Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Prediction of Peritonitis Infection Risk in CAPD Patients using Random Forest Algorithm

Silviani Nadia Gustaman[1*], Rudi Kurniawan[2], Bani Nurhakim[3], Cep Lukman Rohmat[4], Gifthera Dwilestari[5]

[1,2]*Department of Informatics Engineering, STMIK IKMI Cirebon, Indonesi*
[3]*Department of Informatics Management, STMIK IKMI Cirebon, Indonesia*
[4]*Department of Software Engineering, STMIK IKMI Cirebon, Indonesia*
[5]*Department of Information Systems, STMIK IKMI Cirebon, Indonesia*
nadiagustamans@gmail.com[1*], rudi226ikmi@gmail.com[2], baninurhakim@gmail.com[3], ceplukmanrohmat@gmail.com[4], gifthera.ikmi@gmail.com[5]

Abstract

Peritonitis is a serious complication frequently experienced by patients undergoing Continuous Ambulatory Peritoneal Dialysis (CAPD) and may worsen patient outcomes if not detected early. This study aims to develop a machine learning model to predict peritonitis risk using the Random Forest algorithm and to interpret prediction results using Explainable Artificial Intelligence (XAI). The study utilized a secondary dataset obtained from Kaggle consisting of 20,538 clinical records that were transformed to represent CAPD-related clinical parameters. The research stages included data preprocessing, feature selection using SelectKBest (f_classif), dataset splitting into training and testing sets, model development using Random Forest, and performance evaluation using accuracy, precision, recall, F1-score, and Area Under Curve (AUC). Model interpretability was analyzed using SHAP to identify feature contributions. The experimental results demonstrate that the proposed model achieved an accuracy of 98.70%, precision of 98.22%, recall of 99.24%, F1-score of 98.73%, and AUC of 1.00. The findings indicate that Random Forest provides highly reliable predictive performance and interpretable insights into clinical features influencing peritonitis risk. The developed model has potential to support clinical decision-making systems for early detection of peritonitis risk in CAPD patients.

Keywords: *CAPD; Explainable artificial intelligence; Machine learning; Peritonitis; Random forest*

1. Introduction

Machine learning has been widely applied in healthcare to support clinical prediction and medical decision-making processes [1],[2]. The ability of machine learning algorithms to analyze large-scale medical datasets enables improved diagnostic accuracy and risk prediction performance.

Continuous Ambulatory Peritoneal Dialysis (CAPD) is a long-term therapy for chronic kidney disease patients; however, peritonitis remains a major complication causing hospitalization and increased mortality risk [3],[4]. Early prediction of peritonitis risk is therefore essential for preventive intervention.

Random Forest has demonstrated strong predictive performance in clinical risk prediction tasks due to its capability to model nonlinear relationships and handle complex datasets [5]. Nevertheless, many machine learning models suffer from limited interpretability, reducing clinical trust and adoption [6].

Explainable Artificial Intelligence (XAI) addresses this limitation by providing interpretable explanations of model decisions. Methods such as SHAP enable understanding of feature contributions and improve transparency in healthcare AI systems [7],[8]. Therefore, this study integrates Random Forest with SHAP-based explainability to develop an interpretable prediction model for peritonitis risk in CAPD patients.

2. Research Methodology

2.1. Dataset

This study utilizes a secondary dataset obtained from Kaggle containing clinical variables associated with kidney disease indicators. The original dataset was further processed and transformed into a derived dataset representing clinical parameters related to Continuous

Ambulatory Peritoneal Dialysis (CAPD). The final dataset consists of 20,538 records with multiple clinical attributes used to represent patient health conditions and potential risk factors associated with peritonitis occurrence. The use of a large-scale secondary dataset enables broader pattern recognition and improves the robustness of the machine learning model in identifying peritonitis risk patterns [9].

2.2. Data preprocessing

Data preprocessing was conducted to improve data quality and ensure reliable model performance. This stage included data cleaning, handling missing values, feature transformation, and normalization processes. Missing or inconsistent values were addressed to prevent bias during model training. Feature transformation was applied to convert categorical variables into numerical representations suitable for machine learning algorithms. Additionally, normalization was performed to standardize feature scales, allowing the model to learn patterns more effectively and improving generalization capability across unseen data [10], [11].

2.3. Feature selection

Feature selection was performed using the SelectKBest method combined with the ANOVA *f_classif* statistical test to identify the most relevant features contributing to peritonitis risk classification. This approach evaluates the statistical relationship between each feature and the target variable, allowing irrelevant or less informative features to be removed. By reducing dimensionality, feature selection helps improve model efficiency, reduces computational complexity, and enhances predictive performance while minimizing noise within the dataset.

2.4. Model building

The Random Forest algorithm was selected as the primary classification model due to its robustness and strong performance in handling complex and high-dimensional clinical datasets. Random Forest operates by constructing multiple decision trees and aggregating their predictions through an ensemble learning approach, which reduces variance and improves prediction stability. The model was trained using default parameter settings to evaluate its baseline performance in predicting peritonitis risk among CAPD patients. This algorithm is also known for its resistance to overfitting and its ability to capture nonlinear relationships within medical data [5].

2.5. Model evaluation

Model performance was evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). Accuracy measures overall prediction correctness, while precision and recall evaluate the model's effectiveness in identifying positive cases. The F1-score provides a balance between precision and recall, making it suitable for classification tasks involving potential class imbalance. The AUC metric was used to assess the model's discriminative capability across different classification thresholds, providing a comprehensive evaluation of predictive performance.

2.6. Model intepretation

Model interpretability was analyzed using SHapley Additive exPlanations (SHAP), a method designed to quantify the contribution of each feature toward prediction outcomes. SHAP values provide both global and local explanations, enabling visualization of how individual features influence model decisions. This approach enhances transparency and supports clinical interpretability, allowing healthcare practitioners to better understand the reasoning behind model predictions. The use of explainable AI techniques is particularly important in medical applications to increase trust and facilitate decision support systems in clinical environments [7].

3. Results and Discussion

This study developed a prediction model for peritonitis risk in patients undergoing Continuous Ambulatory Peritoneal Dialysis (CAPD) using the Random Forest algorithm based on the *cleaned_peritonitis_risk_data* dataset. The dataset was divided into training and testing sets to ensure proper model generalization and to reduce the risk of overfitting. The modeling process was conducted using the default parameters of the Random Forest algorithm, followed by evaluation using several classification performance metrics. Overall, the results indicate that the model was able to effectively identify patterns associated with peritonitis risk, as demonstrated by near-perfect evaluation performance on the testing dataset. The following table presents the performance of the Random Forest model based on accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) metrics.

3.1. Dataset

Before model development, exploratory data analysis was performed to understand dataset characteristics and class distribution. The target variable *Peritonitis_Risk* was analyzed to ensure that the dataset represented both risk and non-risk classes adequately. Balanced class distribution is essential to prevent model bias toward the majority class and to ensure reliable predictive performance in medical classification tasks.

The dataset was divided into training and testing sets using a 70:30 ratio. This separation ensured that model evaluation was conducted on unseen data, allowing objective measurement of generalization capability and minimizing overfitting risk.

3.2. Data preprocessing

Data preprocessing was conducted to prepare the dataset before model training and to ensure that the input data met the requirements of machine learning algorithms. The preprocessing stage included data cleaning, handling missing values, encoding categorical variables, and feature normalization. These steps were necessary to reduce noise and inconsistencies that could negatively affect model performance.

Missing values were identified and treated using appropriate imputation techniques to maintain dataset completeness without significantly altering data distribution. Categorical variables were transformed into numerical representations using encoding methods so that they could be processed effectively by the Random Forest algorithm. This transformation ensured that all features contributed properly during model learning.

Furthermore, normalization was applied to standardize feature scales and reduce the dominance of variables with larger numerical ranges. Although Random Forest is relatively insensitive to feature scaling, normalization helps improve data consistency and supports better interpretability during feature analysis and model explanation stages. The preprocessing process resulted in a clean and structured dataset suitable for predictive modeling and evaluation.

3.3. Feature selection

Feature selection was performed using the SelectKBest method with the ANOVA (*f_classif*) statistical approach to identify features significantly associated with peritonitis risk. The selection process reduced irrelevant or redundant variables and improved computational efficiency while maintaining predictive performance.

The feature evaluation results indicated that several clinical variables demonstrated strong statistical relevance to the prediction target. This process contributed to improved model stability and enhanced interpretability during subsequent Explainable Artificial Intelligence (XAI) analysis.

The selected features represent clinically meaningful indicators commonly associated with inflammation, nutritional condition, and kidney function, supporting the relevance of the dataset transformation performed in this research.

3.4. Model buiding

The model building phase involved training a classification model using the Random Forest algorithm to predict peritonitis risk in CAPD patients. Random Forest was selected due to its ensemble learning mechanism, which combines multiple decision trees to produce more stable and accurate predictions compared to single-tree models.

During training, the dataset was divided into training and testing subsets using a 70:30 ratio, where the training data were used to learn underlying patterns and the testing data were reserved for performance evaluation. Each decision tree within the Random Forest was constructed using randomly selected subsets of features and samples, allowing the model to reduce variance and improve generalization capability.

The model was initially trained using default hyperparameter settings to evaluate baseline performance and observe how well Random Forest could capture relationships among clinical variables. The ensemble voting mechanism aggregated predictions from multiple trees to determine the final classification outcome. This approach enabled the model to handle nonlinear relationships and complex interactions among clinical features commonly found in medical datasets.

The resulting model demonstrated stable learning behavior and showed strong capability in distinguishing between patients with high peritonitis risk and those without risk, forming the basis for further evaluation and interpretability analysis using Explainable Artificial Intelligence (XAI) techniques.

3.5. Model evaluation

Model performance was evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics were selected because they provide comprehensive evaluation for medical prediction problems, particularly where false negatives may have critical clinical consequences.

The Random Forest model achieved excellent predictive performance with:

a. Accuracy: 98.70%
b. Precision: 98.22%
c. Recall: 99.24%
d. F1-score: 98.73%
e. AUC: 1.00

Figure X illustrates the Receiver Operating Characteristic (ROC) curve of the Random Forest model used for predicting peritonitis risk in CAPD patients. The ROC curve represents the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different classification threshold values.

As shown in Figure X, the ROC curve closely follows the upper-left boundary of the plot, indicating excellent classification performance. The model achieved an Area Under the Curve (AUC) value of 1.0000, which signifies near-perfect discrimination capability between high-risk and non-risk classes. This result demonstrates that the model is highly effective in distinguishing patients with potential peritonitis risk from those without risk.

The steep rise of the curve toward the top-left corner indicates that the model maintains a high true positive rate while keeping the false positive rate extremely low. Such performance suggests that the Random Forest algorithm successfully captured the underlying patterns within the clinical dataset. From a clinical prediction perspective, this outcome is particularly important because accurate identification of high-risk patients can support early preventive intervention.

However, although the obtained AUC value indicates outstanding predictive performance, further validation using independent clinical datasets is necessary to ensure model generalization and robustness in real-world medical applications.

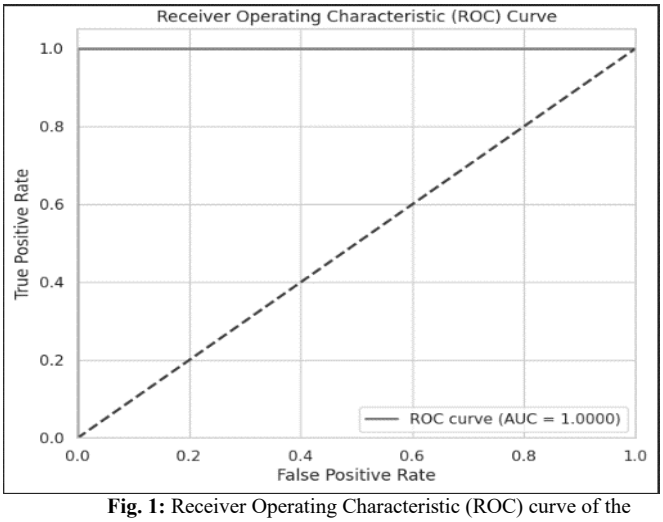

Fig. 1: Receiver Operating Characteristic (ROC) curve of the Random Forest model for peritonitis risk prediction showing an AUC value of 1.0000.

The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at various classification thresholds. As shown in Figure 1, the curve closely follows the upper-left boundary, indicating excellent classification capability.

The obtained AUC value of 1.0000 demonstrates near-perfect discrimination between risk and non-risk classes. This result indicates that the model successfully captured underlying clinical patterns associated with peritonitis risk.

From a clinical perspective, this level of discrimination suggests that the proposed model can effectively support early identification of patients requiring preventive monitoring or intervention.

3.6. Model interpretation

To enhance transparency and clinical interpretability, the trained model was analyzed using SHAP (SHapley Additive exPlanations). The SHAP analysis identified several dominant features influencing prediction outcomes, including:

a. Albumin
b. White Blood Cell (WBC) count
c. Urea level
d. Hemoglobin
e. C-Reactive Protein (CRP)

These variables are clinically associated with inflammation status, immune response, and nutritional condition, which are recognized risk factors for peritonitis in CAPD patients. The consistency between model findings and medical knowledge strengthens the validity of the developed prediction model.

SHAP visualization also enables healthcare professionals to understand how individual features contribute to prediction outcomes, thereby improving trust and adoption of machine learning systems in clinical environments.

4. Conclusion

This study successfully developed an interpretable peritonitis risk prediction model using Random Forest and SHAP. The model achieved high predictive accuracy while providing transparent explanations of prediction outcomes. The proposed approach demonstrates strong potential for implementation in clinical decision support systems for early peritonitis detection in CAPD patients.

References

[1] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges," *Sensors*, vol. 23, no. 9, p. 4178, 2023, doi: 10.3390/s23094178.

[2] S. Yin *et al.*, "Risk Factors and Pathogen Spectrum in Continuous Ambulatory Peritoneal Dialysis- Associated Peritonitis: A Single Center Retrospective Study," *Med. Sci. Monit.*, vol. 28, Aug. 2022, doi: 10.12659/MSM.937112.

[3] X. Li, Y. Zhang, and J. Zhao, "Threshold optimization for imbalanced medical data," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 243, 2022, doi: 10.1186/s12911-022-02041-7.

[4] I. G. Okpechi, "Prevalence of peritonitis and mortality in peritoneal dialysis patients in Africa," *BMJ Open*, vol. 10, no. 12, p. e039970, 2020, doi: 10.1136/bmjopen-2020-039970.

[5] A. Heidarian, "Comparison of Random Forest, logistic regression, and SVM in medical risk prediction," *BMC Med. Inform. Decis. Mak.*, vol. 22, p. 112, 2022, doi: 10.1186/s12911-022-01947-0.

[6] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 181, 2022, doi: 10.1186/s12911-022-01865-1.

[7] Z. Chen, Y. Liu, H. Wang, and J. Kang, "Interpretability in clinical machine learning models using SHAP and LIME," *J. Biomed. Inform.*, vol. 149, p. 104550, 2024, doi: 10.1016/j.jbi.2024.104550.

[8] B. D. Mittelstadt and L. Floridi, "The ethics of artificial intelligence: Mapping the debate," *Minds Mach.*, vol. 32, no. 2, pp. 403–431, 2022, doi: 10.1007/s11023-022-09606-3.

[9] B. M. Pavlyshenko, "Machine-learning models for sales time series forecasting using data from Kaggle datasets," *Data*, vol. 5, no. 1, p. 15, 2020, doi: 10.3390/data5010015.

[10] A. Pfob, "Handling missing data in electronic health record-based machine learning models: A systematic review," *J. Biomed. Inform.*, vol. 128, p. 104059, 2022, doi: 10.1016/j.jbi.2022.104059.

[11] G. Varoquaux, "Cross-validation strategies for data science," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 13854–13866, 2021, [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/ab9a86e