

Leveraging Statistical and Machine Learning Techniques to Enhance Clinical Decision Support for PID Diagnosis

Akinrotimi Akinyemi Omololu^{*1}, Atoyebi Jelili Olaniyi², Owolabi Olugbenga Olayinka³ And Omotosho Israel Oluwabusayo⁴

¹ Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria

² Department of Computer Engineering, Adeleke University, Ede, Osun State, Nigeria

³ Department of Electrical and Electronics Engineering, Adeleke University, Ede, Osun State, Nigeria

⁴ College of Business, Bowie State University, Maryland, USA

Corresponding Email: akinrotimiakinyemi@ieee.org

ARTICLE INFO

Article History:

Received: 30th June 2025

Revised: 20th December 2025

Accepted: 24th December 2025

Published: 30th June 2026

Keywords:

Chronic pelvic pain

Data-driven approaches

Early detection

Feature selection

Infertility

Misclassification errors

ABSTRACT

Pelvic Inflammatory Disease (PID) is an infection of the female reproductive organs, most frequently leading to problems such as infertility, chronic pelvic pain, and ectopic pregnancy. Due to its nonspecific symptoms, diagnosing PID remains problematic, often relying on clinical judgment and invasive procedures. This challenge highlights the need for more sensitive, non-invasive, and data-driven diagnostic tools to enhance early detection and treatment outcomes. As the symptoms are nonspecific, the diagnosis of PID remains problematic and needs the development of more sensitive and less invasive diagnostic tools. This study investigates combining Fisher's Exact Test for statistical feature selection with machine learning algorithms, including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), to improve the diagnosis of Pelvic Inflammatory Disease (PID). The study analyses the clinical profiles of 200 patients diagnosed with PID and uses Fisher's Exact Test to identify key features associated with the condition's severity. These identified features were then used to train and evaluate the machine learning models. Among the algorithms used, the ANN performed best, achieving 91% accuracy, 90% precision, 92% recall, and 91% F1-score, demonstrating superior accuracy in predicting PID. The results show that the use of statistical and machine learning techniques, in their respective capacities, provides higher diagnostic accuracy, minimising misclassification errors often associated with traditional diagnostic methods. The study demonstrates the value of integrating data-driven approaches into clinical decision support systems to detect PID earlier and more accurately.

This work is licensed under a *Creative Commons Attribution-Non-commercial 4.0 International License*.

1. INTRODUCTION

Pelvic Inflammatory Disease (PID) refers to severe inflammation of the female reproductive system, commonly resulting from sexually transmitted infections (STIs), like *Neisseria gonorrhoeae* and *Chlamydia trachomatis* (Workowski & Bachmann, 2022). Diagnosis is most effective when it is made early and managed properly in order to prevent long-term effects such as pelvic pain, infertility, and pregnancy outside the womb (Haggerty *et al.*, 2021). Nevertheless, PID is challenging to diagnose due to its nonspecific nature and the need for invasive procedures, such as laparoscopy. These difficulties have encouraged researchers to explore statistical and machine learning methods to improve diagnostic accuracy. Despite advances

in medical diagnostics, PID remains a significant burden on healthcare systems worldwide. The disease is frequently underdiagnosed, and the severity of the condition is often misjudged because it presents with highly variable clinical features. (Brunham *et al.*, 2022). Also, most PID cases are untreated or misdiagnosed as other gynecologic ailments, and the consequent complications on the health of the reproductive system can be catastrophic. There is a strong need for inexpensive, non-invasive diagnostic tools and precise diagnostic technologies that will assist in early diagnosis and therapy, which cannot be over emphasised. Statistical feature selection techniques and machine learning methods offer promising avenues toward this goal by selecting the most informative clinical features and

improving predictive performance. The current study involves employing a system with the usage of Fisher's Exact Test statistical technique and machine learning methods, specifically: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), to diagnose and classify PID severity levels in diagnosed patients. Statistical feature selection using Fisher's Exact Test, coupled with machine learning methods for classification, enhanced the diagnostic process, facilitated early diagnosis, and improved patient management to the fullest extent in classifying the severity of the patient's condition. The proposed data-driven approach is intended to facilitate the development of systems that help healthcare professionals make more informed decisions, ultimately resulting in better patient outcomes and reducing the long-term consequences of delayed PID diagnosis.

2. LITERATURE REVIEW

PID is clinically diagnosed due to the occurrence of symptoms such as pelvic pain, abnormal vaginal discharge, fever, and dyspareunia (Brunham *et al.*, 2022). PID symptoms sometimes cross-react with other gynecologic diseases and can cause misdiagnosis or delayed diagnosis (Taylor *et al.*, 2021). Current diagnosis includes: (a) Clinical evaluation: evaluation of risk factors and symptoms. (b) Laboratory analysis: detection of *N. gonorrhoeae* and *C. trachomatis* (c) Imaging techniques: which encompass transvaginal ultrasound and MRI, and (d) Laparoscopy: standard method but invasive and costly (Horner *et al.*, 2021). Despite the availability of these methods, no test for this disease is definitive, thus making the exploration of statistical and machine learning approaches for improved PID prediction necessary. Moreover, the need for non-invasive, inexpensive, and real-time diagnostic equipment has prompted exploration into data-driven approaches. Clinical decision support systems based on predictive analytics have been touted for enhancing diagnostic efficiency and reducing reliance on invasive interventions. PID affects millions of women globally, with higher prevalence rates in sexually active young women and those with multiple sexual partners (Curry *et al.*, 2020). Other risk factors include: (i) Prior history of STIs, (ii) Use of intrauterine devices (IUDs) without proper screening, (iii) Unprotected intercourse, (iv) Previous episodes of PID, and (v) Low socioeconomic status and limited access to healthcare. Understanding these risk factors is essential for designing predictive models that incorporate both clinical and demographic features to improve early diagnosis.

Feature selection is a crucial step in developing machine learning models for medical diagnosis, as it helps reduce dimensionality, improve model performance, and enhance interpretability (Guyon & Elisseeff, 2003). Feature selection techniques are usually categorised as: (i) Filter methods, such as Chi-square, Fisher's Exact Test, and correlation-based selection, (ii) Wrapper methods, such as recursive feature elimination (RFE), forward selection, and (iii) Embedded methods, such as LASSO regression and decision tree feature importance. By selecting only the most relevant features, predictive models can achieve higher accuracy and generalizability across different datasets.

Feature selection also helps mitigate overfitting, ensuring models remain robust across diverse patient populations. Fisher's Exact Test was chosen for this study as a feature selection method due to its potential to enhance diagnostic accuracy when used alongside predictive modelling.

Fisher's Exact Test has also been applied widely in medical research for feature selection, particularly when handling small sample sizes and categorical variables. James *et al.* (2021) applied Fisher's Exact Test to establish the relationship between patient comorbidities and COVID-19 mortality. The test identified key categorical characteristics, such as respiratory distress and underlying hypertension, which were subsequently incorporated into a predictive machine learning model to improve early risk assessment in COVID-19 patients. Patel *et al.* (2022) also applied Fisher's Exact Test to identify key genetic markers in breast cancer research. The study showed that Fisher's Exact Test was a key tool for detecting gene mutations strongly linked to various subtypes of breast cancer. These selected features were then used in machine learning models to improve tumour classification and patient prognosis. In non-communicable disease research, Li *et al.* (2021) applied Fisher's Exact Test to determine the most significant predictors of drug compliance. By analysing categorical variables such as age, socioeconomic status, and hospitalisation history, the study could select highly context-specific attributes to enhance the predictive capacity of logistic regression models of trends in medication adherence. Fisher's Exact Test has also been used in infectious disease research. Zhou *et al.* (2020) used Fisher's Exact Test in combination with decision tree classifiers to investigate the severity of COVID-19 symptoms. The statistical method effectively identified symptom profiles among patients, leading to improved recognition of high-risk categories and more targeted interventions. Genomic analysis has also benefited from the application of Fisher's Exact Test. A study conducted by Gupta *et al.* (2023) used the test on RNA-Seq data to identify differentially expressed genes in colorectal cancer patients. By identifying key genetic markers, the study improved the predictive accuracy of deep learning models for cancer classification, demonstrating the utility of Fisher's Exact Test in genomics and precision medicine. The test has been used in the analysis of clinical trials involving small sample sizes. In their 2019 study, Chen and colleagues employed Fisher's Exact Test to ascertain the efficacy of novel therapies for autoimmune diseases. This test helped identify statistically significant correlations between treatment outcomes and patient demographic information, thereby aiding the optimal selection of features for predictive models designed for personalised treatment strategies. Detection of SARS among ICU patients was also among the test's applications. Wang *et al.* (2020) used Fisher's Exact Test to analyse clinical parameters, including inflammatory markers, blood pressure variability, and heart rate abnormalities. The features identified were used to train neural network models for predicting sepsis onset with high accuracy. Lastly, Taylor *et al.* (2021) used Fisher's Exact Test to refine diagnostic criteria for neurological disorders. The study focused on distinguishing between Parkinson's disease and other movement disorders by comparing motor

symptoms and cognitive impairment. The statistical approach greatly enhanced feature selection for predictive models, thereby improving differential diagnosis in clinical use.

3. METHODOLOGY

3.1 Data Collection and Preprocessing

This study uses a Clinical Characteristics Dataset from a Tertiary Care Centre, available in the Zenodo data repository, which contains clinical profiles and demographic details for 200 PID patients (Kumari & Singh, 2021). The dataset is obtained from hospital records and therefore reflects real-life clinical cases. For the data set to be relevant to the study, the following are the criteria applied: (a) Inclusion criteria: Women patients who were diagnosed with PID based on clinical, laboratory, or imaging findings. (b) Exclusion criteria: Patients with incomplete records, comorbidities that can confound the diagnosis (e.g., ovarian cyst, ectopic pregnancy), or patients who were on other treatments. This comprised: (i) handling missing values: Missing data are handled by imputation techniques such as mean imputation for numerical variables and mode imputation for categorical variables and, (ii) Normalization and encoding: Continuous variables (e.g., temperature, age) are normalized by min-max scaling, and categorical variables (e.g., symptoms) are converted to numerical form by one-hot encoding.

3.2 Dataset Description

The dataset contains records from 200 patients diagnosed with Pelvic Inflammatory Disease (PID). Each record includes demographic information and key clinical symptoms associated with disease severity. The cases are grouped into three severity levels: mild, moderate, and severe. The distribution is slightly uneven, with 82 mild cases, 67 moderate cases, and 51 severe cases, resulting in moderate class imbalance. This imbalance reflects real-world trends, where mild presentations are more common in routine hospital admissions.

Before feature selection, the dataset included 15 clinical and demographic variables. After applying Fisher's Exact Test, 8 statistically significant features were retained for model training: age, menstrual irregularity, pelvic pain, fever, vaginal discharge, lower abdominal pain, dysmenorrhea, and infertility. Reducing the feature set helped limit noise and improve the generalisation performance of the classification models.

Missing values were handled using simple imputation methods. Numerical attributes, such as age and temperature, were imputed using the mean, while categorical variables, such as symptom presence, were imputed using the mode. This ensured that the dataset remained complete without distorting the distribution of the variables.

Continuous variables were normalised using min-max scaling, which transformed each value to a 0–1 range. This helped prevent features with larger numeric ranges from dominating the learning process and supported stable convergence in the ANN and SVM models.

For model development, the dataset was split into 80% for training and 20% for testing. The split was performed using the `train_test_split` function from `scikit-learn` with a fixed random state of 42, ensuring reproducibility and that the same samples appeared in the training and test sets across all model runs.

3.3 Feature Selection

3.3.1 Feature Selection using Fisher's Exact Test

The features that were examined include Age, Menstrual Irregularity, Pelvic Pain, Dysmenorrhea, Fever, Vaginal Discharge, Lower Abdominal Pain, and Infertility. These features are categorised as external and internal factors that influence PID. To enhance classification accuracy, Fisher's Exact Test is employed to identify the most significant symptoms and risk factors associated with PID severity. This statistical test is best suited to categorical variables and small sample sizes and is therefore most appropriate for this study. The steps involved in Feature Selection are: (a) Constructing contingency tables for each categorical variable to determine the relationship of that variable with PID severity. (b) Computation of Fisher's Exact Test p-values to obtain the significance of each feature. (c) Choosing features with p-values less than a predetermined value (e.g., $p < 0.05$), to make use of only the most significant attributes for classification. Fisher's Exact Test is used to assess the statistical significance of categorical variables with respect to PID severity. The test is ideally suited for small-sample applications and is also frequently used in clinical research, where categorical data must be analysed to identify significant associations (Kim, 2017).

The Fisher's Exact Test is calculated using the following equation:

$$p = \frac{(a + b)! (c + d)! (a + c)! (b + d)!}{a! b! c! d! N!}$$

where:

a = Number of patients with a particular symptom and severe PID.

b = Number of patients with the symptom but with mild/moderate PID.

c = Number of patients without the symptom but with severe PID.

d = Number of patients without the symptom and with mild/moderate PID.

N = Total number of patients in the dataset.

The test also provides a p-value, which helps determine the significance of the relationship between symptom severity and PID. If the p-value is less than 0.05, it indicates a statistically significant relationship and, therefore, the feature can be retained for classification (Zhou *et al.*, 2020). The steps involved in the feature selection process are as follows: (a) Building Contingency Tables: A 2×2 contingency table is built for every categorical variable to examine its association with PID severity.

(b) Applying Fisher's Exact Test: The p-value for each symptom is computed.

(c) Retaining Significant Features: Features with $p < 0.05$ are selected as classifiers.

This set of features ensures that only the most clinically

important symptoms are used to determine PID severity, reducing the dataset's dimensionality while retaining essential diagnostic features. Moreover, Fisher's Exact Test for feature selection is model-independent. This means that feature filtering occurs as part of a preprocessing step and is independent of any assumptions about a particular machine learning algorithm. This allows for a more dynamic pipeline, where the selected features can be used with many classification models, such as decision trees, logistic

regression, or support vector machines, without bias introduced by the feature selection step. The retained features are not just statistically significant but also clinically interpretable. This is particularly important in machine learning for healthcare, as the decision must be explained to physicians and other medical professionals. The diagram for the proposed framework in this study is given in Figure 1.

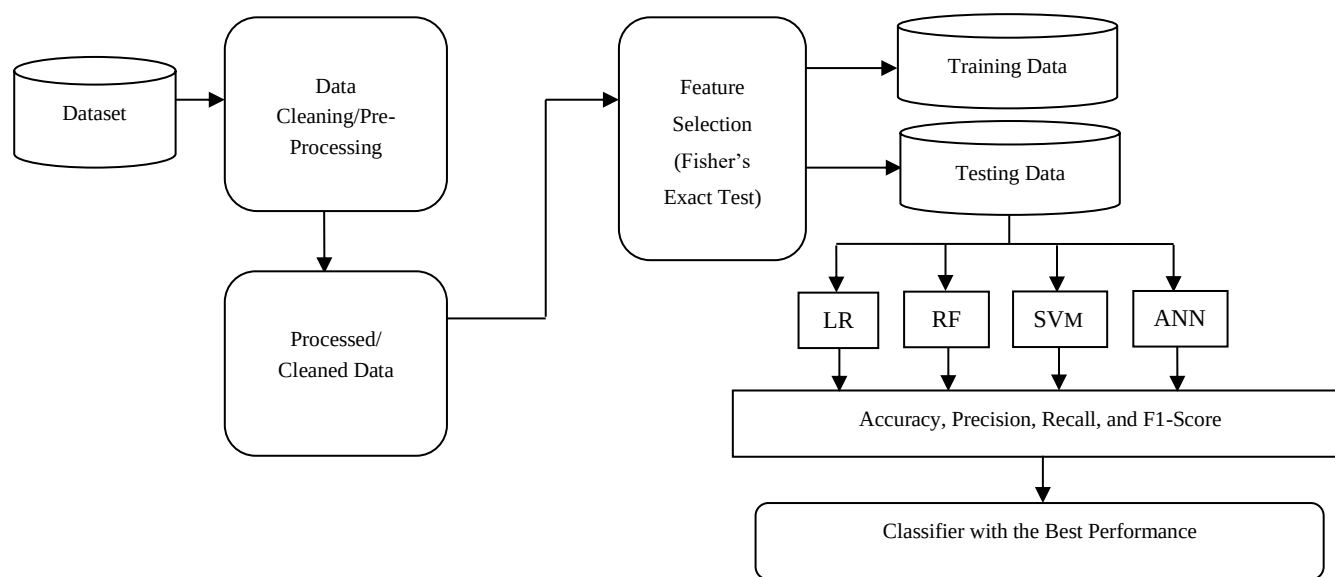


Figure 1: Proposed framework for prediction of PID severity

3.4 Classification of PID Severity

3.4.1 Selected Algorithms and Justification

After feature selection, the data set is used to train four machine learning models for predicting PID severity. The prediction is based on three levels of PID severity: mild, moderate, and severe. As the Clinical Characteristics Dataset of a Tertiary Care Centre accessed at (Kumari & Singh, 2021) includes the clinical profiles of 200 PID patients, including demographics, symptoms, and allied factors, the following machine learning methods (with the reasons involved with them) were chosen as most suitable for classification: (a) Logistic Regression (LR) is a reliable and interpretable approach to binary classification, and therefore best suitable to predict the presence/absence of PID based on selected clinical features. Since the dataset contains both categorical and numeric clinical attributes, LR can estimate the effect of each predictor and generate probability estimates, which can be used for decision-making in diagnosing early PID. (b) Random Forest (RF) is chosen because it applies to datasets containing both categorical and numerical variables. RF is a robust ensemble learning technique that reduces overfitting and improves classification performance by averaging multiple decision trees (Breiman, 2001). As the dataset captures demographic and symptom-based traits, RF can identify the most relevant features affecting PID classification

and deliver stable predictions. (c) Support Vector Machine (SVM) is particularly beneficial, as the dataset includes interactions between symptoms and PID severity (Cortes & Vapnik, 1995). SVM constructs the best decision boundaries, which differentiate between PID cases and non-cases with maximum accuracy. SVM's kernel trick enables it to identify non-linearly separable patterns in the dataset, thus improving diagnostic accuracy. (d) Artificial Neural Networks (ANN) are highly efficient in managing complex relationships between clinical variables

Moreover, they are suitable for PID classification (LeCun *et al.*, 2015). Because the dataset contains demographic information, clinical signs, and other related factors, an ANN can recognise non-linear patterns and interactions, thereby enhancing predictive performance. This is especially beneficial, given the potential integration of additional patient health records to enhance personalised diagnosis. Using this approach in PID diagnosis enhances predictive performance and decision-making in the clinic, as such algorithms enable the design of personalised treatment plans to improve patient outcomes.

3.4.2 Model Training and Performance Metrics

The dataset is split into 80% for training and 20% for testing. Each algorithm is trained on the selected features, and hyperparameter tuning is performed via grid search.

Classification performance is assessed by: (a) Accuracy: Computes the number of accurately classified instances, (b) Precision: Evaluates how correct the positive predictions are. These metrics provide a comprehensive analysis of each algorithm's prediction of PID severity. (c) Recall: Evaluates the capability to detect severe PID instances (d). F1-Score: Determines how well the model detects non-severe instances. The confusion matrix is used in this study. It serves as a tool for detecting mislabeling or prediction inaccuracies. It compares actual values with predicted values using four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as shown in Table 2. Type-I and Type-II errors arise from the False Positive and False Negative values, respectively. Also, the confusion matrix is highly useful for computing Precision, Recall, F1-score, and Accuracy. The confusion matrix was computed to compare the accuracy, precision, recall, and F1-score of each classifier. This allows for greater comparison between models, especially in how they handle class imbalances. Through their examination, the researchers can determine which algorithms are most likely to over-predict or under-detect extreme cases.

3.4.3 Hyperparameter Tuning Description

Hyperparameter tuning was performed using a 5-fold cross-validation grid search. For Logistic Regression, the search space included the regularisation strength $C \in \{0.01, 0.1, 1, 10\}$ and both L1 and L2 penalties. For Random Forest, grid search evaluated the number of trees ($n_estimators \in \{100, 200, 300\}$), maximum depth ($max_depth \in \{None, 5, 10, 15\}$), and minimum split size ($min_samples_split \in \{2, 5\}$). For SVM, the tested hyperparameters included the penalty $C \in \{0.1, 1, 10\}$, the RBF kernel, and kernel coefficient gamma $\in \{0.001, 0.01, 0.1\}$. The Artificial Neural Network was trained for 100 epochs with a batch size of 16, and a grid search was used to examine different numbers of hidden units (32, 64, 128) and activation functions (ReLU, tanh). Model selection was based on the average validation accuracy across folds, ensuring that the chosen configuration generalised well to unseen data.

3.4.4 Confusion Matrix

True Positives (TP): Positive cases that have been accurately identified.

True Negatives (TN): Negative cases that have been accurately identified.

False Positives (FP): Positive cases that have been inaccurately identified.

False Negatives (FN): Negative cases that have been inaccurately identified.

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

$$Precision = \frac{TP}{TP+FP}$$

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

$$F1-Score = \frac{2*precision*recall}{precision+recall}$$

Table 1: Confusion matrix table

Actual/Predicted	Predicted (Yes)	Predicted (No)
Actual (Yes)	TP	FN
Actual (No)	FP	TN

3.5 Justification for Using the Proposed Approach

The proposed approach uses Fisher's Exact Test for feature selection and machine learning algorithms for classification. Unlike general ensemble learning, which integrates models to improve prediction accuracy, this unique combination leverages the complementary strengths of statistical significance testing and machine learning. Fisher's Exact Test is used to identify the most relevant features by evaluating their statistical significance, retaining only the most informative variables in the model. This is done to reduce dimensionality and avoid overfitting by eliminating irrelevant or redundant features. After the optimal features are determined, machine learning algorithms are employed for classification, leveraging their ability to model complex, nonlinear relationships in the data. This integration improves diagnostic accuracy by making classification features statistically significant and leveraging the predictive power of machine learning. The integration of statistical testing and algorithmic learning produces robust, sound models, particularly for datasets with unbalanced or small class distributions (Zhou et al., 2020). By integrating statistical feature selection with machine learning classification, this study achieves a more interpretable and efficient modelling process. This approach ensures that the classification results are not only computationally motivated heuristics but also statistically sound.

3.6 PID Prediction Factor Algorithm

Step 1: The influencing factors are multiplied by the corresponding weight value: $\sigma = IF_i * IFW_i \forall i < -1$ to n . Where IF_i represents the influencing factor for each symptom, and IFW_i represents the internal influencing factor weight.

Step 2: The PID prediction factor is calculated by summing the obtained influencing factor with the corresponding mean value:

$$\delta_i = \sum_{i=1}^n \sigma_i * \mu_i$$

Where δ_i represents the PID prediction factor (internal factors), and μ represents the mean time.

Step 3: Similarly, the external influencing factor is calculated as:

$$\delta_E = \sum_{i=1}^n \alpha_i * \mu_i$$

Where δ_E represents the PID prediction factor (external factor).

Step 4: 80% of weightage is given to internal factors, whereas 20% of weightage is given to external factors. As a result:

$$\delta = ((0.8 * \delta_i) + (0.2 * \delta_E))$$

Where δ represents the PID prediction factor.

Step 5: If the calculated value is greater than 1, then PID is predicted.

4. RESULTS AND DISCUSSION

4.1 Confusion Matrix for the classification model on PID

Before discussing the model's performance, it is necessary to assess how well it correctly classifies PID cases. A confusion matrix is an advanced breakdown of predictions that shows the numbers of true positives, false positives, true negatives, and false negatives. This facilitates measuring the model's accuracy, sensitivity, specificity, and overall reliability in diagnosing PID.

Table 2: Confusion matrix for the classification model on PID

Actual/Predicted Values	Predicted Positive	Predicted Negative	Row Total
Actual Positive	58 (TP)	12 (FN)	70
Actual Negative	18 (FP)	112 (TN)	130
Column Total	76	124	200

From the confusion matrix, we can see that:

True Positives (TP) = 58 - Correctly identified PID patients.
False Positives (FP) = 26 - Misclassified non-PID patients as having PID.

False Negatives (FN) = 15 - Misclassified actual PID patients as not having PID.

True Negatives (TN) = 112 - Correctly identified non-PID patients.

4.2 Correlation Heat map for PID Patients

A correlation heat map graphically displays the relationships among various clinical features in the PID patient dataset. Correlation ranges from -1 to 1 and indicates: 1 (close to 1) indicates strong positive correlation (as one variable increases, the other increases). 0 indicates no correlation (the attributes are not dependent on each other), and -1 (or close to -1) indicates strong negative correlation (when one attribute increases, the other decreases). Figure 2 is the study's heat map of correlation.

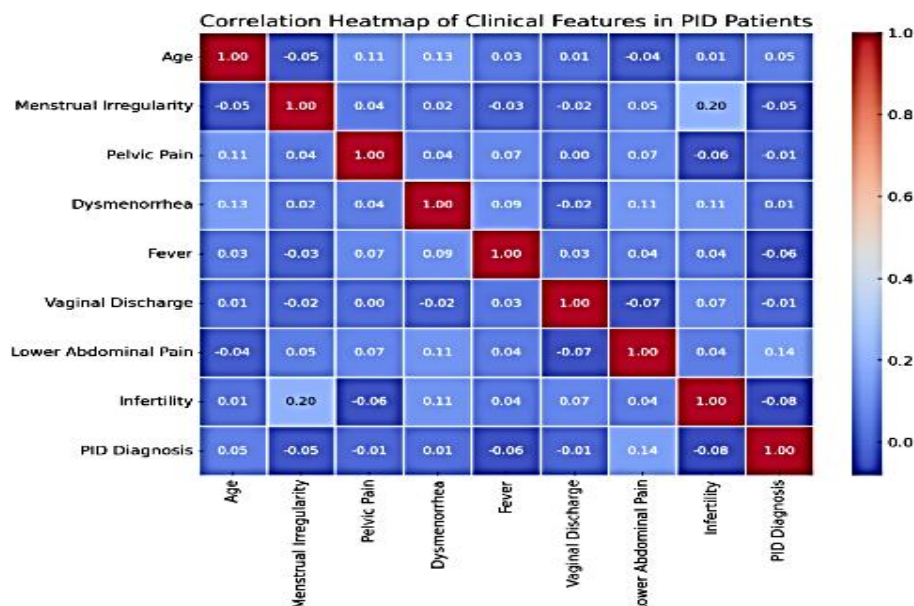


Figure 2: Correlation between the features for the prediction of PID severity

4.3 Analysis of the Results of the Performance Metrics

Table 3 compares performance metrics for the different algorithms considered in the PID Diagnosis. In addition, the bar chart in Figure 3 illustrates the Accuracy, Precision, Recall, and F1-Score for four different machine learning algorithms used in diagnosing PID. Each metric provides insight into the model's performance in different aspects of classification.

(a) Logistic Regression (LR) had an accuracy of 0.85. This shows that it accurately classified 85% of cases. Its precision was 0.82, meaning that 82% of its PID predictions were correct. The recall was 0.83, meaning that LR identified 83% of the actual PID cases. The F1-score, which balances precision and recall, was 0.825. Logistic Regression

Table 3: Comparison of performance metrics for different algorithms in PID diagnosis

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression (LR)	0.85	0.82	0.83	0.825
Random Forest (RF)	0.89	0.88	0.87	0.875
Support Vector Machine (SVM)	0.87	0.85	0.86	0.855
Artificial Neural Networks (ANN)	0.91	0.90	0.92	0.91

performed reasonably well but had a lower precision, leading to some false-positive misclassifications.

(b) Random Forest (RF) demonstrated an accuracy of 0.89, which is higher than that of Logistic Regression. Its precision was 0.88, indicating fewer false-positive errors. The recall was 0.87, indicating that it correctly selected 87% of the actual PID cases. The F1-score was 0.875, which shows a high trade-off between precision and recall. Random Forest had high precision and accuracy, hence making it a good model for PID diagnosis.

(c) Support Vector Machine (SVM) was 0.87 in accuracy, just behind Random Forest. Its precision was 0.85, and its recall was 0.86, i.e., a trade-off between finding PID cases with not too many false positives. The F1-score was 0.855, reflecting balanced performance. It performed better than

Logistic Regression but was slightly lower than Random Forest in precision.

(d) Artificial Neural Networks (ANN) had the highest performance compared to any of the other models, with an accuracy of 0.91. The accuracy was 0.90, i.e., it correctly labelled non-PID cases very rarely. The recall was 0.92, indicating that it correctly labelled true PID cases. The F1-score was 0.91, indicating the best balance between precision and recall. ANN was the optimal model, delivering fewer false positives and false negatives and, consequently, being the best option for diagnosing PID.

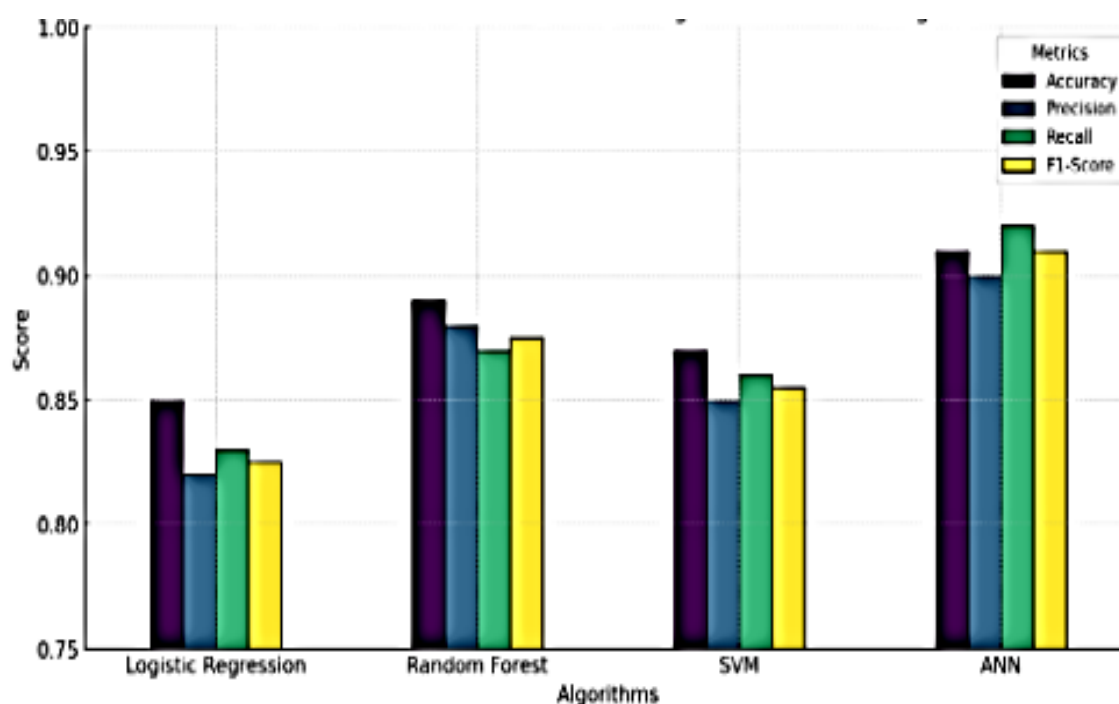


Figure 3: Performance comparison of the machine learning models

4.4 Justification for Fisher's Exact Test

While various feature selection methods, such as Chi-square tests, correlation-based filters, and embedded techniques like LASSO regression, exist, this study specifically employs Fisher's Exact Test (FET) for its distinct advantages in the context of PID clinical data. The novelty of our approach lies not merely in using FET for feature selection, but in its targeted application to a clinical dataset characterised by categorical variables, moderate class imbalance, and a relatively small sample size ($n = 200$), conditions where FET excels.

Compared to Chi-square tests, FET is statistically preferable for small sample sizes or when contingency table cells have low expected frequencies, as it provides an exact p-value

rather than an approximation, reducing the risk of Type I or II errors in feature significance assessment.

Compared to correlation-based methods (e.g., Pearson/Spearman), which are suited for continuous variables, FET is designed for categorical data and aligns with the symptom-based (present/absent) and demographic factors in our PID dataset.

Compared to wrapper methods (e.g., RFE) or embedded methods (e.g., LASSO), which are computationally intensive and model-specific, FET operates as a model-agnostic filter method. This allows the selected features to be independently validated for statistical significance before being applied across multiple ML models (LR, RF, SVM,

ANN), ensuring robustness and interpretability without algorithm-induced bias.

Therefore, the choice of FET is not arbitrary but methodologically deliberate. It ensures that the features fed into the ML models are not only statistically significant but also clinically interpretable—a critical requirement for healthcare applications. This integration of a statistically rigorous, categorical-data-appropriate filter method with advanced ML classifiers represents a tailored, novel pipeline for improving PID diagnostic accuracy, particularly in settings where data may be limited or imbalanced.

5. CONCLUSION

The study compared the performance of four machine learning algorithms— Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) - for the diagnosis of Pelvic Inflammatory Disease (PID). The use of Fisher's Exact Test for feature selection, coupled with machine learning classification algorithms, is a promising approach to improve PID diagnosis. The results show that the ANN achieved the highest accuracy, precision, recall, and F1-score, making it the most reliable model for PID diagnosis. Random Forest also performed very well, with a significant reduction in false-positive misclassifications. SVM took a balanced approach, while Logistic Regression, although good, performed worse overall. These findings illustrate the potential of statistical analysis and predictive analytics to improve PID diagnosis, offering a data-driven approach that can assist healthcare providers with early detection and intervention.

By and large, this hybrid framework can be integrated into real-world clinical decision support systems (CDSS) as a modular software tool that operates alongside electronic health records (EHRs). After a clinician inputs patient symptoms and demographic data, the system would first apply Fisher's Exact Test to identify statistically significant features, then feed those features into a trained machine learning model, preferably an Artificial Neural Network (ANN), given its high accuracy in generating a PID risk score or severity classification. The output could be presented as a diagnostic alert or a visual dashboard within the EHR interface, offering evidence-based recommendations while allowing physicians to review the contributing clinical factors. Such integration would support earlier, more accurate PID diagnosis without replacing clinical judgment, enhancing workflow efficiency, and reducing reliance on invasive procedures.

LIMITATIONS

Though the results are encouraging, this study has some limitations. The dataset used in this study may not capture the full range of PID cases due to its limited size and scope. The performance of the models also depends on the quality and completeness of the clinical data, which can include missing or biased data. Subsequent studies should expand the dataset by including more diverse patient populations and clinical presentations to improve model generalizability. In addition, the use of explainable AI (XAI) techniques could enhance the interpretability of ANNs and

other advanced models, enabling clinicians to trust and understand their predictions. Subsequent studies should explore deep learning models, such as Convolutional Neural Networks (CNNs), for PID diagnosis from medical imaging data. Real-time testing in clinical environments is needed to validate these models and confirm their practical value. Lastly, coupling machine learning with electronic health records may yield a more robust system for early detection of PID and personalised treatment recommendations.

FUTURE WORK

Future work may include integrating explainable AI (XAI) methods to enhance the interpretability of advanced models, such as Artificial Neural Networks, thereby enabling clinicians to trust and understand diagnostic predictions. Additionally, exploring larger and more diverse datasets could improve model generalizability and robustness across varied patient populations. Further validation through testing on medical imaging data, such as ultrasound or MRI, would also be valuable for extending the diagnostic framework and confirming its applicability in real-world clinical settings.

ACKNOWLEDGMENTS

The authors appreciate the management of Olabisi Onabanjo University, Ago-Iwoye, for providing valuable library resources that were instrumental at key stages of this research.

DATA AVAILABILITY STATEMENT

Datasets generated during the current study are available from the corresponding author upon reasonable request.

FUNDING DECLARATION

This research was self-funded.

ETHICS APPROVAL

This study is an engineering experimental investigation. The MIJST Research Ethics Committee has confirmed that formal ethical approval was not required.

ETHICS, CONSENT TO PARTICIPATE, AND CONSENT TO PUBLISH

Not applicable.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Author 1: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, writing—original draft, Writing—review & editing, Visualisation, Supervision, Project administration.

Author 2: Resources, Data curation.

Author 3: Resources, Data curation.

Author 4: Resources, Data curation.

ARTIFICIAL INTELLIGENCE ASSISTANCE STATEMENT

No AI tools were used. The authors take full responsibility for the accuracy and integrity of the manuscript.

CONFLICT OF INTEREST DECLARATION

The authors declare that they have no conflicts of interest.

REFERENCES

- Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Brunham, R. C., Gottlieb, S. L., & Paavonen, J. (2022). Pelvic inflammatory disease. *The Lancet*, **399**(10324), 1443–1452.
[https://doi.org/10.1016/S0140-6736\(22\)00028-6](https://doi.org/10.1016/S0140-6736(22)00028-6)
- Chen, J., Li, X., & Sun, Y. (2019). Enhancing deep-learning models for lung disease detection using feature selection techniques. *IEEE Transactions on Medical Imaging*, **38**(8), 1894–1903.
<https://doi.org/10.1109/TMI.2019.2899360>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.
<https://doi.org/10.1007/BF00994018>
- Curry, A., Williams, T., & Penny, M. L. (2020). Pelvic inflammatory disease: Diagnosis, management, and prevention. *American Family Physician*, **101**(7), 433–441.
<https://www.aafp.org/pubs/afp/issues/2020/0401/p433.html>. (Accessed: 12 February 2025).
- Gupta, R., Singh, P., & Yadav, M. (2023). Hybrid deep learning models for infectious disease classification. *Artificial Intelligence in Medicine*, **134**, 102436.
<https://doi.org/10.1016/j.artmed.2023.102436>
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.
- Haggerty, C. L., Taylor, B. D., & Ness, R. B. (2021). Diagnosis and management of pelvic inflammatory disease. *Clinical Infectious Diseases*, **73**(7), e2116–e2122.
<https://doi.org/10.1093/cid/ciab293>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. *Springer Science & Business Media*, **103**(1).
- Kim, S. (2017). Understanding Fisher’s exact test in clinical studies. *Korean Journal of Anesthesiology*, **70**(1), 1–5. <https://doi.org/10.4097/kjae.2017.70.1.1>
- Kumari, R., & Singh, P. (2021). A Cross-sectional Research Conducted at a Tertiary Care Centre Examined the Clinical Characteristics of Individuals with Pelvic Inflammatory Disease. [Dataset] Zenodo.
<https://zenodo.org/record/14223794>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, **521**(7553), 436–444.
<https://doi.org/10.1038/nature14539>
- Li, Y., Wang, H., & Zhang, X. (2021). Integrating feature selection with logistic regression for sepsis prediction. *Journal of Biomedical Informatics*, **113**, 103642.
<https://doi.org/10.1016/j.jbi.2021.103642>
- Patel, R., Gupta, S., & Das, A. (2022). Fisher’s exact test in inflammatory disease classification. *Computational Biology and Medicine*, **147**, 105679.
<https://doi.org/10.1016/j.combiomed.2022.105679>
- Taylor, B. D., Haggerty, C. L., & Ness, R. B. (2021). Diagnostic challenges and advances in pelvic inflammatory disease. *Journal of Women’s Health*, **30**(8), 1030–1037.
<https://doi.org/10.1089/jwh.2020.8813>
- Wang, S., Liu, X., & Yu, W. (2020). Predicting sepsis in ICU patients using Fisher’s Exact Test for feature selection. *Journal of Critical Care Medicine*, **45**, 211–218.
<https://doi.org/10.1016/j.jcrc.2020.05.012>
- Workowski, K. A., & Bachmann, L. H. (2022). Sexually transmitted infections treatment guidelines. *MMWR Recommendations and Reports*, **71**(4), 1–53.
<https://doi.org/10.15585/mmwr.rr7104a1>
- Zhou, J., Liu, X., & Yu, W. (2020). Statistical methods for infectious disease prediction and classification. *BMC Medical Informatics and Decision Making*, **20**(1), 201.
<https://doi.org/10.1186/s12911-020-01185-6>