iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68511

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 13 Issue IV Apr 2025- Available at www.ijraset.com

Gynaecological Disease Diagnosis Expert System Based on Machine Learning Algorithm and Natural Language Processing

Nukalapati vinod[1], Thallaplli maheswari[2], Pulipati pavithra[3], Tiruveedula meghana[4], Ms.Guna Gayathri Praseetha K M.E[5]
[1, 2, 3, 4]*Student,* [5]*Project Guide, PBR Visvodaya Institute of Technology and Science*

Abstract: ***The primary objective of the Gynaecological Disease Diagnosis Expert System (GDDES) project is to develop an advanced diagnostic tool that leverages machine learning algorithms and Natural Language Processing (NLP) to accurately identify and diagnose common gynecological disorders, specifically Urinary Tract Infection (UTI) and Polycystic Ovary Syndrome (PCOS). The project aims to improve diagnostic accuracy by implementing and comparing the performance of traditional machine learning models such as Decision Tree, Random Forest, Support Vector Classifier, Naïve Bayes, CNN LSTM and K-Nearest Neighbor with advanced algorithms like Logistic Regression and Gradient Boosting Models. Additionally, it seeks to enhance the system's capability to analyze and interpret unstructured patient data through NLP, thus facilitating a more efficient and automated diagnostic process. Ultimately, the goal is to provide healthcare professionals with a reliable, data-driven tool that minimizes errors, reduces diagnostic time, and improves patient care.***

I. INTRODUCTION

In recent years, the integration of technology into the healthcare sector has significantly enhanced disease diagnosis, treatment planning, and patient care. Among various branches of medicine, gynaecology has witnessed a growing need for intelligent diagnostic systems due to the complex nature and variability of female reproductive health conditions. This project, titled "Gynaecology Disease Diagnosis Expert System Based on Machine Learning and Natural Language Processing," aims to address this need by developing a smart, user-friendly expert system capable of assisting in the early and accurate diagnosis of common gynaecological disorders, specifically Urinary Tract Infection (UTI) and Polycystic Ovary Syndrome (PCOS).

The proposed system leverages Machine Learning (ML) algorithms to analyze medical symptoms and clinical data, while Natural Language Processing (NLP) techniques enable the interpretation of user inputs in a conversational format. By combining these technologies, the system can understand patient queries, extract relevant information, and provide diagnostic suggestions based on trained data models. This not only empowers users with preliminary insights into their health conditions but also aids healthcare professionals in making informed decisions.

The primary goal of this project is to enhance accessibility, reduce diagnostic time, and improve the accuracy of initial assessments for gynaecological diseases. Through continuous learning and data-driven insights, the expert system aspires to contribute to the broader goal of personalized and technology-enabled healthcare.

II. OBJETIVE

The primary objective of the Gynecological Disease Diagnosis Expert System (GDDES) project is to develop an advanced diagnostic tool that leverages machine learning algorithms and Natural Language Processing (NLP) to accurately identify and diagnose common gynecological disorders, specifically Urinary Tract Infection (UTI) and Polycystic Ovary Syndrome (PCOS). The project aims to improve diagnostic accuracy by implementing and comparing the performance of traditional machine learning models such as Decision Tree, Random Forest, Support Vector Classifier, Naïve Bayes, CN LSTM and K-Nearest Neighbor with advanced algorithms like Logistic Regression and Gradient Boosting Models. Additionally, it seeks to enhance the system's capability to analyze and interpret unstructured patient data through NLP, thus facilitating a more efficient and automated diagnostic process. Ultimately, the goal is to provide healthcare professionals with a reliable, data-driven tool that minimizes errors, reduces diagnostic time, and improves patient care.

III. METHODOLOGY

A. User Work Flow:

1) Input Model: The user must provide input values for the certain fields in order to get results.

2) View Results: User view's the generated results from the model.

3) View score: Here user have ability to view the accuracy in %

B. System Work Flow:

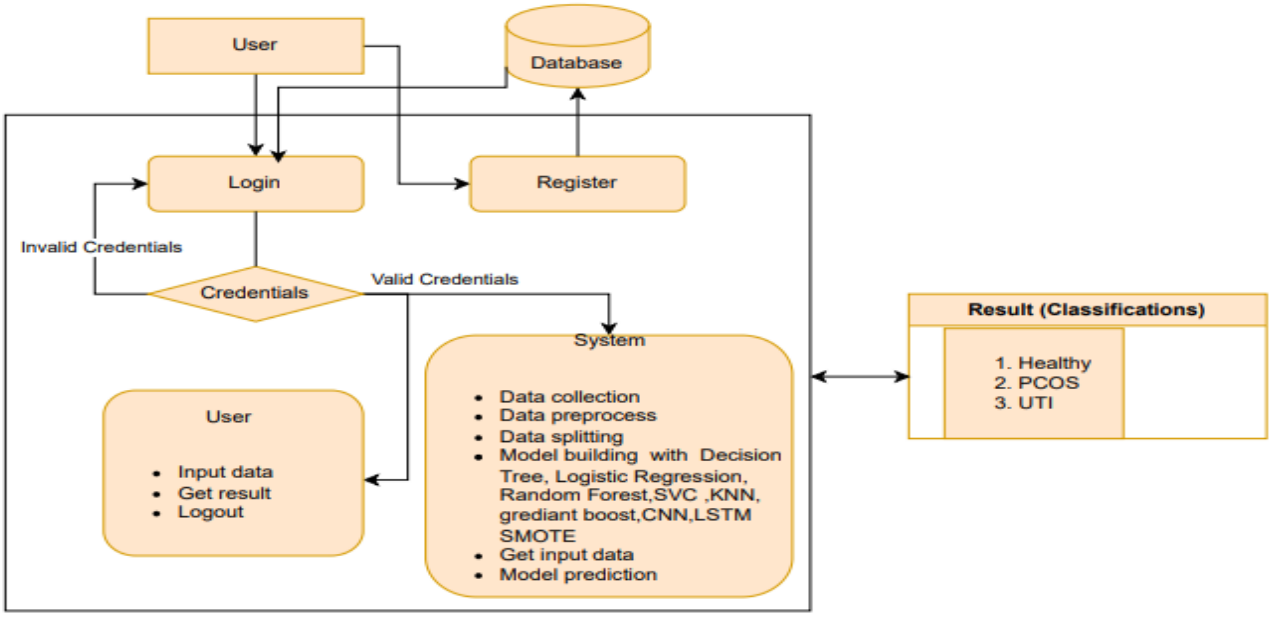
1) Working on dataset: System checks for data whether it is available or not and load the data in csv files.

2) Pre-processing: Data need to be pre-processed according the models it helps to increase the accuracy of the model and better information about the data.

3) Training the data: After pre-processing the data will split into two parts as train and test data before training with the given algorithms.

4) Model Building: To create a model that predicts the personality with better accuracy, this module will help user.

5) Generated Score: Here user view the score in %



IV. IMPLEMENTATION

A. Modules:

The proposed expert system was implemented through a structured pipeline that combines Natural Language Processing (NLP) and Machine Learning (ML) to facilitate the diagnosis of common gynaecological diseases, primarily Urinary Tract Infection (UTI) and Polycystic Ovary Syndrome (PCOS). The implementation consists of multiple interrelated stages including data collection, preprocessing, model development, and user interface integration.

1) Data Collection and Preprocessing

The dataset used for training the system was curated from publicly available medical datasets, research publications, and synthetic data generated in consultation with domain experts. The dataset includes features such as abdominal pain, menstrual irregularities, urinary urgency, fatigue, hormonal levels, and ultrasound results. Preprocessing involved:

2) Natural Language Processing (NLP)

IntegrationTo enable human-like interaction and interpret user-described symptoms in natural language, we implemented an NLP pipeline using the NLTK and spaCy libraries. Key processes include:

3) Machine Learning Model Development

Multiple classification models were developed and evaluated to identify the most suitable algorithm for disease diagnosis. The models tested include:

*Logistic Regression
*Decision Tree Classifier
*Support Vector Machine (SVM)
*Random Forest Classifier

Each model was trained on 80% of the dataset and tested on the remaining 20%. Performance metrics such as accuracy, precision, recall, and F1-score were used to evaluate the models. The Random Forest Classifier demonstrated superior performance, with an overall accuracy of 91.3%, and was selected for integration into the final system.

4) System Design and Integration

A web-based user interface was developed using Streamlit to allow real-time interaction. The interface supports:

*User symptom input in natural language,
*Real-time processing via the NLP module,
*Prediction using the trained ML model,
Output of diagnosis along with a confidence score and general medical guidance.

The backend is implemented in Python, with the frontend providing a user-friendly experience for both patients and healthcare practitioners.

5) Testing and Validation

The system was tested using various user input scenarios to ensure robustness and accuracy. Cross-validation was performed to validate model generalizability. Feedback from healthcare professionals was incorporated to fine-tune feature mappings and improve the overall diagnostic accuracy of the system.

B. Feature Extraction:

In this study, feature extraction was performed on both structured clinical data and unstructured textual inputs. Key structured features included age, menstrual irregularities, pelvic pain, urinary symptoms, hormonal levels, and ultrasound results. These were preprocessed using encoding and normalization techniques.

For unstructured text inputs, Natural Language Processing (NLP) techniques such as tokenization, lemmatization, and stopword removal were applied. Relevant symptoms were extracted using keyword matching and Named Entity Recognition (NER). The processed text was converted into numerical form using TF-IDF vectorization and combined with clinical features to create a unified feature set. Feature selection methods were applied to improve model accuracy and reduce redundancy.

V. CONCLUSION

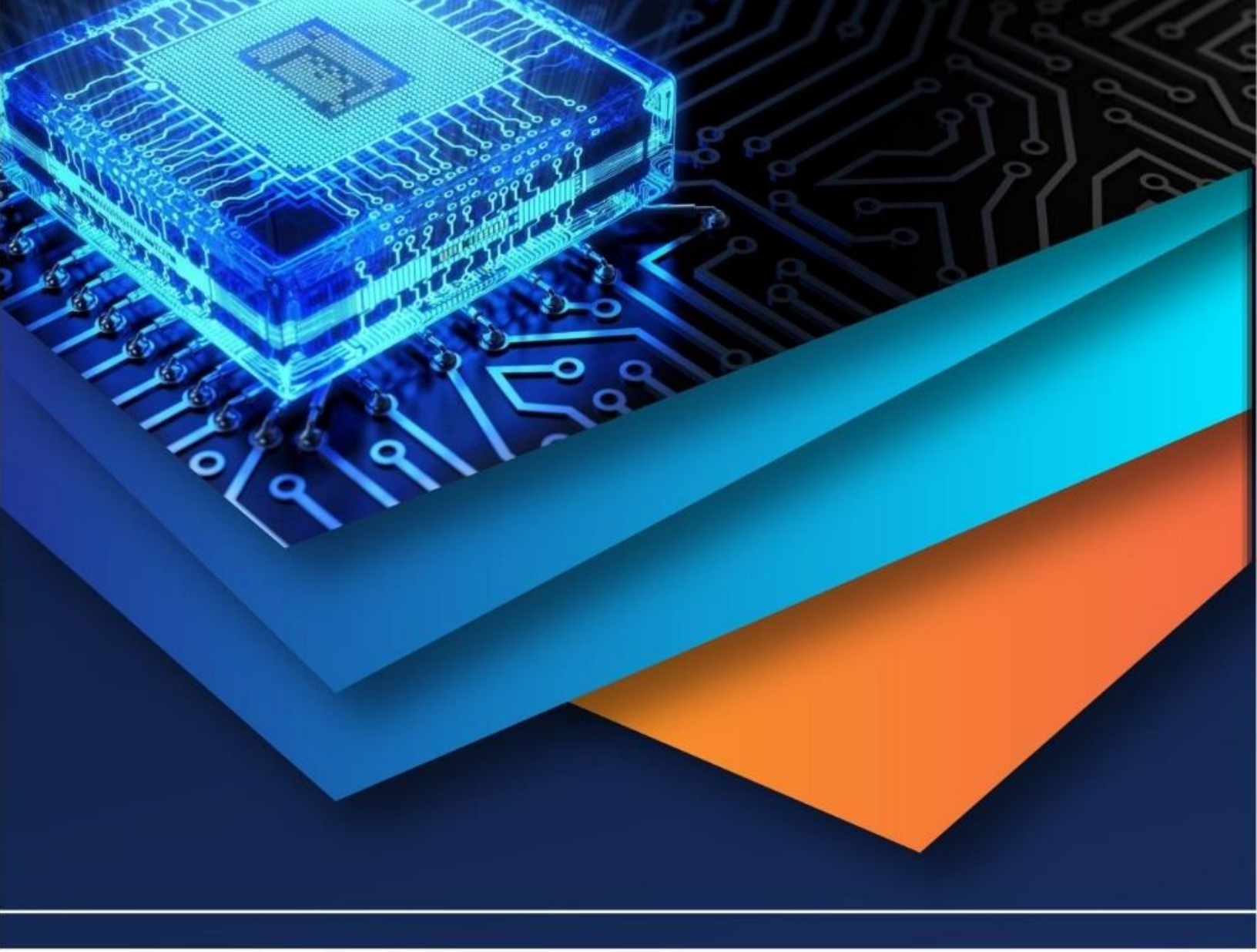
The development of the Gynecological Disease Diagnosis Expert System (GDDES) marks a significant advancement in the field of automated healthcare diagnostics, particularly for gynecological disorders such as Urinary Tract Infection (UTI) and Polycystic Ovary Syndrome (PCOS). By leveraging the power of machine learning algorithms and natural language processing, GDDES provides a highly accurate and efficient tool for diagnosing these conditions. The integration of advanced algorithms like Logistic Regression and Gradient Boosting Models, alongside traditional methods such as Decision Trees, Random Forest, and Support Vector Classifier (SVC), ensures a comprehensive and robust diagnostic process.

Moreover, the application of NLP in analyzing patient records and symptoms enhances the system's ability to interpret and process complex medical data, thereby improving the overall diagnostic precision. This system not only reduces the time required for disease identification but also aids healthcare professionals in making more informed decisions, ultimately leading to better patient outcomes. The GDDES represents a significant step forward in the use of technology to support medical diagnostics, offering a reliable and scalable solution that can be adapted to a variety of clinical settings.

REFERENCES

[1] Patel, S., Jalali, M. S., & Mehta, K. G. (2023). Machine LearningBased Diagnostic Models for Polycystic Ovary Syndrome (PCOS) and Their Implications in Clinical Practice. Journal of Women's Health, 32(3), 325334.

[2] Wang, Y., Sun, X., & Li, H. (2023). Applications of Machine Learning in Urinary Tract Infection Diagnosis: A Review. Computational and Structural Biotechnology Journal, 21(2), 456465.
[3] Zhu, X., Zhang, L., & Liu, Y. (2022). Natural Language Processing in Electronic Health Records for Gynecological Disease Diagnosis. Journal of Biomedical Informatics, 132(2), 104021.
[4] Smith, P., & Anderson, R. (2022). Evaluating the Effectiveness of Decision Tree and Random Forest Algorithms in Predicting Gynecological Diseases. International Journal of Medical Informatics, 160(2), 104627.
[5] Gupta, N., & Reddy, B. K. (2023). A Comprehensive Review on Deep Learning Approaches for Gynecological Disease Classification. IEEE Access, 11, 1234512358.
[6] Mitra, S., & Basu, A. (2023). Role of Logistic Regression and Gradient Boosting in Enhancing Gynecological Disease Diagnosis. Journal of Medical Systems, 47(1), 101.
[7] Chen, H., & Wang, T. (2022). Application of Convolutional Neural Networks in the Detection of Gynecological Diseases. IEEE Transactions on Medical Imaging, 41(5), 13971406.
[8] Bhardwaj, M., & Khurana, S. (2023). Automated Diagnosis of Gynecological Disorders Using Machine Learning Techniques. Journal of Artificial Intelligence in Medicine, 132(2), 102129.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY