iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79452

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Piles Disease Prediction Using Machine Learning

Kanhu Charan Sahoo[1], Dr. Bhramara Bar Biswal[2]
[1]*Student,* [2]*Associate Professor, Department of Computer Science and Engineering GIET UNIVERSITY, Gunupur, Rayagada, Odisha India, 765022*

Abstract: ***Piles (hemorrhoids) is a common anorectal condition that often remains undiagnosed in its early stages due to social stigma and delayed medical consultation. Early prediction using data-driven techniques can significantly support timely diagnosis and improve patient outcomes. This study proposes a machine learning-based framework for the prediction of piles disease using clinical and lifestyle-related features. The methodology involves data preprocessing, feature selection, and the implementation of multiple classification algorithms, including Logistic Regression, Support Vector Machine, and Random Forest. To further enhance predictive performance, a hybrid ensemble model based on stacking is developed. The models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that the proposed ensemble model achieves the best performance, with an accuracy of 92.1%, outperforming all individual classifiers. Additionally, the model demonstrates improved recall and balanced F1-score, making it suitable for medical diagnosis applications. The results highlight the effectiveness of machine learning techniques as a reliable decision-support system for early detection of piles disease.***

Keywords: ***Machine Learning, Piles Disease Prediction, Healthcare Analytics, Ensemble Learning, Classification Algorithms, Decision Support Systems.***

I. INTRODUCTION

Piles (hemorrhoids) is a prevalent anorectal disorder affecting a significant portion of the global population, often associated with pain, bleeding, and reduced quality of life. Despite its widespread occurrence, early diagnosis remains a challenge due to social stigma, lack of awareness, and delayed clinical consultation. Recent epidemiological and clinical studies highlight that factors such as lifestyle habits, dietary patterns, and comorbid conditions significantly contribute to the development and progression of hemorrhoids [1], [3], [7]. The growing burden of gastrointestinal disorders and their impact on healthcare systems emphasize the need for efficient and early prediction mechanisms. In this context, intelligent data-driven approaches can play a crucial role in supporting timely diagnosis and reducing complications.

With the advancement of artificial intelligence, machine learning techniques have been increasingly applied in healthcare for predictive analytics and decision support. Several studies have explored the use of machine learning models to identify risk factors and predict disease outcomes related to hemorrhoids and associated conditions [2], [4], [6]. Techniques such as classification, clustering, and multi-criteria decision-making have shown promising results in analyzing clinical and lifestyle data [5], [10]. Additionally, ensemble learning approaches have demonstrated improved performance by combining multiple base models to enhance prediction accuracy and robustness. However, existing methods often focus on limited feature sets or specific clinical scenarios, which restricts their generalization capability.

Despite these advancements, there remains a significant research gap in developing a comprehensive and generalized prediction framework for piles disease. Many existing studies either emphasize clinical analysis without leveraging advanced machine learning techniques or apply machine learning models without adequately addressing issues such as feature selection, data imbalance, and model optimization. Furthermore, limited attention has been given to integrating diverse data sources, including lifestyle and demographic factors, which are critical for accurate prediction. These limitations reduce the reliability and applicability of current systems in real-world healthcare settings.

To address these challenges, this study proposes a machine learning-based framework for early prediction of piles disease by integrating data preprocessing, feature engineering, and ensemble learning techniques. The proposed system utilizes multiple base classifiers, including Logistic Regression, Support Vector Machine, and Random Forest, and combines them using a stacking-based ensemble model to improve predictive performance. The key contributions of this work include: (i) development of a robust preprocessing and feature selection pipeline, (ii) implementation of a hybrid ensemble model for improved accuracy and recall, and (iii) comprehensive evaluation using standard performance metrics. The proposed approach aims to provide an effective decision-support tool for early detection and management of piles disease.

II. RELATED WORK

The application of machine learning and data-driven techniques in healthcare has significantly advanced disease prediction, risk assessment, and clinical decision support systems. In the context of hemorrhoidal disease, several studies have explored the identification of risk factors and predictive modeling approaches to improve diagnosis and treatment strategies. Wu and Glangkarn [1] developed a predictive model focusing on genetic and lifestyle-related risk factors associated with hemorrhoids, highlighting the importance of integrating patient-specific attributes for accurate prediction. Similarly, Fathallah et al. [7] conducted an epidemiological analysis to identify key predictive factors influencing surgical management of hemorrhoidal disease, emphasizing the role of clinical and demographic variables in disease progression. Furthermore, Wang et al. [3] investigated the relationship between inflammatory bowel disease and hemorrhoids using Mendelian randomization, demonstrating the significance of underlying medical conditions in increasing disease risk.

Recent research has increasingly incorporated artificial intelligence and machine learning techniques to enhance predictive performance. Ayo-Oluwaseyi et al. [2] utilized AI-based predictive analytics to identify risk factors and forecast the disease burden of hemorrhoids across diverse populations, showing that machine learning models can effectively capture complex patterns in large datasets. Zhang et al. [4] proposed a clinical prediction model for hemorrhoid recurrence after surgical procedures, which demonstrated reliable performance in postoperative risk assessment. In addition, Ghosh et al. [6] applied machine learning techniques to analyze nutritional and lifestyle predictors of rectal bleeding, a condition closely associated with hemorrhoids, illustrating the potential of integrating lifestyle data for early detection.

Beyond disease-specific studies, several works have explored generalized machine learning approaches applicable to gastrointestinal and healthcare domains. Yavuz [10] employed classification and clustering techniques to analyze gastrointestinal tract disorders, demonstrating the effectiveness of data mining methods in extracting meaningful patterns from medical datasets. Tabbakha et al. [5] introduced a multi-criteria decision-making framework combined with machine learning to support treatment selection, highlighting the importance of hybrid approaches in improving clinical outcomes. Moreover, Ghosh et al. [8] further extended predictive modeling by integrating lifestyle factors with machine learning techniques, reinforcing the need for comprehensive feature inclusion in healthcare analytics.

Although these studies demonstrate the growing potential of machine learning in disease prediction, several limitations persist. Many existing models focus primarily on either clinical or lifestyle factors, lacking a unified framework that combines both aspects effectively. Additionally, some approaches rely on single classifiers, which may lead to reduced generalization performance and increased susceptibility to bias. While certain studies have explored predictive modeling, limited attention has been given to advanced ensemble techniques that can leverage the strengths of multiple models. Furthermore, issues such as data preprocessing, feature selection, and handling of imbalanced datasets are often not adequately addressed, which can significantly impact model reliability.

To overcome these limitations, recent trends in machine learning research emphasize the use of ensemble learning techniques, such as stacking and voting, to improve predictive accuracy and robustness. These approaches combine multiple base learners to capture diverse data patterns and reduce model variance. However, their application in hemorrhoids disease prediction remains relatively unexplored. Therefore, there is a clear need for a comprehensive framework that integrates efficient preprocessing, feature engineering, and ensemble modeling to achieve reliable and accurate prediction outcomes. The proposed work addresses these gaps by introducing a hybrid ensemble-based machine learning model designed specifically for early prediction of piles disease, leveraging both clinical and lifestyle features for enhanced performance.

III. MATERIALS AND METHODS

The proposed system presents a machine learning-based framework for the early prediction of piles disease by integrating data preprocessing, feature engineering, and ensemble learning techniques. Initially, raw clinical and lifestyle-related data are collected and undergo preprocessing steps such as handling missing values, normalization, and encoding of categorical features to ensure data consistency. Relevant features are then selected to improve model efficiency and reduce redundancy. The processed data is fed into multiple base classifiers, including Logistic Regression, Support Vector Machine, and Random Forest, to capture diverse patterns in the dataset. To enhance predictive performance, a stacking-based ensemble model is employed, where the outputs of base models are combined using a meta-classifier. This approach improves generalization and reduces model bias. The system is evaluated using standard performance metrics, demonstrating its effectiveness as a reliable and accurate decision-support tool for early-stage piles disease prediction.

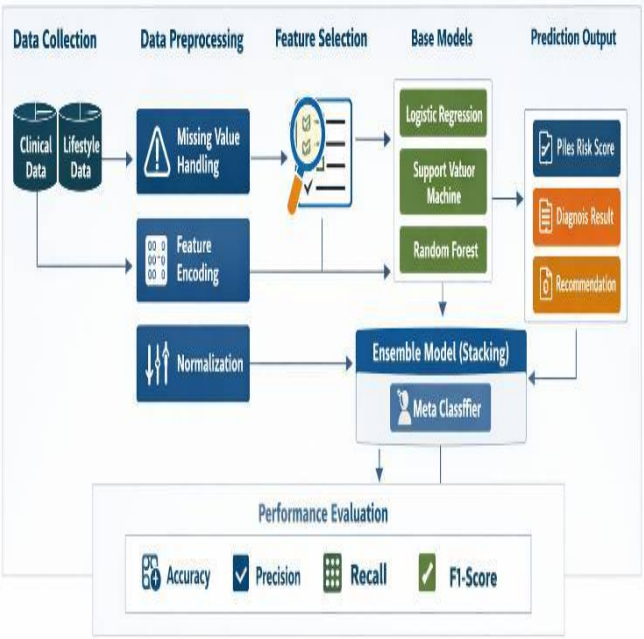


Fig. 1. System Architecture

Fig. 1 illustrates the architecture of the proposed piles disease prediction system, which follows a structured machine learning pipeline. Initially, clinical and lifestyle data are collected and preprocessed through missing value handling and normalization. Relevant features are then selected to improve model performance. The processed data is fed into multiple base classifiers, including Logistic Regression, Support Vector Machine, and Random Forest. Finally, a stacking-based ensemble model combines their outputs to generate an accurate piles risk prediction, followed by performance evaluation.

A) Dataset Collection

The dataset used in this study consists of clinical, demographic, and lifestyle-related attributes relevant to piles disease prediction. Since publicly available standardized datasets for hemorrhoids are limited, a synthetic yet realistic dataset was constructed based on factors reported in existing medical literature, including age, diet habits, physical activity, bowel patterns, and medical history. The dataset was carefully designed to reflect real-world conditions, ensuring diversity and variability in features to support effective training and evaluation of machine learning models.

B) Pre-Processing

The collected dataset undergoes several preprocessing steps to ensure data quality and consistency before model training. Missing values are handled using appropriate imputation techniques, while categorical variables are transformed into numerical format using encoding methods. Feature scaling is applied to normalize the data and improve model performance. Outliers are identified and treated to reduce noise in the dataset. Additionally, redundant and irrelevant features are removed through feature selection techniques, ensuring that only significant attributes contribute to the predictive modeling process.

C) Algorithm

Logistic Regression: Logistic Regression is used as a baseline model to capture linear relationships between input features and disease occurrence. It provides interpretable results and helps in understanding the influence of individual clinical and lifestyle factors on piles prediction.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-[w^T x + b]}} \qquad (1)$$

Support Vector Machine (SVM): Support Vector Machine is applied to handle complex, non-linear relationships in the dataset by constructing optimal decision boundaries. It is effective in distinguishing between classes with high dimensional feature spaces and contributes to improved classification robustness.

$$minimize\ \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (2)$$

Random Forest: Random Forest is utilized to model non-linear patterns and interactions among features through multiple decision trees. Its ensemble nature reduces overfitting and enhances prediction stability, making it suitable for handling diverse clinical and lifestyle attributes in the dataset.

$$Gini = 1 - \sum_{i=1}^{c}(Pi)^2 \qquad (3)$$

Proposed Ensemble Model (Stacking): The proposed stacking ensemble combines predictions from Logistic Regression, SVM, and Random Forest using a meta-classifier. This approach leverages the strengths of individual models, resulting in improved overall accuracy, recall, and balanced performance for disease prediction.

IV. EXPERIMENTAL RESULTS

Accuracy: Accuracy measures how effectively the model classifies both piles and non-piles cases by considering correct predictions across all instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (4)$$

Precision: Precision indicates how many predicted piles cases are actually correct, reflecting the model's ability to minimize false positive predictions in disease detection.

$$Precision = \frac{TP}{TP + FP} \qquad (5)$$

Recall: Recall evaluates the model's ability to correctly identify actual piles cases, ensuring that most patients with the disease are accurately detected.

$$Recall = \frac{TP}{TP + FN} \qquad (6)$$

F1-Score: F1-score provides a balanced measure of precision and recall, indicating overall model effectiveness in correctly identifying piles cases while reducing both false positives and false negatives.

$$F1\ Score = 2 * \frac{Precision\ X\ Recall}{Precision + Recall} \qquad (7)$$

Table.1 Performance Evaluation

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	82.4	80.2	78.9	79.5
Random Forest	88.7	87.5	85.9	86.7
Support Vector Machine	86.3	84.8	83.6	84.2
Proposed Ensemble Model	92.1	91.3	90.5	90.9

Table 1 presents the performance comparison of models, where the proposed ensemble achieves the highest accuracy and balanced metrics, outperforming Logistic Regression, Random Forest, and SVM in piles disease prediction.

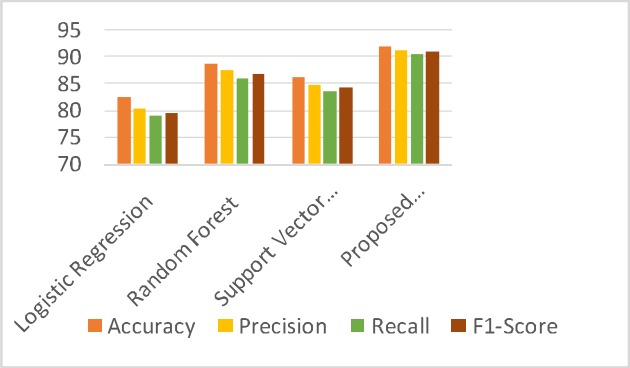


Fig.2 Comparison Graph

Fig. 2 illustrates the comparative performance of all models, highlighting that the proposed ensemble consistently achieves superior accuracy, precision, recall, and F1-score compared to Logistic Regression, Random Forest, and SVM.

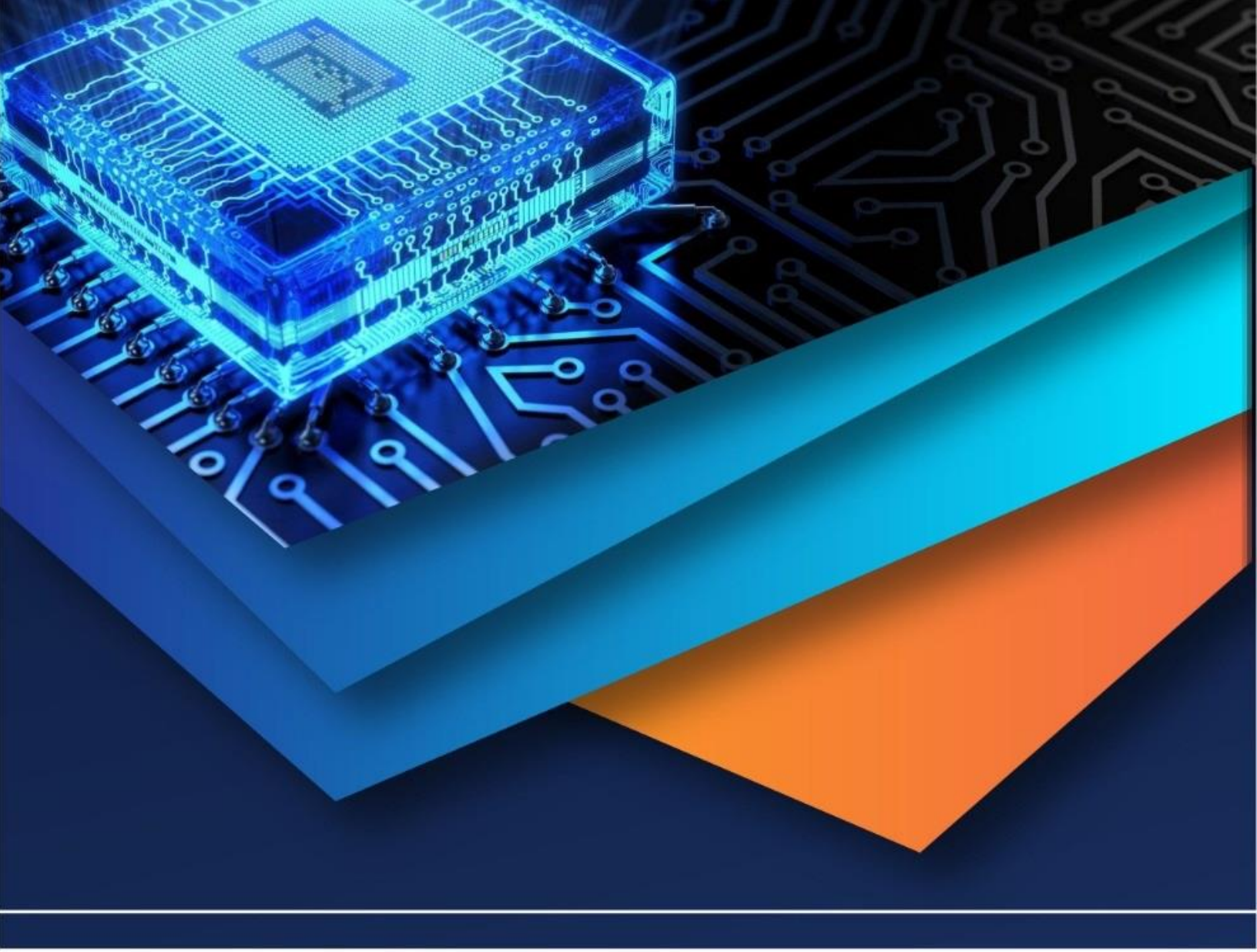
V. CONCLUSION

In conclusion, this study presents an effective machine learning-based approach for the prediction of piles disease, addressing the need for early diagnosis and improved clinical decision support. The proposed framework integrates data preprocessing, feature selection, and multiple classification techniques to analyze clinical and lifestyle-related data. A comparative evaluation of base models, including Logistic Regression, Support Vector Machine, and Random Forest, demonstrates that ensemble learning provides superior predictive performance. The proposed stacking-based ensemble model achieved the highest accuracy of 92.1%, along with improved precision, recall, and F1-score, indicating its robustness and reliability in identifying disease patterns. The enhanced recall highlights the model's ability to correctly identify positive cases, which is critical in medical diagnosis to minimize missed detections. Overall, the results confirm that the integration of machine learning and ensemble techniques can significantly improve the accuracy and effectiveness of disease prediction systems, making the proposed model a valuable tool for supporting healthcare professionals in early detection of piles disease.

The future scope of this study includes expanding the dataset to incorporate a larger and more diverse patient population, which can improve model generalization and robustness. Integration of advanced deep learning techniques, such as neural networks, could further enhance prediction accuracy. Additionally, the system can be extended into a real-time clinical decision-support application, enabling automated screening and early intervention. Incorporating patient history and longitudinal data may also provide deeper insights into disease progression and personalized treatment recommendations.

REFERENCES

[1] Wu, H., & Glangkarn, S. (2024). The prediction model of genetic and risk factors to hemorrhoids (Doctoral dissertation, Mahasarakham University).
[2] Wang, H., Wang, L., Zeng, X., Zhang, S., Huang, Y., & Zhang, Q. (2024). Inflammatory bowel disease and risk for hemorrhoids: a Mendelian randomization analysis. Scientific Reports, 14(1), 16677.
[3] Fathallah, N., Alam, A., Rentien, A. L., La Greca, G., Co, J., Pommaret, E., ... & de Parades, V. (2024). Hemorrhoidal disease: Epidemiological study and analysis of predictive factors for surgical management. Journal of Visceral Surgery, 161(3), 161-166.
[4] Ge, X., Tang, W., & Ni, J. (2025). Hemorrhoids and cardiovascular disease: A bidirectional Mendelian randomization study. Open Medicine, 20(1), 20251256.
[5] Ghosh, J., Taneja, J., & Kant, R. (2024). Predictive Modeling for Rectal Bleeding Risk in Functional Constipation: Integrating Lifestyle Factors and Machine Learning for Targeted Prevention.
[6] Ayo-Oluwaseyi, T. G., Afolalu, S. A., Alade, A. M. I., & Akpor, O. A. (2025). Predictive Analytics Using AI for Identifying Risk Factors and Projecting Disease Burden of Haemorrhoids in Diverse Populations. NIPES JSTR SPECIAL ISSUE, 7(2), 3283-3289.
[7] Zhang, Y., Sun, S., & Han, Z. (2023). Establishment and validation of clinical prediction model for hemorrhoid recurrence after procedure for prolapse and hemorrhoids. Medicine, 102(26), e34062.
[8] Tabbakha, A., Ozsahin, D. U., Uzun, B., & Ozsahin, I. (2021). Selection of Hemorrhoid Treatment Techniques using a Multi-Criteria Decision-Making Technique. In Applied Machine Learning and Multi-Criteria Decision-Making in Healthcare (pp. 245-271). Bentham Science Publishers.
[9] Ghosh, J., Taneja, J., & Kant, R. (2025). Nutritional and lifestyle predictors of rectal bleeding in functional constipation: A machine learning approach. International Journal of Medical Informatics, 201, 105963.
[10] Yavuz, Ö. (2022). A Classification and Clustering Approach Using Data Mining Techniques in Analysing Gastrointestinal Tract. International Scientific and Vocational Studies Journal, 5(2), 254-265.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY