Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

Comparison of XGBoost and Random Forest for Prediction of Male Fertility Status Based on Semen Analysis Parameters

Kecitaan Harefa*, Joko Priambodo

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: 1,*dosen00842@unpam.ac.id, 2dosen00276@unpam.ac.id
Correspondence Author Email: dosen00842@unpam.ac.id
Submitted: **10/06/2026**; Accepted: **30/06/2026**; Published**: 30/06/2026**

Abstract−Male infertility is a major reproductive health problem that contributes to approximately half of infertility cases among couples of reproductive age. Accurate evaluation of male fertility status commonly relies on semen analysis, including parameters such as semen volume, sperm concentration, motility, morphology, and vitality. However, manual interpretation of these parameters remains time-consuming and is highly dependent on clinical expertise. This study aims to compare the performance of the XGBoost and Random Forest algorithms in predicting male fertility status based on semen analysis parameters. The study employed a secondary dataset consisting of 1,000 semen analysis records with 11 predictor variables and one target variable representing fertility status. Data preprocessing included categorical encoding, data cleaning, and an 80:20 train–test split before model development and evaluation using accuracy, precision, recall, F1-score, and ROC-AUC. Experimental results showed that XGBoost outperformed Random Forest, achieving an accuracy of 99.00%, precision of 100.00%, recall of 96.88%, F1-score of 98.41%, and ROC-AUC of 99.86%, while Random Forest achieved an accuracy of 97.50%. Feature importance analysis identified Total Motility, Vitality, and Progressive Motility as the most influential predictors of male fertility status. The main contribution of this study is the direct empirical comparison of Random Forest and XGBoost under identical experimental settings using comprehensive semen analysis parameters, providing evidence on the relative effectiveness of ensemble learning algorithms for male fertility status prediction. Although the proposed model demonstrates excellent predictive performance, it was developed using secondary data and is intended to support, rather than replace, clinical decision-making. Future studies should validate the model using larger multicenter clinical datasets to improve its generalizability and practical applicability.

Keywords: XGBoost; Random Forest; Male Fertility Status; Semen Analysis; Machine Learning

1. INTRODUCTION

Infertility remains one of the major reproductive health problems worldwide and has become an important public health concern due to its impact on individuals, families, and society. According to the World Health Organization (WHO), infertility is defined as the inability of a couple to achieve pregnancy after at least 12 months of regular unprotected sexual intercourse [1]. This condition affects millions of couples globally and is associated not only with reproductive disorders but also with psychological, social, and economic consequences. Previous studies have reported that male factors contribute to approximately 40–50% of infertility cases, highlighting the importance of accurate evaluation of male reproductive health during infertility assessment [2].

Male fertility status is commonly assessed through semen analysis, which evaluates several laboratory parameters, including semen volume, pH, sperm concentration, total motility, progressive motility, morphology, and sperm vitality [3]. These parameters provide essential information regarding the functional quality of sperm and are widely used as clinical indicators for determining male fertility status. However, the interpretation of semen analysis results is still largely performed manually by andrologists or laboratory specialists. This conventional approach is time-consuming, highly dependent on examiner expertise, and susceptible to inter-observer variability, potentially leading to inconsistent clinical interpretations. Furthermore, the increasing volume of laboratory data has created a growing need for automated analytical tools capable of providing faster, more objective, and reproducible decision support [4].

Recent advances in Artificial Intelligence (AI), particularly Machine Learning, have significantly improved predictive modeling in healthcare. Machine learning algorithms are capable of discovering complex nonlinear relationships within clinical datasets and have demonstrated promising performance in disease diagnosis, prognosis prediction, and clinical decision support systems [5][6]. In reproductive medicine, machine learning provides an opportunity to analyze semen examination data objectively and to assist clinicians in identifying male fertility status more efficiently than conventional manual approaches [7]. Among various machine learning algorithms, Random Forest and Extreme Gradient Boosting (XGBoost) have received considerable attention because of their excellent predictive performance on structured tabular medical datasets [8][9]. Random Forest offers robust classification through ensemble decision trees and effectively reduces overfitting, whereas XGBoost enhances prediction accuracy by sequentially correcting previous prediction errors through gradient boosting and regularization mechanisms and has consistently demonstrated competitive performance in structured biomedical datasets [10].

Several recent studies have demonstrated the effectiveness of machine learning techniques in reproductive health. Mehrjerd et al. developed ensemble learning models, including Random Forest and XGBoost, to predict clinical pregnancy success in Assisted Reproductive Technology (ART). Although the study demonstrated promising predictive capability, the prediction target was pregnancy outcome rather than male fertility status based on semen examination parameters [11]. Huang et al. utilized Random Forest and other machine learning algorithms to predict sperm count using general health screening indicators. Their findings confirmed the usefulness of machine learning

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

in reproductive health prediction; however, the proposed model relied primarily on general health variables instead of comprehensive semen analysis parameters [12].

Similarly, GhoshRoy et al. developed explainable machine learning models for predicting male fertility using lifestyle-related variables combined with SMOTE and SHAP analysis. While their approach improved model interpretability, the prediction was mainly based on demographic and lifestyle information rather than laboratory semen parameters [13]. Mohammadi et al. investigated occupational exposure factors affecting male reproductive health using several machine learning algorithms, including Random Forest and XGBoost. Their study achieved excellent predictive performance; nevertheless, occupational and environmental variables were the primary predictors instead of routine semen examination results [14]. Chen et al. applied XGBoost to predict infertility treatment outcomes in patients undergoing IVF/ICSI procedures. Although the proposed model successfully predicted treatment outcomes, the study focused on assisted reproductive therapy rather than direct classification of male fertility status from semen analysis data [15].

A critical analysis of previous studies reveals several important research gaps. First, most existing studies evaluated only a single machine learning algorithm or focused on optimizing one predictive model, making it difficult to determine which algorithm performs better when applied to the same semen analysis dataset. Second, previous studies frequently employed heterogeneous predictor variables such as demographic information, lifestyle factors, occupational exposure, hormonal measurements, or ART-related clinical indicators, whereas comprehensive routine semen analysis parameters have received comparatively less attention as the primary predictors of male fertility status. Third, although Random Forest and XGBoost are recognized as two of the most powerful ensemble learning algorithms for structured medical data, direct comparative studies between these algorithms using identical experimental settings and standardized evaluation metrics remain limited [16]. Consequently, there is still insufficient empirical evidence regarding which algorithm is more suitable for predicting male fertility status using routine semen analysis data, highlighting the need for further validation and comparative studies [17].

Male infertility remains a significant and growing global health challenge, making the development of rapid, accurate diagnostic tools essential. Considering current research gaps particularly the lack of robust, head-to-head algorithmic assessments a direct comparison between Random Forest and XGBoost has become increasingly important. Both of these tree-based ensemble methods have shown immense promise in handling complex, non-linear medical data, yet their relative performance in reproductive health remains underexplored. Accurate prediction of male fertility status using routinely available semen examination parameters holds tremendous practical value. Because these standard macroscopic and microscopic laboratory tests are already commonly performed in fertility clinics and hospitals worldwide, leveraging them through machine learning maximizes the utility of existing data. Identifying the most appropriate algorithm can substantially improve diagnostic efficiency, reduce the inherent subjective interpretation associated with manual semen analysis, and provide reliable, real-time decision support for clinicians. Crucially, this can be achieved without imposing the financial or physical burden of additional laboratory examinations on patients. Furthermore, conducting a comparative evaluation using identical datasets and standardized evaluation metrics provides much stronger, highly reproducible scientific evidence compared to studies that evaluate a single algorithm independently

Therefore, this study compares the performance of Random Forest and XGBoost for predicting male fertility status based on comprehensive semen analysis parameters obtained from a secondary dataset consisting of 1,000 records. Both algorithms are evaluated using identical preprocessing procedures and the same performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to ensure a fair comparison. The main contribution of this study is a direct empirical comparison of Random Forest and XGBoost using identical preprocessing procedures, data partitions, and evaluation metrics on comprehensive semen analysis parameters. Unlike previous studies that primarily focused on lifestyle factors, occupational exposures, general health indicators, or assisted reproductive treatment outcomes, this study specifically evaluates the relative performance of two ensemble learning algorithms for direct male fertility status classification. The findings provide comparative evidence regarding algorithm selection and identify the semen analysis parameters that contribute most strongly to fertility status prediction, thereby supporting the development of data-driven decision-support approaches in andrology.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study aims to compare the performance of the Extreme Gradient Boosting (XGBoost) and Random Forest algorithms in predicting male fertility status based on semen analysis parameters. The proposed methodology consists of six sequential stages: dataset collection, data preprocessing, data splitting, model development, model evaluation, and comparative performance analysis. A secondary tabular dataset containing routine semen examination parameters was used as the input for both machine learning models. The workflow of the proposed study illustrated in Figure 1.

Based on Figure 1, the research begins with dataset collection, followed by data preprocessing to improve data quality and prepare the variables for machine learning. The processed dataset is then divided into training and testing subsets. Random Forest and XGBoost models are subsequently trained using the same training dataset and evaluated

Copyright © 2026 **Author**, Page 539
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

on the same testing dataset to ensure a fair comparison. Finally, the predictive performance of both algorithms is compared using several evaluation metrics to determine the most effective model for predicting male fertility status.

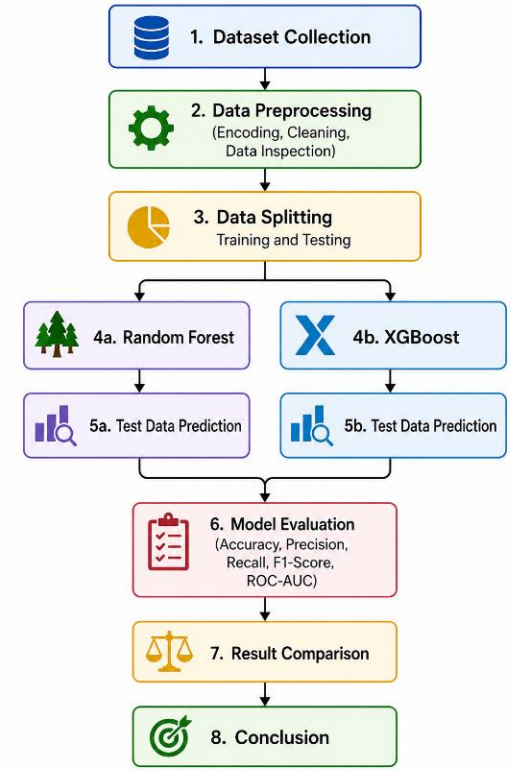

Figure 1. Research Stages

2.2 Dataset Collection

This study employed a secondary dataset obtained from the Kaggle platform, which contains routine semen analysis records for male fertility assessment. The dataset consists of 1,000 observations, including 11 predictor variables and one target variable (Fertility_Status). The predictor variables comprise Age, Body Mass Index (BMI), Smoking, Alcohol Consumption, Semen Volume, Semen pH, Sperm Concentration, Total Motility, Progressive Motility, Normal Morphology, and Vitality. The dataset was selected because it contains laboratory parameters routinely used in clinical andrological examinations according to WHO recommendations, making it appropriate for evaluating machine learning algorithms for male fertility status prediction. Since the dataset is publicly available and anonymized, no personally identifiable information was included, and no additional ethical approval was required.

Table 1. Research Variables

No	Variable	Remarks
1	Age	Patient's age
2	BMI	Body Mass Index
3	Smoking	Smoking status
4	Alcohol	Alcohol consumption status
5	Semen_Volume_ml	Semen volume (mL)
6	pH	Acidity level of semen
7	Sperm_Concentration_Million_ml	Sperm concentration
8	Total_Motility_pct	Total motility percentage
9	Progressive_Motility_pct	Progressive motility percentage
10	Morphology_Normal_pct	Normal morphological percentage
11	Vitality_pct	Percentage of sperm vitality
12	Fertility_Status	Fertility status (Target)

All experiments were conducted using a consistent software environment to ensure the reproducibility of the proposed models. The experimental environment used in this study is presented in Table 2.

Table 2. Experimental Environment

Component	Specification
Programming Language	Python 3.11

Copyright © 2026 **Author**, Page 540
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

Component	Specification
Development Platform	Google Colaboratory
Operating Environment	Cloud-based Jupyter Notebook
Data Analysis Library	Pandas 2.x
Numerical Computing	NumPy 2.x
Machine Learning Library	Scikit-learn 1.5
Gradient Boosting Library	XGBoost 2.x
Visualization Library	Matplotlib
Random State	42
Validation Method	Stratified 5-Fold Cross Validation

The use of a fixed software environment, library versions, and random state ensures that the experimental procedure can be reproduced consistently by other researchers.

2.3 Data Preprocessing

The preprocessing stage was conducted to ensure data quality before model training, following recommended practices for structured healthcare datasets [18]. Initially, data inspection was performed to examine the number of records, data types, and the presence of missing values. This process ensured that the dataset was suitable for subsequent analysis and model development. Categorical variables, including Smoking, Alcohol, and Fertility_Status, were subsequently transformed into numerical representations using label encoding. The Smoking and Alcohol variables were encoded as No = 0 and Yes = 1, while the target variable Fertility_Status was encoded as Fertile = 0 and Infertile = 1. The encoding results are presented in Table 3.

Table 3. Results of Encoding Categorical Variables

Variable	Original Value	Label Encoding
Smoking	No	0
Smoking	Yes	1
Alcohol	No	0
Alcohol	Yes	1
Fertility_Status	Fertile	0
Fertility_Status	Infertile	1

After the data preprocessing stage, the dataset was separated into independent variables (X) and a dependent variable (y). The independent variables consisted of 11 predictor attributes, including Age, BMI, Smoking, Alcohol, Semen Volume, pH, Sperm Concentration, Total Motility, Progressive Motility, Normal Morphology, and Vitality, while Fertility_Status was defined as the target variable. Subsequently, the dataset was divided into training and testing subsets using an 80:20 ratio. The training dataset, comprising 80% of the total data, was used to train the Random Forest and XGBoost models, whereas the remaining 20% was reserved as an independent testing dataset for evaluating model performance on previously unseen data. The data splitting process was performed using stratified sampling to preserve the class distribution of Fertile and Infertile samples in both subsets, with a fixed random state of 42 to ensure reproducibility.

2.4 Random Forest Algorithm

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their prediction results through a majority voting mechanism. The process begins with bootstrap sampling to generate different subsets of the training data. Multiple decision trees are then constructed using these subsets, with a random selection of features considered at each node during the splitting process. Each decision tree independently produces a class prediction, and the final classification result is determined by selecting the class that receives the majority of votes from all decision trees. The final prediction of the Random Forest model is formulated in Equation (1).

$$\hat{Y} = Mode(T_1, T_2, T_3, \dots, T_n) \quad (1)$$

In Equation (1), $\hat{Y}$ represents the final prediction generated by the Random Forest model, Ti represents the prediction produced by the i-th decision tree, and n denotes the total number of decision trees used in the ensemble. The majority voting mechanism selects the most frequently predicted class among all decision trees as the final classification result.

2.5 XGBoost Algorithm

XGBoost is an advanced implementation of the Gradient Boosting algorithm that constructs models sequentially, where each new model is developed to correct the prediction errors generated by the previous model. This iterative learning mechanism enables the algorithm to progressively improve its predictive performance. The XGBoost objective function combines the loss function and a regularization term, as formulated in Equation (2).

Copyright © 2026 **Author**, Page 541
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (2)$$

In Equation (2), Obj(θ) represents the objective function of the XGBoost model, L(θ) denotes the loss function used to measure the difference between the predicted and actual values, and Ω(θ) represents the regularization term used to control model complexity. The regularization mechanism helps reduce the risk of overfitting and improves model generalization. Furthermore, XGBoost is well suited for processing structured tabular data and provides efficient computational performance through its optimized boosting mechanism. These characteristics make XGBoost a suitable classification algorithm for analyzing semen analysis parameters and predicting male fertility status.

2.6 Model Hyperparameter Configuration

Before model training, the hyperparameters of both Random Forest and XGBoost were configured to obtain stable predictive performance while ensuring reproducibility of the experiments. The selected hyperparameter values were consistently applied throughout the model training and testing processes. In addition, a fixed random_state value of 42 was used for both algorithms to ensure that the experimental results could be reproduced.

The Random Forest model was configured using 200 decision trees (n_estimators = 200) with no restriction on the maximum tree depth (max_depth = None), allowing each decision tree to grow until all leaves were pure or no further splitting was possible. A fixed random_state = 42 was applied to ensure consistent model initialization and reproducibility across different experimental runs.

The XGBoost model was configured using 200 boosting trees (n_estimators = 200), a learning rate of 0.05 (learning_rate = 0.05), and a maximum tree depth of 4 (max_depth = 4). Furthermore, subsample = 0.8 and colsample_bytree = 0.8 were employed to improve model generalization by randomly sampling training instances and input features during the boosting process. The evaluation metric was set to logloss (eval_metric = 'logloss'), and a fixed random_state = 42 was used to ensure reproducible experimental results.

To further improve the reliability of model evaluation, both algorithms were assessed using Stratified 5-Fold Cross Validation, ensuring that each fold preserved the original class distribution. After cross-validation, the final models were trained using the training dataset and evaluated on an independent testing dataset comprising 20% of the total observations. This evaluation strategy helps reduce the risk of data leakage while providing a more reliable estimate of model performance.

Table 4. Hyperparameter Configuration

Parameter	Random Forest	XGBoost
n_estimators	200	200
max_depth	None	4
learning_rate	–	0.05
subsample	–	0.8
colsample_bytree	–	0.8
eval_metric	–	logloss
random_state	42	42

2.7 Model Testing

After the training process was completed, both models were evaluated using testing data that had not been used during model training. To reduce the risk of overfitting and data leakage, the dataset was divided into training and testing subsets using a stratified train-test split with a ratio of 80:20. Model development and hyperparameter configuration were performed exclusively using the training dataset, whereas the testing dataset was reserved for final model evaluation. In addition, Stratified 5-Fold Cross Validation was conducted to assess the stability and generalization capability of both models, as this evaluation strategy is commonly used to reduce optimistic bias in predictive modeling [19].

The performance of the Random Forest and XGBoost models was evaluated using accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Accuracy measures the proportion of correctly classified samples relative to the total number of samples and is formulated in Equation (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision measures the proportion of correctly predicted positive samples among all samples classified as positive. Precision is formulated in Equation (4).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall measures the model's ability to correctly identify positive samples from all actual positive samples. In the context of this study, recall is particularly important because it reflects the model's ability to identify infertile cases and minimize false-negative predictions. Recall is formulated in Equation (5).

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

The F1-score represents the harmonic mean of precision and recall and provides a balanced evaluation of both metrics. The F1-score is formulated in Equation (6).

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

ROC-AUC was used to evaluate the ability of each model to distinguish between the Fertile and Infertile classes across different classification thresholds. A higher ROC-AUC value indicates a stronger discriminative ability of the model. In this study, these evaluation metrics were used collectively to provide a comprehensive comparison of the predictive performance of Random Forest and XGBoost.

2.8 Results Analysis

The final stage of the study was carried out by comparing the accuracy, precision, recall, F1-score, and ROC-AUC values of the Random Forest and XGBoost algorithms. The algorithm that produces the highest evaluation value will be considered to have the best performance in predicting male fertility status based on semen analysis parameters. The results of the comparison are then used as a basis for the preparation of research conclusions [20].

3. RESULT AND DISCUSSION

3.1 Dataset Processing Results

This study uses a semen analysis dataset for male fertility status prediction consisting of 1000 records with 11 independent variables and 1 dependent variable. Independent variables include Age, BMI, Smoking, Alcohol, Semen Volume, pH, Sperm Concentration, Total Motility, Progressive Motility, Normal Morphology, and Vitality. Meanwhile, the dependent variable used is Fertility Status, which consists of two classes, namely Fertile and Infertile.

The initial stage of the research was carried out by reading the dataset using the Pandas library on Google Colab. Next, an examination of the data structure, data type, and the existence of missing values was carried out. Based on the results of the examination, it is known that all attributes have complete data, so there is no need for an imputation process or handling of blank data. This condition indicates that the dataset is ready to be used for machine learning processes.

In the next stage, categorical data transformation is carried out using the Label Encoding method. The Smoking, Alcohol, and Fertility Status variables are converted into numerical forms so that they can be processed by the Random Forest and XGBoost algorithms. After the encoding process is complete, all variables are in numerical format so that they can be used as the input to the classification model.

The dataset is then divided into two parts, namely 80% training data and 20% testing data. With a total of 1000 data points, 800 data points were obtained for the model training process and 200 data points for the model testing process. The dataset was divided using the train-test split method with fixed random state parameters so that the results of the study could be reproduced.

3.2 Results of Applying the Random Forest Algorithm

The first algorithm used in this study was Random Forest. The algorithm is an ensemble learning method that builds many decision trees and then combines the prediction results of all trees using a majority voting mechanism. In this study, as many as 200 decision trees (n_estimators = 200) were used to improve the stability of the model.

After the training process is complete, the model is used to make predictions on the test data. The results of the evaluation of the Random Forest model are shown in Table 5.

Table 5. Random Forest Evaluation Results

Metric	Value
Accuracy	97,50%
Precision	98,36%
Recall	93,75%
F1-Score	96,00%
ROC-AUC	99,48%

Based on Table 5, it can be seen that the Random Forest algorithm achieved an accuracy of 97.50%. This value shows that as many as 97.50% of the test data was successfully classified correctly by the model. A precision value of 98.36% indicates that almost all of the positive predictions produced by the model correspond to the actual conditions. Meanwhile, a recall value of 93.75% indicates that the model is able to recognize most of the data that falls into the positive class.

An F1-score of 96.00% indicates a good balance between precision and recall. In addition, the ROC-AUC value of 99.48% indicates that the model has a strong ability to distinguish between the Fertile and Infertile classes.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

Overall, the results indicate that Random Forest provides a robust and stable classification model for Overall, the results indicate that Random Forest provides a robust and stable classification model for predicting male fertility status.

3.3 Random Forest Confusion Matrix Analysis

To find out the distribution of prediction results in more detail, the Confusion Matrix shown in Figure 2 is used.

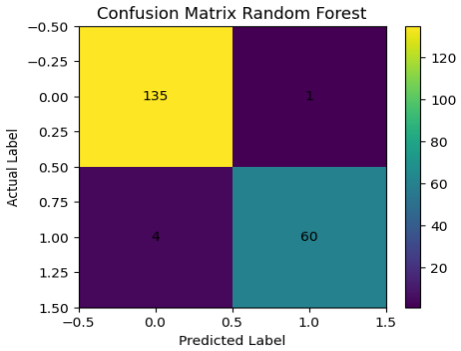

Figure 2. Confusion Matrix Random Forest

Based on the Confusion Matrix, the following results were obtained:

Table 6. Confusion Matrix Random Forest

Actual	Predicted Fertile	Predicted Infertile
Fertile	135	1
Infertile	4	60

Based on the confusion matrix results, the Random Forest model correctly classified 135 fertile samples as Fertile and 60 infertile samples as Infertile. However, one fertile sample was incorrectly classified as Infertile, while four infertile samples were incorrectly predicted as Fertile. These four misclassified infertile samples represent false-negative cases, which require particular attention in the context of male fertility assessment because infertile individuals may be incorrectly identified as fertile.

From a clinical perspective, the false negative cases are more critical than the false positive cases because infertile patients are incorrectly predicted as fertile. Such misclassification may delay further diagnostic examinations and appropriate fertility treatment. In this study, Random Forest produced four false negative cases out of 200 testing samples, indicating that although the overall classification performance was high, minimizing false negative predictions remains an important consideration for clinical decision-support systems.

The total classification errors made by the model were 5 out of a total of 200 test data points. The relatively small number of errors indicates that Random Forest provides reliable classification performance on the semen analysis dataset for semen analysis data.

3.4 Results of the Implementation of the XGBoost Algorithm

The second algorithm used in this study is Extreme Gradient Boosting (XGBoost). This algorithm is a development of the Gradient Boosting method, which works by building models gradually to correct the prediction errors of previous models. In addition, XGBoost is equipped with a regularization mechanism that helps reduce the risk of overfitting.

After the training process is complete, the model is used to make predictions on the test data. The results of the evaluation are shown in Table 7.

Table 7. XGBoost Evaluation Results

Metric	Value
Accuracy	99,00%
Precision	100,00%
Recall	96,88%
F1-Score	98,41%
ROC-AUC	99,86%

Based on Table 7, it is known that the XGBoost algorithm achieved an accuracy of 99.00%. This value is higher than that of Random Forest, which obtained an accuracy of 97.50%.

The precision value reaches 100.00%, which indicates that all positive predictions given by the model correspond to the actual conditions without generating positive errors. A recall value of 96.88% indicates that almost all positive data is successfully recognized by the model.

Copyright © 2026 **Author**, Page 544
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

An F1-score of 98.41% indicates a good balance between precision and recall. In addition, the ROC-AUC value of 99.86% indicates the model's high discriminative ability to distinguish between fertility and infertility classes.

Although XGBoost achieved an accuracy of 99.00%, these results should be interpreted with caution. The excellent predictive performance may be influenced by the strong discriminative characteristics of the selected secondary dataset and the close relationship between semen analysis parameters and fertility status. Therefore, the reported performance should not be interpreted as evidence of perfect real-world clinical performance. External validation using larger multicenter clinical datasets is required before the proposed model can be considered for routine clinical application.

3.5 XGBoost Confusion Matrix Analysis

The results of the Confusion Matrix for the XGBoost algorithm are shown in Figure 3.

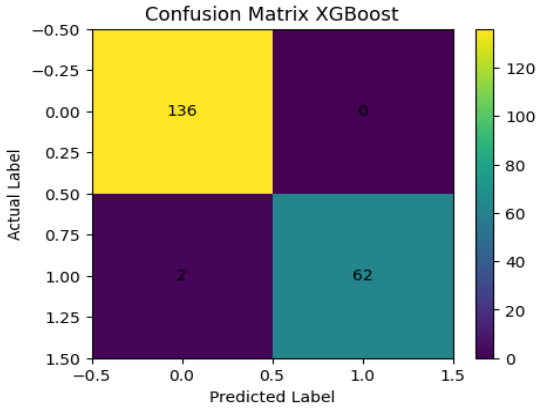

Figure 3. Confusion Matrix XGBoost

Based on the test results, the following data were obtained:

Table 8. Confusion Matrix XGBoost

Actual	Predicted Fertile	Predicted Infertile
Fertile	136	0
Infertile	2	62

Based on the confusion matrix results, the XGBoost model correctly classified 136 fertile samples and 62 infertile samples. No fertile samples were incorrectly classified as Infertile, while two infertile samples were misclassified as Fertile. These two misclassified infertile samples represent false-negative cases. Although the number of false-negative predictions was relatively small, such errors remain clinically important because infertile individuals may be incorrectly identified as fertile, potentially delaying further diagnostic evaluation or appropriate clinical assessment.

The XGBoost model generated only two false negative cases, which is lower than Random Forest. This result indicates that XGBoost has a better ability to correctly identify infertile patients. Nevertheless, even a small number of false-negative predictions should be carefully considered in clinical practice because affected patients may not receive timely medical evaluation or fertility treatment. Therefore, the proposed model should be regarded as a decision-support tool that assists clinicians rather than replacing expert diagnosis.

Out of a total of 200 test data points, only 2 data points experienced misclassification. These results indicate that the XGBoost model demonstrates good predictive performance on the testing dataset used in this study.

3.6 Random Forest and XGBoost Performance Comparison

A comparison of the performance of the two algorithms is shown in Table 9.

Table 9. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	97,50%	98,36%	93,75%	96,00%	99,48%
XGBoost	99,00%	100,00%	96,88%	98,41%	99,86%

Based on Table 9, it can be seen that XGBoost scored higher on all evaluation metrics than Random Forest. The difference in accuracy between the two algorithms reaches 1.50%. Although it may seem small, the improvement shows that XGBoost is capable of more precise classification of the test data. In addition, the precision XGBoost value reaches 100%, which indicates the absence of positive prediction errors.

The model performance comparison graph in Figure 4 also shows that all of the XGBoost evaluation metrics are above those of Random Forest.

Copyright © 2026 **Author**, Page 545
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

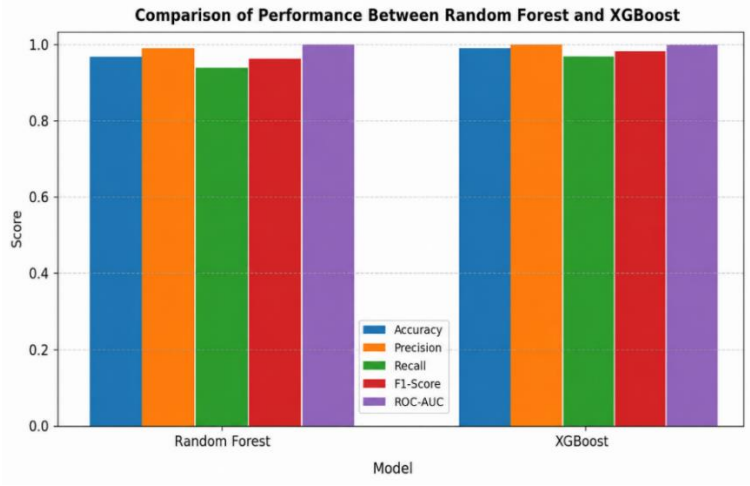

Figure 4. Random Forest and XGBoost Performance Comparison Chart

These results show that XGBoost's boosting mechanism is able to improve the quality of predictions by correcting errors in each iteration of the training. Therefore, XGBoost demonstrated better predictive performance in modeling the relationships among semen analysis variables under the experimental settings used in this study.

3.7 Analysis of Feature Importance Random Forest

Feature importance analysis was carried out to find out the variables that had the most influence on the prediction of male fertility status.

Table 10. Feature Importance Random Forest

Variable	Importance
Total_Motility_pct	0,318034
Vitality_pct	0,214972
Progressive_Motility_pct	0,145048
Morphology_Normal_pct	0,090088
Semen_Volume_ml	0,08784
Sperm_Concentration_Million_ml	0,06402
BMI	0,028506
Age	0,022613
pH	0,021768
Smoking	0,003954
Alcohol	0,003158

Based on Table 10, it is known that the variable that has the greatest influence is Total Motility with an importance value of 0.318034. These results show that sperm motility is the main factor that determines male fertility status.

The next important variables are Vitality and Progressive Motility. Both parameters are directly related to the ability of sperm to survive and move towards the egg, so that it has a major contribution to the fertilization process.

3.8 Analysis of Feature Importance XGBoost

Table 11. Feature Importance XGBoost

Variable	Importance
Total_Motility_pct	0,312812
Vitality_pct	0,193292
Progressive_Motility_pct	0,128273
Morphology_Normal_pct	0,101786
Semen_Volume_ml	0,092936
Sperm_Concentration_Million_ml	0,075588
BMI	0,0204
Smoking	0,020375
Age	0,019782
Alcohol	0,018048
pH	0,016707

Copyright © 2026 **Author**, Page 546
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

XGBoost's results show almost the same pattern as Random Forest. The variables Total Motility, Vitality, Progressive Motility, and Normal Morphology are the dominant factors in determining male fertility status. The consistency of results between the two algorithms showed that the parameters of the semen laboratory had a much greater influence than lifestyle factors such as smoking and alcohol consumption in this study dataset.

3.9 Discussion

Based on the experimental results, both Random Forest and XGBoost demonstrated strong predictive performance in classifying male fertility status based on semen analysis parameters. Random Forest achieved an accuracy of 97.50%, whereas XGBoost achieved an accuracy of 99.00%. The higher performance obtained by XGBoost under the experimental settings of this study indicates that the gradient boosting mechanism may capture complex and nonlinear relationships among semen analysis variables more effectively than the bagging-based mechanism employed by Random Forest. Nevertheless, the performance difference between the two algorithms should be interpreted within the characteristics of the secondary dataset used in this study.

The findings of this study are consistent with previous machine learning research in male reproductive health. Mehrjerd et al. [11] compared several ensemble learning algorithms, including Random Forest and XGBoost, for evaluating sperm quality in relation to clinical pregnancy outcomes. Their study reported that Random Forest achieved the best performance, with an average accuracy of approximately 0.72 and an AUC of 0.80. In contrast, the present study found that XGBoost achieved higher predictive performance than Random Forest. This difference may be associated with the prediction targets and predictor variables used. Mehrjerd et al. focused on clinical pregnancy success in assisted reproductive techniques, whereas the present study directly classified male fertility status using routine semen analysis parameters. Therefore, the more direct relationship between semen characteristics and fertility status may have contributed to the higher classification performance observed in this study.

The results also extend the findings reported by Huang et al. [12], who applied machine learning models to identify risk factors affecting sperm count using general health screening indicators. Their study demonstrated the potential of machine learning for reproductive health prediction; however, the predictor variables were primarily general health indicators. In comparison, the present study used semen-specific laboratory parameters, including Total Motility, Progressive Motility, Morphology, Sperm Concentration, and Vitality. The high predictive performance obtained in this study suggests that laboratory semen parameters may provide stronger discriminative information for fertility status classification than indirect general health indicators.

Furthermore, GhoshRoy et al. [13] employed XGBoost with SMOTE and explainable artificial intelligence techniques to predict male fertility using demographic and lifestyle-related variables. Their findings demonstrated the effectiveness of XGBoost for male fertility prediction. The present results support the predictive capability of XGBoost but differ in the type of input variables used. While GhoshRoy et al. emphasized lifestyle-related predictors, this study identified Total Motility, Vitality, Progressive Motility, and Normal Morphology as the dominant predictive features. This comparison indicates that XGBoost can effectively model different dimensions of male fertility data, while semen laboratory parameters may provide more direct information regarding male reproductive status.

The findings are also relevant to the study by Mohammadi et al. [14], which compared several machine learning algorithms, including Random Forest and XGBoost, for analyzing occupational exposures and male fertility indicators. Their study reported an AUC of approximately 0.99 for XGBoost and Random Forest, demonstrating strong discrimination performance. Similarly, the present study obtained ROC-AUC values of 99.48% for Random Forest and 99.86% for XGBoost. However, Mohammadi et al. primarily investigated occupational and environmental exposure variables, whereas this study focused on routine semen analysis parameters. The similarity in discriminative performance suggests that ensemble learning algorithms have substantial potential for analyzing male reproductive health data, although the predictors and clinical objectives differ between studies.

In addition, Chen et al. [15] demonstrated the effectiveness of XGBoost in predicting cumulative live birth outcomes among patients undergoing IVF/ICSI procedures. Although their prediction target was treatment outcome rather than male fertility status, their findings support the ability of XGBoost to model complex reproductive health data. The present study extends this evidence by directly comparing XGBoost with Random Forest under identical experimental settings and standardized evaluation metrics. Therefore, the contribution of this study lies not only in achieving high predictive performance but also in providing a direct empirical comparison of two ensemble learning algorithms using the same semen analysis dataset.

Feature importance analysis showed consistent patterns across both algorithms. Total Motility was identified as the most influential variable, followed by Vitality, Progressive Motility, and Normal Morphology. These findings are consistent with the semen quality indicators described in the WHO laboratory manual [1] and are also related to the findings of Aykaç et al. [3], who emphasized the importance of semen characteristics in evaluating male reproductive conditions. The consistency between the machine learning feature importance results and established andrological indicators suggests that the models captured clinically relevant relationships rather than relying solely on demographic or lifestyle variables.

Despite the high predictive performance, the 99.00% accuracy achieved by XGBoost should be interpreted cautiously. Biological and clinical datasets generally contain substantial variability and measurement uncertainty. The strong performance observed in this study may be influenced by the characteristics of the secondary dataset and the close relationship between the predictor variables and the Fertility_Status target. Moreover, the false negative analysis

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

showed that Random Forest misclassified four infertile samples as Fertile, whereas XGBoost produced two false negative cases. From a clinical perspective, these errors remain important because an infertile individual incorrectly classified as Fertile may experience delayed diagnostic evaluation or specialist consultation. Therefore, model performance should not be assessed solely based on overall accuracy but should also consider recall and false negative cases.

Overall, the findings of this study indicate that XGBoost achieved higher predictive performance than Random Forest under the experimental conditions used. Compared with previous studies that focused on pregnancy outcomes, lifestyle factors, occupational exposures, or general health indicators [11]–[15], this study specifically evaluates male fertility status using comprehensive routine semen analysis parameters. The direct comparison of Random Forest and XGBoost using identical data splitting, preprocessing procedures, and evaluation metrics provides additional empirical evidence regarding the application of ensemble learning in andrology. Nevertheless, external validation using independent multicenter clinical datasets remains necessary before the proposed model can be integrated into routine clinical practice.

4. CONCLUSION

This study demonstrates that machine learning techniques can effectively support the prediction of male fertility status using routine semen analysis parameters. The comparative evaluation indicates that both Random Forest and XGBoost are capable of learning meaningful patterns from laboratory semen examination data, making them suitable as computational approaches for supporting fertility assessment. Rather than replacing conventional laboratory examinations, the proposed models are intended to assist clinicians by providing objective and consistent predictions that may improve the efficiency of the initial evaluation process. Among the evaluated models, XGBoost achieved higher predictive performance than Random Forest under the experimental settings used in this study. In addition, feature importance analysis consistently identified Total Motility, Vitality, Progressive Motility, Normal Morphology, Semen Volume, and Sperm Concentration as the most influential predictors of male fertility status. These findings are consistent with established andrological knowledge, indicating that the developed models are capable of producing clinically meaningful predictions while maintaining interpretability. The main contribution of this study is the provision of comparative empirical evidence on the performance of Random Forest and XGBoost for direct male fertility status prediction using comprehensive semen analysis parameters under identical experimental conditions. In contrast to previous studies that predominantly investigated pregnancy outcomes, lifestyle factors, occupational exposures, or general health indicators, this study focuses specifically on routine semen examination variables and demonstrates the relative predictive performance of two widely used ensemble learning algorithms. Furthermore, the consistent identification of Total Motility, Vitality, and Progressive Motility as dominant predictors provides additional data-driven evidence regarding the importance of these parameters in male fertility assessment. From a practical perspective, the proposed prediction model can be integrated into the routine andrology examination workflow as a clinical decision-support tool. After routine semen analysis is completed, the laboratory parameters can be entered into the prediction system to generate an estimated fertility status. The prediction results may assist andrologists in identifying patients who require further diagnostic evaluation, additional laboratory examinations, or specialist consultation. Nevertheless, the proposed model should not be considered a replacement for clinical judgment. Instead, it is intended to complement the expertise of healthcare professionals by providing additional objective information during the decision-making process. Although the proposed model demonstrated promising predictive performance, several limitations should be acknowledged. The study utilized a secondary public dataset and therefore requires external validation using independent multicenter clinical datasets before being implemented in routine healthcare settings. Future research should also investigate additional clinical variables, including hormonal profiles, genetic factors, and medical history, as well as compare other machine learning algorithms. Furthermore, integrating Explainable Artificial Intelligence (XAI) techniques such as SHAP may improve model transparency and increase clinicians' confidence in the prediction results.

REFERENCES

[1] World Health Organization, “WHO Laboratory Manual for the Examination and Processing of Human Semen,” 6th ed., H. R. P. World Health Organization, Ed., Geneva: World Health Organization, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240030787

[2] S. Dehghan, R. Rabiei, H. Choobineh, K. Maghooli, M. Nazari, and M. Vahidi-Asl, “Comparative study of machine learning approaches integrated with genetic algorithm for IVF success prediction,” *PLoS One*, vol. 19, no. 10 October, pp. 1–19, 2024, doi: 10.1371/journal.pone.0310829.

[3] A. Aykaç, C. Kaya, Ö. Çelik, M. E. Aydın, and M. Sungur, “The prediction of semen quality based on lifestyle behaviours by the machine learning based models,” *Reprod. Biol. Endocrinol.*, vol. 22, p. 112, 2024, doi: 10.1186/s12958-024-01268-w.

[4] S. S. Cao, X. M. Liu, B. T. Song, and Y. Y. Hu, “Interpretable machine learning models for predicting clinical pregnancies associated with surgical sperm retrieval from testes of different etiologies: a retrospective study,” *BMC Urol.*, vol. 24, no. 1, pp. 1–12, 2024, doi: 10.1186/s12894-024-01537-1.

[5] Gyorgy J. Simon & Constantin Aliferis, “Artificial Intelligence and Machine Learning in Health Care and Medical Sciences,”

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 538−549
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10264

Springer, 2024, p. 810. [Online]. Available: https://link.springer.com/book/10.1007/978-3-031-39355-6
[6] A. Alanazi, "Using machine learning for healthcare challenges and opportunities," *Informatics Med. Unlocked*, vol. 30, p. 100924, 2022, doi: 10.1016/j.imu.2022.100924.
[7] L. Lu, Y. Qian, Y. Dong, H. Su, Y. Deng, and L. Lu, "A systematic study of the performance of machine learning models on analyzing the association between semen quality and environmental pollutants," *Front. Phys.*, vol. 11, p. 1259273, 2023, doi: 10.3389/fphy.2023.1259273.
[8] I. D. Mienye, Y. Sun, and S. Member, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.
[9] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.
[10] M. Niazkar *et al.*, "Applications of XGBoost in water resources engineering : A systematic literature review (Dec 2018 – May 2023)," *Environ. Model. Softw.*, vol. 174, p. 105971, 2024, doi: 10.1016/j.envsoft.2024.105971.
[11] A. Mehrjerd, T. Dehghani, M. Jajroudi, S. Eslami, H. Rezaei, and N. K. Ghaebi, "Ensemble machine learning models for sperm quality evaluation concerning success rate of clinical pregnancy in assisted reproductive techniques," *Sci. Rep.*, vol. 14, no. 1, pp. 1–10, 2024, doi: 10.1038/s41598-024-73326-7.
[12] H. Huang, S. Hsieh, M. Chen, M. Jhou, and T. Liu, "Machine Learning Predictive Models for Evaluating Risk Factors Affecting Sperm Count : Predictions Based on Health Screening Indicators," *J. Clin. Med.*, vol. 12, no. 3, p. 1220, 2023, doi: 10.3390/jcm12031220.
[13] D. Ghoshroy, P. A. Alvi, and K. C. Santosh, "Explainable AI to Predict Male Fertility Using Extreme Gradient Boosting Algorithm with SMOTE," *Electronics*, vol. 12, no. 1, p. 15, 2023, doi: 10.3390/electronics12010015.
[14] H. Mohammadi, S. Khoddam, and F. Golbabaei, "Analyzing the impact of occupational exposures on male fertility indicators : A machine learning approach," *Reprod. Toxicol.*, vol. 136, p. 108959, 2025, doi: 10.1016/j.reprotox.2025.108959.
[15] Z. Chen, D. Zhang, J. Zhen, Z. Sun, Q. Yu, and Y. Yin, "Predicting cumulative live birth rate for patients undergoing in vitro fertilization (IVF)/intracytoplasmic sperm injection (ICSI) for tubal and male infertility: a machine learning approach using XGBoost," *Chin. Med. J. (Engl).*, vol. 135, no. 8, pp. 997–999, 2022, doi: 10.1097/CM9.0000000000001874.
[16] J. Kim, "The research gap in evaluating community-based mental health interventions in Korea : A comparative analysis with the United Kingdom," *Asian J. Psychiatr.*, vol. 103, p. 104348, 2025, doi: 10.1016/j.ajp.2024.104348.
[17] I. Chouvarda *et al.*, "Differences in technical and clinical perspectives on AI validation in cancer imaging : mind the gap !," *Eur. Radiol. Exp.*, vol. 9, no. 7, 2025, doi: 10.1186/s41747-024-00543-0.
[18] A. Tawakuli, B. Havers, V. Gulisano, D. Kaiser, and T. Engel, "Survey : Time-series data preprocessing : A survey and an empirical analysis," *J. Eng. Res.*, vol. 13, no. 2, pp. 674–711, 2025, doi: 10.1016/j.jer.2024.02.018.
[19] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection : A review with examples from ecology," *Ecol. Monogr.*, vol. 93, no. 1, pp. 1–24, 2023, doi: 10.1002/ecm.1557.
[20] W. Hong, X. Zhou, S. Jin, Y. Lu, J. Pan, and Q. Lin, "A Comparison of XGBoost , Random Forest , and Nomograph for the Prediction of Disease Severity in Patients With COVID-19 Pneumonia : Implications of Cytokine and Immune Cell Pro fi le," *Front. Cell. Infect. Microbiol.*, vol. 12, no. April, pp. 1–13, 2022, doi: 10.3389/fcimb.2022.819267.

This Journal is licensed under a Creative Commons Attribution 4.0 International License