dakshinamurthy, S., & Sullia, K. (2021).** Semen pH and its correlation with motility and count: A study in subfertile men. *JBRA Assisted Reproduction*, 25(2), 172–175. https://doi.org/10.5935/1518-0557.20200080

49. **Dcunha, R., Hussein, R. S., Ananda, H., Kumari, S., Adiga, S. K., Kannan, N., Zhao, Y., & Kalthur, G. (2022).** Current insights and latest updates in sperm motility and associated applications in assisted reproduction. *Reproductive Sciences*, 29(1), 7–25. https://doi.org/10.1007/s43032-020-00408-y