CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 9(2) (2024), Pages 352-367
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Hybrid Methods Random Forest and FOX-Inspired Optimization Algorithm for Selecting Features in Cervical Cancer Data

Afidatul Masbakhah, Umu Sa'adah*, Mohamad Muslikh

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Brawijaya University, Malang, Indonesia

Email: u.saadah@ub.ac.id

ABSTRACT

Cervical cancer remains one of the leading causes of mortality among women worldwide, necessitating effective early detection methods. This study aims to improve cervical cancer prediction through the integration of Random Forest (RF) and FOX-Inspired Optimization (FOX) for feature selection, addressing class imbalance with SMOTE-NC and refining model accuracy. he research utilized cervical cancer patient data from the UCI repository, applying data preprocessing techniques like handling missing values, normalization, and SMOTE-NC for balancing classes. The RF-FOX hybrid method was implemented to select key features, followed by classification using the RF model. The RF-FOX model demonstrated improved classification performance, with accuracy exceeding 95% across diagnostic categories, are Hinselmann, Schiller, Cytology, and Biopsy. Feature selection focused on significant factors like the number of pregnancies, age, and hormonal contraceptive use, enhancing the model's precision, recall, and AUC values. The RF-FOX approach offers a robust tool for early cervical cancer detection, reducing misclassification and computational complexity. Its successful application highlights its potential for broader use in healthcare diagnostics, paving the way for future hybrid model adaptations to other medical conditions.

Keywords: feature selection; cervical cancer; Random Forest; FOX-inspired optimization

INTRODUCTION

Cervical cancer is one of the most significant global health challenges and remains a leading cause of death among women worldwide. According to the World Health Organization (WHO), cervical cancer accounts for approximately 604,000 new cases and 324,000 deaths annually, ranking as the fourth most common cause of cancer-related mortality among women [1], [2]. This alarming prevalence underscores the urgency of effective prevention and early detection strategies to reduce morbidity and mortality rates associated with the disease. Early detection plays a critical role in preventing the progression of cervical cancer, especially since early stages are often asymptomatic [3]. Routine Pap smear and Human Papillomavirus (HPV) tests have proven to be effective screening methods, enabling early intervention by identifying abnormal cervical cells changes [4]. HPV, particularly oncogenic strains like types 16 and 18, is responsible for nearly 99% of cervical cancer cases. While many HPV infections resolve spontaneously, persistent infections can lead to cancerous developments. Additionally, other risk factors,

Submitted: September 12, 2024 Reviewed: September 23, 2024 Accepted: October 29, 2024
DOI: http://dx.doi.org/10.18860/ca.v9i2.29582

such as smoking, prolonged use of oral contraceptives, and weakened immune systems, contribute to the onset of cervical cancer [5]. WHO emphasizes that HPV vaccination and enhanced screening programs are essential, especially in developing countries, to control the spread of cervical cancer [2], [6].

The integration of machine learning algorithms in healthcare has demonstrated considerable potential in improving the early detection of cervical cancer. Predictive modeling is used to analyze risk factors and enhance the accuracy of diagnosis. However, the imbalanced distribution of data classes remains a significant challenge in building effective predictive models for cervical cancer. For instance, Mudawi and Alazeb's research showed that the Random Forest (RF) algorithm has superior predictive capabilities compared to other algorithms like Support Vector Machine (SVM) [7]. On the other hand, Abdoh’s study proposed a hybrid approach combining RF with the *Synthetic Minority Over-sampling Technique* (SMOTE), effectively increasing prediction accuracy by balancing the imbalanced target classes [8]. Another critical aspect of improving model performance is feature selection, which involves identifying the most relevant attributes in a dataset. Nithya’s study emphasized the importance of feature selection in enhancing both classification accuracy and processing speed, making it an essential component of cervical cancer detection models [9]. By focusing on core risk factors such as HPV infection history, smoking habits, and contraceptive use, predictive models become more efficient [10]. In recent years, Swarm Intelligence (SI) algorithms, inspired by natural collective behavior, have emerged as effective methods for feature selection [11]. SI algorithms, like the FOX-Inspired Optimization Algorithm (FOX), are particularly promising in the context of hyperparameter optimization and feature selection, showing superior results compared to traditional algorithms like Grey Wolf Optimization (GWO) and Particle Swarm Optimization (PSO) [12],[13],[14]. FOX leverages elements from the Fox Hunting Algorithm (FHA) and Red Fox Optimization (RFO), demonstrating better accuracy and efficiency in detecting relevant features in cervical cancer prediction models [15], [16]. By incorporating FOX-based feature selection, models can identify significant risk factors more accurately and optimize parameters, leading to improved prediction performance.

This study has several critical differences compared to previous research. While earlier studies [8] focused on using RF and oversampling techniques like SMOTE to handle data imbalances, cervical cancer data contains categorical data, requiring the use of the SMOTE-NC variant to address class imbalances. Moreover, few studies have combined RF with SI-based optimization algorithms like FOX for feature selection. Some studies employing SI algorithms, such as *Particle Swarm Optimization* (PSO) or *Grey Wolf Optimization* (GWO), have shown improvements in prediction accuracy and efficiency. However, this study proposes the combination of RF with FOX, which offers a better balance between exploration and exploitation in feature selection, thereby significantly improving prediction outcomes [17], [18]. The FOX algorithm, inspired by the hunting behavior of foxes (FHA and RFO), offers a unique approach to feature selection that can accurately identify the most relevant features and optimize model parameters which has been compared with other SI algorithms [15].

The main novelty of this research is the integration of the RF algorithm with FOX and SMOTE-NC as a hybrid method for more effective feature selection and handling of data imbalances in cervical cancer prediction. This approach is applied to cervical cancer detection, providing a new contribution to the development of more accurate and efficient predictive models. Therefore, the contributions of this study can be summarized as follows:

1. Identifying the performance of the RF model in predicting cervical cancer as a comparative method.
2. Selecting the most influential features for cervical cancer diagnosis using the FOX algorithm.
3. Evaluating the performance of the Hybrid RF-FOX method in predicting cervical cancer patients based on the selected features.
4. Comparing the model performance between the hybrid method and the basic RF method.

The proposed hybrid method is expected to address challenges such as data imbalances and redundant features, as well as enhance the accuracy of risk factor identification and early detection of cervical cancer. With this approach, the model can better identify significant risk factors, improve prediction efficiency and accuracy, and support efforts to refine predictive models in healthcare. This research aligns with current efforts to enhance cervical cancer detection through more effective screening and early intervention, potentially reducing cervical cancer mortality significantly.

METHODS

Preprocessing Data

1. The dataset used for this study, "Risk Factors Cervical Cancer" was retrieved from the UCI Machine Learning reporsitory. It contains risk factor data and various features that contribute to cervical cancer prediction.
2. Handling missing values [19]. features with fewer missing values, imputation was applied using the mean of non-missing values [20] calculated as:
$$Imputasi_{mean} = \frac{\Sigma_{i=1}^{n} \boldsymbol{x_i}}{n}.$$
where $\boldsymbol{x_i}$ is non missing values and $\boldsymbol{n}$ is sum of non-missing values.
3. Data normalization. To ensure that all features are on a similar scale, Min-Max normalization was applied to transform the data into the [0, 1] range, enhancing model performance and convergence [21].
4. Managing class imbalance. The target classes in the dataset were imbalanced, which could skew model performance [22], [23]. To address this, the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) was applied to balance the target class distribution, increasing the model's ability to generalize across all classes [8].

Random Forest

Random Forest (RF) is one of the machine learning algorithms used for the classification and regression process. RF is Breiman's proposal that combines predictors from classification trees. RF is defined by Breiman as "Random Forest is a classifier consisting of a structured collection of classification trees where random vectors $\{h(\boldsymbol{x}, \boldsymbol{\Theta_k}) | \forall k \in N\}$ where $\boldsymbol{\Theta_k}$ are independently and identically distributed and each tree assigns a unit vote for the most popular class on the input vector $\boldsymbol{x}$) "[24]. The final outcome of this system is determined by an ordinary majority vote and the function of the decision [25].

In the concept of RF, several decision trees are built by randomly selecting features and observations on their predictions. The formula for building a RF and determining the prediction is as follows [26].

1. Tree construction
   a) Select a random sample with substitution from the training dataset (bootstrap sampling).
   b) Choose a random number of features.
   c) Build a decision tree using the selected features, by splitting each node based on the selected criteria (In this context, impurity and entropy are chosen).
   d) Repeat the previous steps to build a number of decision trees.
2. Prediction
   Predictions are made by taking a majority of the votes from the predictions generated by each decision tree that has been built.

The most commonly used parameters in RF models are *n_estimators, max_features, min_samples_split, min_samples_leaf, max_depth, and criterion*. These parameters can affect the accuracy prediction values [27].

FOX-Inspired Algorithm

The FOX algorithm combines principles from two optimization methods: FHA and RFO [13],[14]. FOX maintains a balance between exploration (searching for new solutions) and exploitation (refining the current best solution), making it effective for finding optimal solutions rapidly [15]. The algorithm simulates fox behavior in finding prey, where foxes move within the search space to either explore new areas or exploit the best-known areas, based on a fitness function. In the context of feature selection, FOX identifies the most relevant attributes, eliminating less significant or redundant features to improve model accuracy and reduce computational overhead.

Feature Selection Process with the FOX algorithm

The FOX algorithm was applied to the training data for feature selection, focusing on identifying the most relevant features that significantly influence cervical cancer prediction. The FOX algorithm uses both exploration and exploitation strategies to search for the optimal set of features, thereby enhancing model efficiency and accuracy [28]. The steps involved are:

1. Initialization. The algorithm initializas a population of foxes, each representing a potential solution.
2. Fitness evaluation. The fitness of each fox (solution) is calculated to evaluate its effectiveness in selecting important features.
3. FOX movement. Foxes move within the search space, guided by global and local best solutions, to explore new areas or refine the best-known solutions.
4. Binary conversion. In the binary space, feature selestion is based on converting fox positions into binary vectors, where selected features are represented as “1” and unselected ones as “0”.
5. Selection of best features. The final selection includes features with the highest fitness scores, which are then used to reduce the dimensionality of the dataset.

Proposed Algorithm: RF-FOX

The proposed method integrates the RF model with FOX-based feature selection and SMOTE-NC resampling to improve the classification of cervical cancer risk factors. The steps are:

1. Prepocessing Data. The dataset is preprocessed by handling missing values, normalizing features, and managing class imbalances.
2. Feature selection with FOX. The FOX algorithm selects the most relevant features from the training data, optimizing the model for more accurate and efficient predictions.
3. Model training with RF. The RF model is trained using reduced feature set and evaluated based on accuracy, precision, recall and F1-score [29], [30].
4. Make Conclusion.

The hybrid RF-FOX model aims to deliver better predictive accuracy, faster processing, and improved early detection of cervical cancer risk factors by addressing data imbalances and optimizing feature selection.

Evaluation Metrix

The evaluation metric formulas used are as follows based on canfusion matrix's component, which are True Negative (TN), False Negative (FN), False Positive (FP) and True Negative (TN) [29], [30]. Table of confusion matrix shown as Table 1.

Table 1. Table of *Confusion Matrix*

Actual	Prediktion	
	Negative (0)	Positive (0)
Negative (0)	TN	FP
Positif (0)	FN	TP

The accuracy, precision, recall and F1-score are obtained using equation as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$

$$Presision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 - score = \frac{2(presision \times recall)}{presision + recall}.$$

Experimental Design

The data used in this study is in the form of secondary data of cervical cancer patients obtained from the UCI website which was accessed on September 09, 2024, the link for acces dataset as follow: https://archive.ics.uci.edu/dataset/383.

Table 2. Variables of Cervical Cancer Dataset

No	Variable	No	Variable	No	Variable
0	Age	12	STDs (number)	24	STDs:HPV
1	Number of sexual partners	13	STDs:condylomatosis	25	STDs:Number of diagnosis
2	First sexual intercourse	14	STDs:cervical condylomatosis	26	STDs: Time since first
3	Num of pregnancies	15	STDs:vaginal condylomatosis	27	STDs: Time since last
4	Smokes	16	SDTs:vulvo-perineal condylo.	28	Dx: Cancer
5	Smokes(years)	17	STDs:syphilis	29	Dx: CIN
6	Smokes (Packs/year)	18	STDs:pelvic inflammatory	30	Dx: HPV
7	Hormonal Contraceptives	19	STDs:genital herpes	31	Dx
8	Hormonal Contraceptives (years)	20	STDs:molluscum contagiosum	32	Hinselmann (Target)
9	IUD	21	STDs:AIDS	33	Schiller (Target)
10	IUD (years)	22	STDs:HIV	34	Citology (Target)
11	STDs	23	STDs:Hepatitis B	35	Biopsy (Target)

The cervical cancer data used in this study consisted of 858 rows and 36 columns its mean 32 features and 4 category target. Table 2 shown the variables of cervical cancer dataset. The cervical cancer dataset has missing values and imbalanced data. Missing values were identified in columns such as “STDs: Time since first diagnosis” and “STDs: Time since last diagnosis,” with over 90% missing data in these features (787 out of 868 records). These columns were dropped due to their high percentage of missing values [31]. After handling it, dataset can be split to 80% training data and 20% testing data. The research stages are shown as in Figure 1.

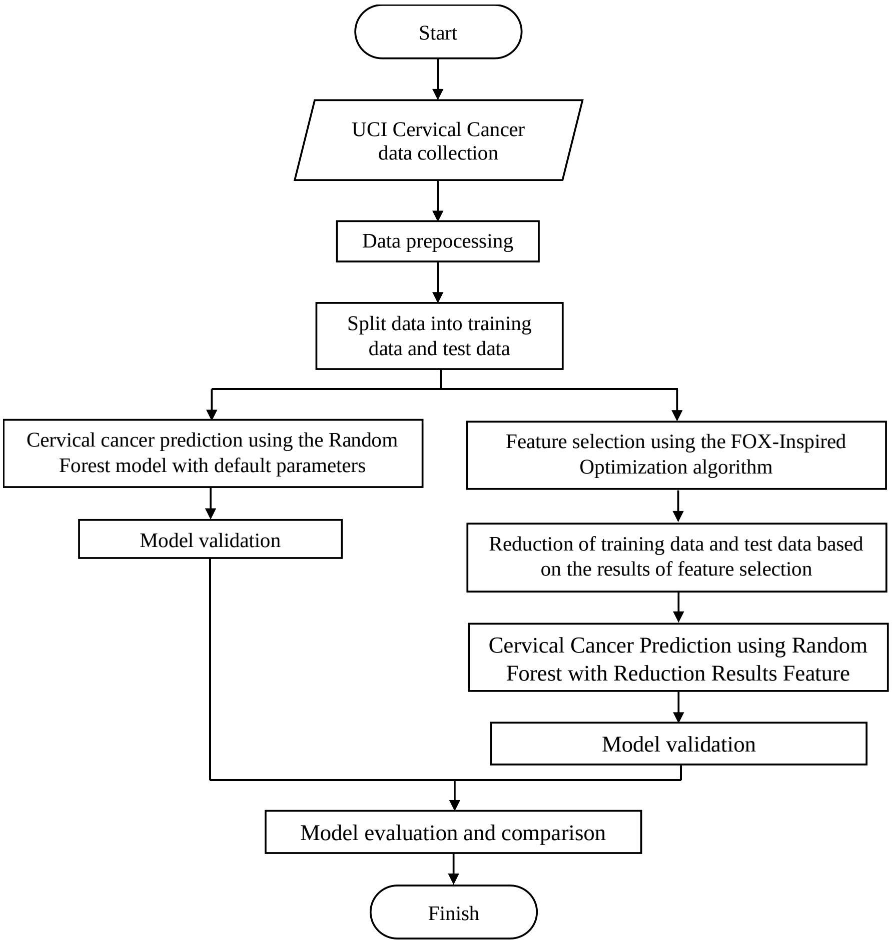

Start
UCI Cervical Cancer data collection
Data prepocessing
Split data into training data and test data
Cervical cancer prediction using the Random Forest model with default parameters
Model validation
Feature selection using the FOX-Inspired Optimization algorithm
Reduction of training data and test data based on the results of feature selection
Cervical Cancer Prediction using Random Forest with Reduction Results Feature
Model validation
Model evaluation and comparison
Finish

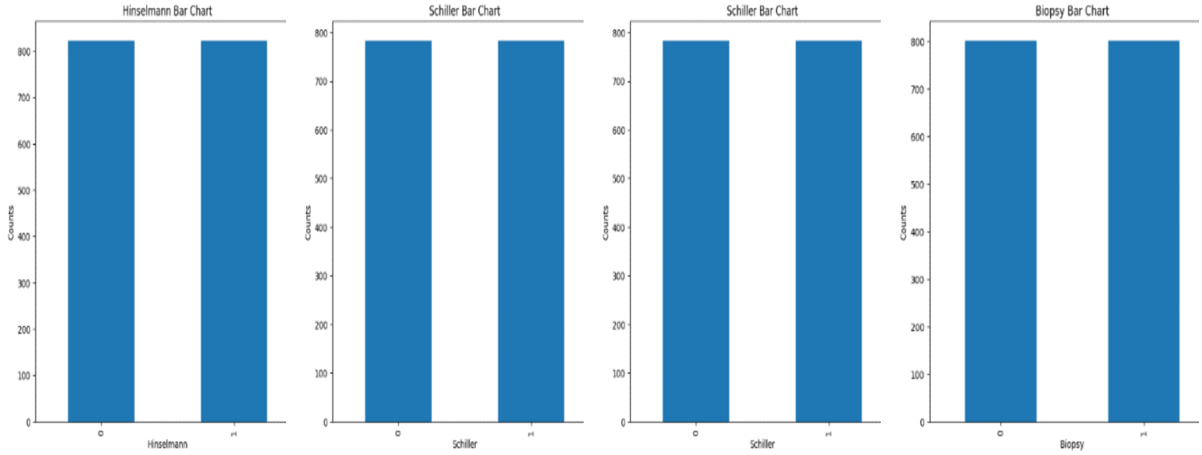
Figure 1. Research Flowchart

RESULTS AND DISCUSSION

Data Analysis

There are four categories of response variables, Hinselmann, Schiller, Citology and Biopsy with two classes, namely negative(0) and positive (1), while there are 32 predictor variables as features. There are 26 features that have missing values in the dataset. Handling missing values is done by imputing the mean value of each feature's data. However, the SDTs: Time since first diagnosis and SDTs: Time since last diagnosis features were removed because they had too many missing values, which was 787 out of 858. Furthermore, the target data is checked for data balance.

In the four target data class, members of each target showed imbalanced data. The comparison between the need to be examined and not in the biopsy category showed 93%:7%, the Citology category showed 95%:5%, the Schiller category showed 91%:9%, and the Hinselmann category showed 96%:4%. To respond to the imbalance of target data classes, resampling can be carried out with an oversampling method, namely the Synthetic Minority Over-sampling Technique for Numeric and Continuous (SMOTE-NC). The results of handling imbalanced classes are as shown in Figure 2.



(a) **(b)** **(c)** **(d)**

Figure 2. SMOTE-NC Results Target Data Class (a) Hinselmann Category (b) Schiller Category (c) Citology Category (d) Biopsy Category

Result of Random Forest Classification

After data analysis, this study predicted the data of cervical cancer patients using a RF model with default parameters and all features. The confusion matrix (Table 2) reveals that the model is able to classify samples correctly, with only a small number of samples being incorrectly predicted as false negatives. The low number of errors, in both false positives and false negatives, indicates that the modelhas been effective in detecting both positive and negative classes, with minimal errors inpredictions.

Table 2. Confusion Matrix of Test Data

		Predicted Values							
	Category	**Biopsy**		**Citology**		**Schiller**		**Hinselmann**	
	Class	**(1)**	**(0)**	**(1)**	**(0)**	**(1)**	**(0)**	**(1)**	**(0)**
Actual Values	**Positive (1)**	149	4	149	7	152	8	149	3
	Negative (0)	16	153	12	158	5	149	9	169

Based on table of confusion matrix, the results of the model evaluation on the test data can seen at Table 3. The RF model shows good performance with the average accuracy level of the four categories. The model was able to correctly classify 97% of the Biopsy category, 93% of the Citology category, 96% of the Schiller category and 98% of the Hinselmann category, of each sample in each category correctly. Overall, this model shows strong performance in classifying cervical cancer risk factors.

Table 3. Evaluation Metrix of RF Model

Category	Class	Split Data	Accuracy	Presision	Recall	F1-Score	AUC
Hinselmann	0	Train	0.98	0.97	0.98	0.98	0.98
		Test	0.95	0.93	0.98	0.96	0.97
	1	Train	0.98	0.98	0.97	0.98	0.98
		Test	0.95	0.98	0.92	0.95	0.97
Schiller	0	Train	0.96	0.94	0.98	0.96	0.96
		Test	0.90	0.87	0.94	0.90	0.95
	1	Train	0.96	0.98	0.93	0.96	0.96
		Test	0.90	0.94	0.86	0.90	0.95
Citology	0	Train	0.95	0.96	0.95	0.91	0.95
		Test	0.94	0.93	0.91	0.90	0.92
	1	Train	0.95	0.96	0.96	0.91	0.95
		Test	0.94	0.94	0.90	0.90	0.92
Biopsy	0	Train	0.96	0.96	0.97	0.96	0.97
		Test	0.93	0.90	0.97	0.94	0.96
	1	Train	0.96	0.97	0.95	0.96	0.97
		Test	0.94	0.97	0.91	0.94	0.96

Feature Selection Risk Factors Cervical Cancer with RF-FOX

a. Hinselmann Category

Classification analysis using feature selection with RF-FOX for the Hinselmann category, as shown in Table 4 and Figure 3, shows that the feature "3" (Num of pregnancies) has the greatest influence with an importance level of 0.204. This indicates that this feature has a significant contribution in detecting cervical cancer in the Hinselmann category.

Table 4. Result of Feature Ranking of Category Hinselmann

Feature	Rank	Feature	Rank
Num of Pregnancies	0.203736	Smokes	0.017757
Number of sexual	0.191636	STDs:HIV	0.016661
First sexual intercourse	0.184287	STDs:vaginal condylomatosis	0.015667
age	0.140594	STDs:genital herpes	0.011376
IUD (years)	0.048673	STDs:Hepatitis B	0.010180
Smokes (Packs/year)	0.038546	Dx: Cancer	0.008263
Smokes (years)	0.033218	Dx: CIN	0.003504
STDs: condylomatosis	0.024995	Dx: HPV	0.002603
SDTs:vulvo-perineal condylomatosis	0.024116	Dx	0.001348
STDs:syphilis	0.022840		

Based on Figure 3 and Table 4, shown that the features "1" (Number of sexual partners) and "2" (First sexual intercourse) followed with importance values of 0.192 and 0.184, respectively, indicating that they also have an important role in the classification. The "0" (*Age*) feature also has a considerable influence with an importance value of 0.141.

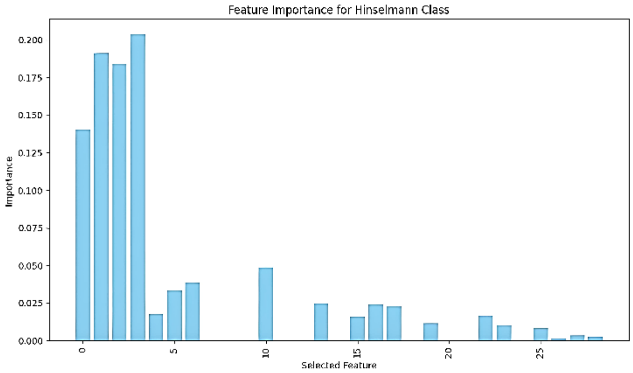


Figure 3. Plot of Feature Importance for Hinselmann Category

These features overall support the cancer detection process by providing essential information to the model. Meanwhile, other features, such as "10" (IUD per year) and "6" (smokes per pack per year) have lower levels of importance, suggesting that they have a weaker correlation with cervical cancer risk in this category. These features appear to have a smaller role in the context of cervical cancer diagnosis or prediction in the Hinselmann category.

On the other hand, some other features like "10" (IUD per year) and "6" (Smoke per packs per year) show lower importance values, suggesting that these features have a weaker correlation with cervical cancer risk in this category. This implies that these features play a smaller role in the context of cervical cancer diagnosis or prediction in the Hinselmann category. Meanwhile, features with a ranking value of 0 indicate that they do not contribute to the classification in this category, meaning they have no significant impact on model predictions. Therefore, excluding such features from the model could help improve efficiency by reducing computational complexity**.**

b. Schiller Category

The classification in the Schiller category uses feature selection with RF-FOX, as shown in Table 5 and Figure 4. Identifying the feature "1" (Number of sexual partners) as the most significant with a significance value of 0.192. This suggests that these factors have a strong association with cervical cancer detection in the context of schiller diagnosis. The "8" feature (Hormonal Contraceptives) also shows a great influence with a value of 0.187. The "3" (First sexual intercourse) and "0" (Age) features followed with important values of 0.168 and 0.161, respectively, indicating that the model relied on a combination of these factors for cervical cancer detection.

Table 5. The Result of Feature Selection Schiller Category with RF-FOX

Feature	Rank	Feature	Rank
Number of sexual	0.192363	STDs:condylomatosis	0.028529
Hormonal Contraceptives (years)	0.186831	STDs:vaginal condylomatosis	0.013977
Num of pregnancies	0.167784	STDs:genital herpes	0.011673
Age	0.161460	STDs:HIV	0.009968
Smokes (Packs/year)	0.053082	STDs:Hepatitis B	0.009767
Hormonal Contraceptives	0.046179	STDs:molluscum contagiosum	0.006453
STDs	0.037365	Dx	0.005368
STDs (number)	0.034732	Dx: CIN	0.002833
STDs: syphilis	0.029284	Dx: HPV	0.002350

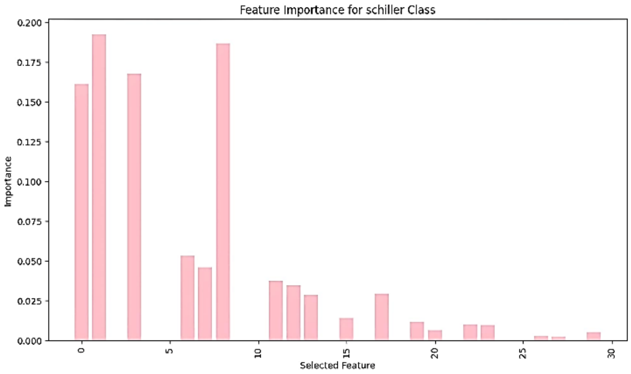


Figure 4. Plot of Feature Importance for Schiller Category

On the other hand, based on Figure 4, shown that features such as "6" (smoke per year) and "7" (hormonal contraceptives) which have lower importance values, may indicate a weaker correlation with cervical cancer risk in the cytology category. This means that some factors may be less relevant or have little effect in the context of cervical cancer prediction or diagnosis in the Schiller category.

c. Citology Category

The results of the Citology classification with feature selection using RF-FOX, as shown in Table 6 and Figure 5. Based on Table 6 and Figure 5, shown that the feature "8": Hormonal Contraceptives (years) has the highest importance level with 0.189, meaning that this factor has a stronger correlation with cervical cancer detection, compared to other features in the citology diagnosis category. Other features of high importance, such as those indexed "1": number of sexual partners, "2": first sexual intercourse, and "0": age, also represent risk factors or other important indicators as in the biopsy category.

Table 6. The Result of Feature Selection Citology Category with RF-FOX

Feature	Rank	Feature	Rank
Hormonal Contraceptives (years)	0.188049	Smokes (years)	0.029558
Number of sexual partners	0.164019	Smokes (Packs/year)	0.023630
First sexualintercourse	0.115908	STDs:syphilis	0.016773
Age	0.102562	STDs:Number of diagnosis	0.013898
Num of pregnancies	0.096582	STDs:HPV	0.012559
Hormonal Contraceptives	0.060272	STDs:vaginal condylomatosis	0.012334
STDs	0.054170	STDs:Hepatitis B	0.008515
STDs:condylomatosis	0.051071	Dx: HPV	0.004881
Smokes	0.040908	Dx	0.003411

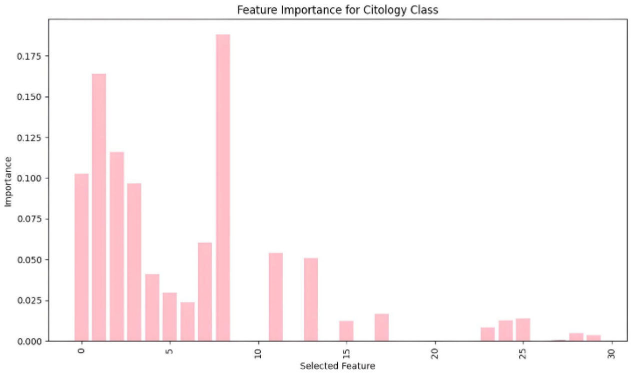


Figure 5. Plot of Feature Importance for Citology Category

In contrast, features with lower importance values, such as those indexed as "23" (STDs:Hepatitis B) and "24" (STDs:HPV), like as Figure 5, suggest a weaker or no direct correlation with cervical cancer risk in the cytology category. This implies that these features have minimal impact on the model's performance in detecting or predicting cervical cancer within this diagnostic context. Consequently, the model primarily relies on significant features like hormonal contraceptive use, age, and sexual behavior-related factors to enhance detection accuracy in cytology diagnosis. Meanwhile, features with a ranking value of 0 which can seen at Figure 5, indicate no contribution to the classification process in the Cytology category. These features lack a significant impact on prediction accuracy and do not enhance the model's detection capabilities. Excluding these features can improve model efficiency by reducing computational load without affecting performance.

d. Biopsy Category

The results of feature selection using RF-FOX in the biopsy classification show the features identified as the most important features. This means that these risk factors are the most influential in the development or detection of cervical cancer after the diagnosis of the biopsy category.

Table 7. The Result of Feature Selection Biopsy Category with RF-FOX

Feature	Rank	Feature	Rank
Num of pregnancies	0.214464	Smokes	0.018320
First sexualintercourse	0.164122	STDs:vaginal condylomatosis	0.017925
Number of sexual	0.151198	STDs:HPV	0.014131
Age	0.137007	STDs:molluscum contagiosum	0.013891
STDs (number)	0.070886	STDs:HIV	0.009940
IUD (years)	0.052413	STDs:vaginal condylomatosis	0.017925
Smokes(years)	0.032237	STDs:Hepatitis B	0.008336
Smokes (Packs/year)	0.029058	STDs:genital herpes	0.005254
IUD	0.026881	Dx: HPV	0.004965
Hormonal Contraceptives	0.025096	Dx	0.003876

Table 7 show that the feature with the index "3" (Num of pregnancies) has the highest importance level of around 0.21, so it can be said to be a major risk factor. This means that this factor has a strong correlation with the risk of cervical cancer. The high importance of this feature in the model suggests that it has a major influence in differentiating between positive and negative cases of cervical cancer.

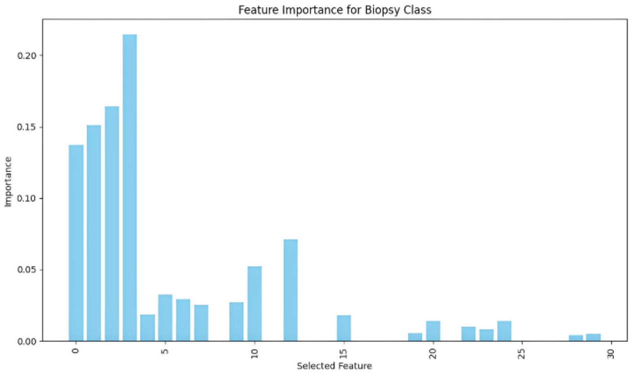


Figure 6. Plot of Feature Importance for Biopsy Category

Other features of considerable importance can be seen at Figure 6, such as those indexed "2": first sexual intercourse, "1": number of sexual partners and "0": age, also represent risk factors or other important indicators. This suggests that the model on biopsy diagnosis relies on a combination of Num of pregnancies risk factors to perform cancer detection. In contrast, other low-importance features likely represent factors that have little or no direct correlation with cervical cancer risk in this dataset. This means that the data may not have a significant role in the diagnosis or prediction of cervical cancer.

On the contrary, based on Figure 6, there are features with lower importance values, such as "25"(STDs:Number of diagnosis) and "27" (Dx:CIN), exhibit weaker correlations with cervical cancer risk in this context. This implies that these features contribute minimally to the classification process and have a limited impact on the model's performance. Additionally, features with a ranking value of 0 indicate no contribution to the classification in the biopsy category. These features have no significant influence on prediction accuracy, meaning they can be excluded to enhance model efficiency by reducing computational complexity without compromising accuracy.

Evaluation Model After Feature Selection

After the feature selection is carried out and the most influential features are known, the data is then reduced using the results of the feature selection and predicted again and then evaluated using metric evaluation. The evaluation of the metrics of all category after using the hybrid method has increased in each metric can see on Table 8. Table 8 shown that after applying feature selection, the Hinselmann Category shows strong performance, especially for Class 0 with high accuracy, precision, and recall on both training (0.99–1.00) and testing data (0.95).

Table 8. Evaluation Metrix After Feature Selection with RF-FOX

Category	Class	Split Data	Accuracy	Presision	Recall	F1-Score	AUC
Hinselmann	0	Train	0.99	0.99	1.00	0.99	1.00
		Test	0.96	0.96	0.98	0.96	0.99
	1	Train	0.99	1.00	0.99	0.00	1.00
		Test	0.98	0.98	0.95	0.96	0.99
Schiller	0	Train	0.99	0.99	1.00	0.99	1.00
		Test	0.96	0.97	0.95	0.96	0.98
	1	Train	0.99	1.00	0.99	0.99	1.00
		Test	0.96	0.95	0.97	0.96	0.98
Citology	0	Train	0.99	0.98	0.99	0.99	0.99
		Test	0.94	0.94	0.96	0.96	0.98
	1	Train	0.99	0.99	0.98	0.99	0.99
		Test	0.94	0.96	0.96	0.96	0.98
Biopsy	0	Train	0.99	0.99	1.00	0.99	1.00
		Test	0.97	0.96	0.97	0.98	0.99
	1	Train	0.99	1.00	0.99	0.99	1.00
		Test	0.97	0.97	0.96	0.98	0.99

For Class 1, the performance is slightly lower, with testing accuracy at 0.89 and recall at 0.85, but the model still maintains a high AUC of 0.98, indicating good classification ability. In the Schiller category, Class 0 achieves an accuracy of 0.98 on training data and 0.90 on testing, with balanced precision and recall. Class 1 shows a slight decline, with test accuracy at 0.87 and recall at 0.81, but maintains a high precision (0.92), indicating effective handling of false positives. The AUC remains high (0.99), reflecting the model's reliable classification. For the Citology category, Class 0 achieves good results with testing accuracy at 0.91 and F1-score at 0.90, suggesting effective generalization. Class 1 shows similar performance, with a test accuracy of 0.90 and balanced precision and recall, demonstrating robust detection across both classes. Beside that, the Biopsy category shows near-perfect performance for Class 0, with testing accuracy at 0.97 and an AUC of 1.00. For Class 1, accuracy drops to 0.87 with lower recall (0.68), indicating some challenges in detecting positive cases. However, high precision (0.97) and a strong AUC of 0.99 reflect good overall classification.

Comparison Evaluation Metrix of RF and RF-FOX

After knowing and evaluating the results of the analysis using the default RF method and RF-FOX for feature selection, it can be seen that the evaluation metrics show a difference in the increase in values across all metrics. This shows that feature selection plays an important role in selecting features that are influential in cervical cancer detection based on the rank of feature importance. A comparison of the analysis results can be seen in Table 9.

Table 9. Comparison Evaluation Metrix RF and RF-FOX

Category	Cls	Split Data	Accuracy		Presision		Recall		F1-Score		AUC	
			RF	RF-FOX	RF	RF-FOX	RF	RF-FOX	RF	RF-FOX	RF	RF-FOX
Hinselmann	0	Train	0.98	0.99	0.97	0.99	0.98	1.00	0.98	0.99	0.98	1.00
		Test	0.95	0.96	0.93	0.96	0.98	0.98	0.96	0.96	0.97	0.99
	1	Train	0.98	0.99	0.98	1.00	0.97	0.99	0.98	0.00	0.98	1.00
		Test	0.95	0.98	0.98	0.98	0.92	0.95	0.95	0.96	0.97	0.99
Schiller	0	Train	0.96	0.99	0.94	0.99	0.98	1.00	0.96	0.99	0.96	1.00
		Test	0.90	0.96	0.87	0.97	0.94	0.95	0.90	0.96	0.95	0.98
	1	Train	0.96	0.99	0.98	1.00	0.93	0.99	0.96	0.99	0.96	1.00
		Test	0.90	0.96	0.94	0.95	0.86	0.97	0.90	0.96	0.95	0.98
Citology	0	Train	0.95	0.99	0.96	0.98	0.95	0.99	0.91	0.99	0.95	0.99
		Test	0.94	0.94	0.93	0.94	0.91	0.96	0.90	0.96	0.92	0.98
	1	Train	0.95	0.99	0.96	0.99	0.96	0.98	0.91	0.99	0.95	0.99
		Test	0.94	0.94	0.94	0.96	0.90	0.96	0.90	0.96	0.92	0.98
Biopsy	0	Train	0.96	0.99	0.96	0.99	0.97	1.00	0.96	0.99	0.97	1.00
		Test	0.93	0.97	0.90	0.96	0.97	0.97	0.94	0.98	0.96	0.99
	1	Train	0.96	0.99	0.97	1.00	0.95	0.99	0.96	0.99	0.97	1.00
		Test	0.94	0.97	0.97	0.97	0.91	0.96	0.94	0.98	0.96	0.99

Based on Table 9, the results clearly demonstrate that the RF-FOX method consistently outperforms the default RF method in all categories and metrics, both in training and testing phases. The feature selection process by RF-FOX effectively enhances the model's performance by focusing on the most significant features, which is evident in the improvements in accuracy, precision, recall, F1-score, and AUC values across four diagnostic categories. The RF-FOX hybrid method, by focusing on feature selection, improves classification performance across all diagnostic categories. The increases in accuracy, precision, recall, F1-score, and AUC demonstrate that selecting the most relevant features significantly enhances the model's ability to detect cervical cancer. RF-FOX's superior performance is especially evident in recall and AUC values, highlighting its strength in reducing false negatives and achieving better class distinction.

Discussion

This study demonstrates the effectiveness of using the RF model for predicting cervical cancer cases, with significant enhancements when combined with the RF-FOX hybrid method for feature selection. The standard RF model initially showed strong classification performance, with average accuracy exceeding 93% across all diagnostic categories (Table 3). The confusion matrix (Table 2) confirmed minimal errors, indicating reliable detection of both positive and negative cases. The introduction of RF-FOX further improved metrics across all categories—Hinselmann, Schiller, Citology, and Biopsy—by focusing on the most influential features. Notably, accuracy, precision, recall, F1-score, and AUC all increased, with the Hinselmann and Biopsy categories achieving up to 99% accuracy and an AUC of 1.00 (Table 8). Key features like "number of pregnancies," "age," and "number of sexual partners" emerged as crucial risk factors in these categories, boosting detection accuracy. The comparison (Table 9) clearly shows that RF-FOX outperforms the standard RF model, not only by improving overall performance but also by reducing computational complexity through the exclusion of less relevant features. Higher recall and AUC values reflect the model's improved ability to detect true positive cases and reduce false negatives.

This study underscores the importance of targeted feature selection for enhancing cervical cancer detection. By concentrating on the most relevant features, the RF-FOX method achieves more accurate predictions, lowers the risk of misclassification, and supports early diagnosis-essential for effective treatment. The improved recall and AUC metrics suggest that RF-FOX has strong potential for real-world applications where minimizing false negatives is critical. The findings contribute to predictive modeling literature by showing the impact of feature selection in cervical cancer detection. The RF-FOX method not only refines the prediction process but also offers an efficient approach to identifying significant risk factors, paving the way for further research in applying hybrid models to other medical conditions.

CONCLUSIONS

Based on the results of research that has been carried out, This study demonstrates that the RF-FOX hybrid method, when applied to cervical cancer detection, significantly enhances classification performance compared to the default RF model. Key findings include: Improved accuracy and recall. The RF-FOX method showed higher accuracy, recall, and AUC values across all diagnostic categories (Hinselmann, Schiller, Citology, and Biopsy). The ability to accurately identify true positive cases and reduce false negatives was especially evident, underscoring the method's effectiveness in improving early detection of cervical cancer. Effectiveness of feature selection. By focusing on the most influential features, the RF-FOX hybrid model streamlined the classification process, reducing computational complexity while maintaining or improving predictive accuracy. Critical risk factors, such as "number of pregnancies," "age," and "hormonal contraceptive use," were consistently identified as key contributors to cervical cancer detection across categories. Broader implications. The findings emphasize the importance of targeted feature selection in medical diagnostics, suggesting that hybrid models like RF-FOX can be valuable tools in predictive healthcare. The increased precision and recall metrics indicate the potential of RF-FOX to be used effectively in real-world clinical settings to aid early diagnosis and timely treatment.

The RF-FOX model should be considered for implementation in clinical settings, given its strong performance in detecting cervical cancer risk factors, supporting early intervention and reducing morbidity. Healthcare practitioners should focus on key risk factors identified by the model, such as "number of pregnancies," "age," and "hormonal contraceptives," to enhance screening strategies. Future research should aim to integrate other algorithms with RF-FOX, optimize parameters, and expand data collection for diverse populations to improve model robustness. Regular evaluation and refinement are necessary to maintain accuracy and adaptability in real-world applications.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[2] A. Mathis, U. D. Smith, V. Crowther, T. Lee, and S. Suther, "An Epidemiological Study of Cervical Cancer Trends among Women with Human Immunodeficiency Virus," *Healthcare (Switzerland)*, vol. 12, no. 12, Jun. 2024, doi: 10.3390/healthcare12121178.

[3] WHO, "World Health Organization," Risk Factors Cervical Cancer. Accessed: Mar. 06, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[4] L. Rahangdale, C. Mungo, S. O'Connor, C. J. Chibwesha, and N. T. Brewer, "Human papillomavirus vaccination and cervical cancer risk," 2022, *BMJ Publishing Group*. doi: 10.1136/bmj-2022-070115.

[5] N. Y. Ozturk, S. Z. Hossain, M. Mackey, S. Adam, and P. Brennan, "HPV and Cervical Cancer Awareness and Screening Practices among Migrant Women: A Narrative Review," Apr. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/healthcare12070709.

[6] S. Choi, A. Ismail, G. Pappas-Gogos, and S. Boussios, "HPV and Cervical Cancer: A Review of Epidemiology and Screening Uptake in the UK," Feb. 01, 2023, *MDPI*. doi: 10.3390/pathogens12020298.

[7] N. Al Mudawi and A. Alazeb, "A Model for Predicting Cervical Cancer Using Machine Learning Algorithms," *Sensors*, vol. 22, no. 11, Jun. 2022, doi: 10.3390/s22114132.

[8] S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, pp. 59475–59485, 2018, doi: 10.1109/ACCESS.2018.2874063.

[9] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," *SN Appl Sci*, vol. 1, no. 6, Jun. 2019, doi: 10.1007/s42452-019-0645-7.

[10] C. P. Vandana and A. A. Chikkamannur, "Feature selection: An empirical study," *International Journal of Engineering Trends and Technology*, vol. 69, no. 2, pp. 165–170, Feb. 2021, doi: 10.14445/22315381/IJETT-V69I2P223.

[11] L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: A review," Sep. 01, 2018, *MDPI AG*. doi: 10.3390/app8091521.

[12] G. Senthilkumar *et al.*, "Incorporating Artificial Fish Swarm in Ensemble Classification Framework for Recurrence Prediction of Cervical Cancer," *IEEE Access*, vol. 9, pp. 83876–83886, 2021, doi: 10.1109/ACCESS.2021.3087022.

[13] M. Onay, "A New and Fast Optimization Algorithm: Fox Hunting Algorithm (FHA)," 2016.

[14] D. Połap and M. Woźniak, "Red fox optimization algorithm," *Expert Syst Appl*, vol. 166, Mar. 2021, doi: 10.1016/j.eswa.2020.114107.

[15] H. Mohammed and T. Rashid, "FOX: a FOX-inspired optimization algorithm," *Applied Intelligence*, vol. 53, pp. 1030–1050, 2023, doi: 10.1007/s10489-022-03533-0/Published.

[16] R. Sharma *et al.*, "Comparative performance analysis of binary variants of FOX optimization algorithm with half-quadratic ensemble ranking method for thyroid cancer detection," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-46865-8.

[17] N. K. Chauhan and K. Singh, “Performance Assessment of Machine Learning Classifiers Using Selective Feature Approaches for Cervical Cancer Detection,” *Wirel Pers Commun*, vol. 124, no. 3, pp. 2335–2366, Jun. 2022, doi: 10.1007/s11277-022-09467-7.

[18] R. Alizadehsani *et al.*, “Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020,” Jan. 01, 2021, *Elsevier Ltd*. doi: 10.1016/j.compbiomed.2020.104095.

[19] R. Ashtagi *et al.*, “Cervical Cancer Prediction Using Machine Learning,” 2024.

[20] R. Alsmariy, G. Healy, and H. Abdelhafez, “Predicting Cervical Cancer using Machine Learning Methods,” 2020. [Online]. Available: www.ijacsa.thesai.org

[21] Y.-Long. Jiang, T.-Ao. Tang, and Ye. Fan, *ICSICT-2018 : 2018 14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT) : proceedings :Oct. 31- Nov. 3, 2018, Qingdao, China*. IEEE, 2018.

[22] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, “Machine learning for assisting cervical cancer diagnosis: An ensemble approach,” *Future Generation Computer Systems*, vol. 106, pp. 199–205, May 2020, doi: 10.1016/j.future.2019.12.033.

[23] A. Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, “Evolutionary Machine Learning: A Survey,” Nov. 01, 2022, *Association for Computing Machinery*. doi: 10.1145/3467477.

[24] L. Breiman, “Random Forests,” 2001.

[25] A. Cutler, D. R. Cutler, and J. R. Stevens, “Random Forests,” in *Ensemble Machine Learning*, New York, NY: Springer New York, 2012, pp. 157–175. doi: 10.1007/978-1-4419-9326-7_5.

[26] P. Bhargav and K. Sashirekha, “A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms,” 2023.

[27] S. Gupta and M. K. Gupta, “Computational Prediction of Cervical Cancer Diagnosis Using Ensemble-Based Classification Algorithm,” *Computer Journal*, vol. 65, no. 6, pp. 1527–1539, Jun. 2021, doi: 10.1093/comjnl/bxaa198.

[28] B. Majhi and Prastavana, “A feature selection model using binary FOX optimization and v-shaped transfer function for network IDS,” *Peer Peer Netw Appl*, 2024, doi: 10.1007/s12083-024-01720-z.

[29] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[30] M. Liao, H. Wen, L. Yang, G. Wang, X. Xiang, and X. Liang, “Improving the model robustness of flood hazard mapping based on hyperparameter optimization of random forest,” *Expert Syst Appl*, vol. 241, May 2024, doi: 10.1016/j.eswa.2023.122682.