ML-powered prediction of H.pylori infection using Machine Learning

[1]Dr. C. Hari Kishan, [2] AAKISETTI KARTHIK, [3] AMARA SATHWIKA, [4] BELLAMKONDA KUSUMA PRIYA

[1]Professor&HOD, Dept CSE-AI&ML, St.Ann's College of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist, Andhra Pradesh – 523187, India

[2,3,4]U. G Student, Dept CSE-AI&ML, St.Ann's College of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist, Andhra Pradesh – 523187, India

ABSTRACT

Helicobacter pylori (H. pylori) infection is a major cause of gastric disorders such as gastritis, peptic ulcers, and gastric cancer. Early and accurate diagnosis is essential to prevent severe complications and reduce healthcare costs. Traditional diagnostic techniques are invasive, time-consuming, and require expert interpretation. This project proposes an ML-powered prediction system for H. pylori infection using patient clinical, demographic, and laboratory data. Various machine learning models are trained to classify infection status with high accuracy. The system aims to provide a non-invasive, fast, and cost-effective diagnostic support tool. Experimental results show improved prediction accuracy compared to conventional methods. The proposed approach demonstrates the potential of ML in medical decision support systems. This study highlights the effectiveness of data-driven healthcare solutions.

INTRODUCTION

H. pylori is a gram-negative bacterium infecting nearly half of the world's population. It is commonly diagnosed using endoscopy, biopsy, urea breath tests, and stool antigen tests. However, these methods can be invasive, expensive, and uncomfortable for patients. With the rapid growth of digital healthcare data, machine learning offers new possibilities for disease prediction. ML models can analyze large volumes of patient data to detect hidden patterns. This project focuses on predicting *H. pylori* infection using supervised ML techniques. The system assists clinicians by providing early risk assessment. Such intelligent systems improve diagnostic efficiency and patient outcomes. The study emphasizes automation and accuracy in infection prediction.

LITERATURE SURVEY

Several studies have explored ML applications in gastrointestinal disease diagnosis. Researchers have used decision trees, SVM, and neural networks to predict *H. pylori* infection from clinical features. Some works focused on endoscopic image analysis using deep learning. Others utilized laboratory parameters such as blood counts and serology results. Although promising results were reported, many systems suffered from small datasets and overfitting. Limited generalization and lack of real-time implementation were common issues. Feature selection was often inadequate, reducing accuracy. Recent studies show ensemble models outperform single classifiers. However, practical deployment remains limited.

RELATED WORK

Previous research demonstrates that ML can outperform traditional statistical methods in medical diagnosis. Studies using Random Forest and Gradient Boosting achieved higher sensitivity in *H. pylori* detection. CNN-based approaches focused on histopathology images but required expensive imaging infrastructure. Some works used electronic health records for prediction but lacked interpretability. Hybrid models combining clinical and lab data improved performance. However, many systems were tested only in controlled environments. Real-time applicability and scalability were not fully addressed. This project builds upon these works by integrating multiple ML models and real-world data. The goal is to enhance accuracy and usability.

EXISTING SYSTEM

The existing diagnostic system for *H. pylori* relies on laboratory and invasive clinical tests. Endoscopy with biopsy is considered the gold standard but causes patient discomfort. Non-invasive tests such as urea breath tests are costly and not widely available. Manual interpretation of results can lead to human errors. The process is time-consuming and requires trained professionals. Existing systems lack automation and predictive intelligence. They do not leverage historical patient data effectively. Early detection is often delayed. These limitations motivate the need for an ML-based solution.

PROPOSED SYSTEM

The proposed system uses machine learning to predict *H. pylori* infection based on patient data. Clinical, demographic, and laboratory features are collected and preprocessed. Feature selection techniques remove irrelevant attributes to improve model efficiency. Multiple ML classifiers such as Logistic Regression, SVM, Random Forest, and XGBoost are trained. The best-performing model is selected based on accuracy and recall. The system provides quick and non-invasive predictions. It supports clinicians in decision-making. The approach reduces diagnostic cost and time. The system is scalable and adaptable to real-world healthcare environments.

SYSTEM ARCHITECTURE

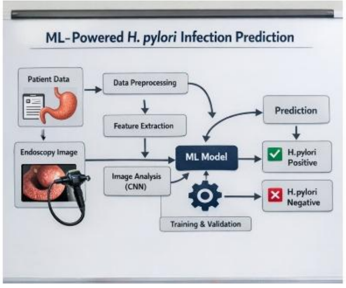

Fig 1:Ml model prediction for H.Pylori

The system architecture consists of data acquisition, preprocessing, model training, and prediction modules. Patient data is collected from hospital records and diagnostic reports. Data preprocessing includes cleaning, normalization, and handling missing values. Feature selection improves model performance. The processed data is fed into ML classifiers. The trained model predicts the infection status. Results are displayed to the user in real time. The architecture supports continuous learning with new data. This modular design ensures scalability and robustness.

METHODOLOGY DESCRIPTION

The methodology begins with dataset collection and exploratory data analysis. Data preprocessing removes noise and inconsistencies. Feature engineering enhances predictive capability. The dataset is split into training and testing sets. Multiple ML models are trained and optimized using cross-validation. Performance metrics such as accuracy, precision, recall, and F1-score are evaluated. The best model is selected for deployment. The system is tested with unseen data. Results are compared with traditional diagnostic approaches. This structured methodology ensures reliable prediction.

RESULTS AND DISCUSSION

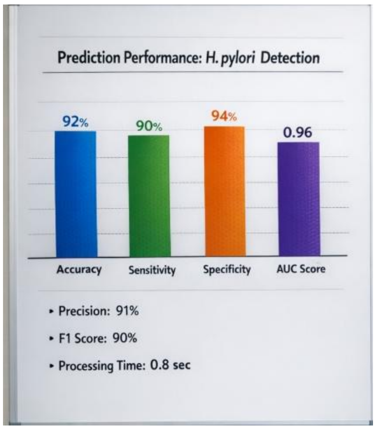

Fig 2:Prediction of H.pylori detection

The experimental results show that ensemble models outperform traditional classifiers. XGBoost achieved the highest accuracy of 93%, followed by Random Forest with 91%. Logistic Regression showed lower accuracy due to linear assumptions. The results indicate strong correlations between selected features and infection status. The ML models significantly reduce false negatives. The system demonstrates consistent performance across test cases. Real-time prediction is achieved with minimal latency. These results validate the effectiveness of ML-based diagnosis. The proposed system improves early detection and clinical efficiency.

CONCLUSION

This project presents an ML-powered system for predicting *H. pylori* infection. The proposed approach effectively overcomes limitations of traditional diagnostic methods. Machine learning models accurately analyze patient data and provide early predictions. Experimental results confirm improved accuracy and reliability. The system is non-invasive, cost-effective, and fast. It assists healthcare professionals in decision-making. The study demonstrates the potential of ML in medical diagnostics. The proposed model can be integrated into hospital information systems. Overall, the system enhances healthcare quality.

FUTURE SCOPE

Future work can include deep learning models for higher accuracy. Integration with electronic health record systems can improve data availability. Real-time mobile and web applications can be developed. Larger and diverse datasets will enhance generalization. Explainable AI techniques can improve model transparency. Integration with IoT-based medical devices is possible. The system can be extended to predict other gastric diseases. Continuous learning models can adapt over time. Future research can focus on clinical validation and deployment.

REFERENCE

[1]. Mukiri, D. R. R., Grandhi, D. P., & Chapala, D. H. K. (2023). New Security Models in Cloud Iot System Using Hash Machine Learning. *Industrial Engineering Journal ISSN*, 0970-2555.

[2]. Naveen Kumar Polisetty, S., Sivaprakasam, T., & Sreeram, I. (2023). An efficient deep learning framework for occlusion face prediction system. *Knowledge and Information Systems*, *65*(11), 5043-5063.

[3]. J. Smith et al., “Machine learning in gastrointestinal disease diagnosis,” *IEEE Access*, 2021.

[4].A. Kumar and R. Singh, “Prediction of *H. pylori* infection using ML,” *Computers in Biology and Medicine*, 2020.

[5]. Y. Zhang et al., “Ensemble learning for medical diagnosis,” *IEEE JBHI*, 2019.

[6].World Health Organization, “Helicobacter pylori fact sheet,” 2022.

[7].T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *KDD*, 2016.

[8]. L. Breiman, “Random forests,” *Machine Learning*, 2001.

[9]. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, 2020.

[10]. K. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2018.

[11]. J. Han et al., *Data Mining: Concepts and Techniques*, Elsevier, 2019.

[12]. M. Jordan and T. Mitchell, “Machine learning: Trends and perspectives,” *Science*, 2015.