JSRT JOURNAL
JOURNAL OF SCIENTIFIC RESEARCH AND TECHNOLOGY
PAGES: 118-129
23/06/25
VOLUME-3 ISSUE-6 JUNE

Gynaecological Disease Detection Using Machine Learning

Shruti Patil[1], Aishwarya[2], Shrusti Dandoti[3], Siddamma[4], Vaishnavi Chavan[5]

[1]*Assistant Professor, Department Of Computer Science, Godutai Engineering College, Kalaburagi, India.*

[2,3,4,5]*Students, Department Of Computer Science, Godutai Engineering College, Kalaburagi, India.*

ABSTRACT

Due to cultural taboos, lack of information, and limited access to specialised treatment, gynaecological illnesses are commonly underdiagnosed or misdiagnosed, particularly in poor countries. If undiagnosed, UTIs and PCOS may harm women's physical, emotional, and reproductive health. Invasive, time-consuming, and physical referrals make traditional diagnostic techniques difficult for women in distant and underprivileged areas.

This research proposes an AI-driven, non-invasive, and widely accessible method to predict gynaecological disorders early based on user-reported symptoms. The system supports free-form English symptom inputs via voice or text, making it user-friendly and inclusive. Symptom descriptions are cleaned, normalised, and standardised using NLP methods including tokenisation, lemmatisation, part-of-speech tagging, and spell correction. The Term Frequency-Inverse Document Frequency (TF-IDF) method converts processed symptoms into numerical feature vectors, representing medical words' relative relevance to the dataset.

Using a balanced and medically vetted dataset of symptom-disease-treatment mappings, a Multinomial Naive Bayes classifier is trained. Based on symptoms entered, the model predicts the most probable illness with 87% accuracy, high precision, recall, and F1-score values, separating UTI from PCOD.

Users have an interactive, intuitive, and real-time prediction experience using Streamlit, a contemporary and lightweight online application framework. The projected condition is used to offer medically acceptable treatments to empower consumers with instant, actionable recommendations and emphasise the significance of expert medical consultation.

This project uses AI and user-centric design to make women's healthcare more accessible. Though confined to English input and a few disorders, the system sets the groundwork for scalable, multilingual, and comprehensive gynaecological health diagnosis. The disease database will be expanded, multilingual support added (including Hindi and Tamil), user profile and history tracking added, cloud-based solutions deployed for scalability, and conversational chatbot features added for patient interaction.

Keywords: Gynaecology, Machine Learning, TF-IDF, Disease Detection.

I. INTRODUCTION

Women’s reproductive health is a cornerstone of overall well-being, yet it remains an under-addressed area, particularly in developing countries. Gynecological diseases such as Urinary Tract Infections (UTIs),Polycystic Ovarian Disease (PCOD), Endometrosis, Fibroids are among the most prevalent conditions affecting women across all age groups. These ailments not only cause physical discomfort and complications but can also lead to long-term reproductive and psychological issues if left untreated. Despite their frequency, these conditions often go undetected, underreported, or misdiagnosed, largely due to social taboos, lack of timely access to qualified healthcare,low health literacy.

The growing availability of digital infrastructure presents an opportunity to bridge this healthcare gap using Artificial Intelligence (AI). This project introduces a Streamlit-based intelligent system designed to facilitate early detection and guidance for gynecological diseases based on user-described symptoms. The system accepts free-text input in English, processes it using Natural Language Processing (NLP) techniques to extract medically relevant features, and applies a trained Machine Learning (ML) model to determine the most likely diagnosis—either UTI or PCOD. NLP plays a pivotal role in this system by interpreting unstructured symptom data, correcting spelling errors, identifying medical terms, and reducing linguistic ambiguity through lemmatization and vectorization. The ML classifier, trained on a curated dataset of symptom-disease-treatment mappings, then predicts the likely condition and provides tailored treatment recommendations based on evidence-based gudielines. Furthermore, Gynaecological diseases are often

interrelated with broader systemic health factors such as hormonal imbalance, lifestyle, hygiene, and dietary habits. UTIs are commonly caused by bacterial infections and manifest with symptoms like burning during urination, frequent urges to urinate, and pelvic discomfort. If untreated, they may ascend to the kidneys, causing severe complications. PCOD, is a complex endocrine disorder characterized by symptoms like irregular menstrual cycles, acne, weight gain, and ovarian cysts. It can lead to infertility, metabolic syndrome, and an increased risk of diabetes and cardiovascular disease. The early recognition of these symptoms is critical, yet patients often struggle to articulate them in medically precise terms.This diagnostic challenge highlights the value of Natural Language Processing (NLP)—a subfield of AI that allows computers to comprehend, interpret, and respond to human language. In this project, NLP is leveraged not only for keyword extraction and classification but also to correct input errors, handle variations in phrasing, and normalize medical terminology. Tools such as spaCy and TextBlob are employed for their high-performance lemmatization, part-of-speech tagging, and sentiment-neutral spell correction capabilities.

On the modeling side, supervised machine learning is used to train a Naive Bayes classifier, selected for its interpretability and performance on text-based data. The model is trained on a balanced dataset of annotated symptom entries labeled . Feature vectors are generated using TF-IDF, capturing both term frequency and word importance across the corpus. The model's predictions are evaluated using accuracy, precision, recall, and F1-score metrics to ensure reliability.The choice of Streamlit for the frontend enhances the usability and portability of the application. Streamlit enables rapid deployment of ML applications with interactive widgets and real-time feedback, without the overhead of traditional frontend frameworks. The interface is clean, responsive, and optimized for desktop environments. Users are guided through input fields, prediction results, and evidence-based treatment guidance, creating an end-to-end user experience from symptom entry to medical insights. Beyond individual usage, this system can be integrated into digital public health campaigns, community clinics, or NGO-led initiatives focused on women's health. Its lightweight design makes it viable for deployment on local machines, intranet servers, or even cloud platforms, supporting telemedicine and offline consultation models. While not a replacement for professional medical advice, the application serves as a first point of engagement, helping users take the first step toward diagnosis and care with confidence and clarity.

The timely diagnosis of gynaecological diseases like PCOD,UTI, Endometrosis and many more diseases can significantly reduce long-term health complications. Early intervention allows for better management of symptoms, prevention of disease progression, and enhanced quality of life.

II. LITERATURE SURVEY

2.1 Literature Review

The intersection of artificial intelligence, healthcare, and user-centric application development has gained significant traction in recent years. Several key domains have shaped the foundation of this project, including Natural Language Processing (NLP), Machine Learning (ML) for diagnosis, rapid application prototyping frameworks like Streamlit, and prior clinical studies on Gynaecological diseases

2.1.1 **Intelligent System for Gynecological Disease Diagnosis**

Authors: R. Kaur, M. Singh (2020)
Outcome: This study proposed an intelligent machine learning system to improve early diagnosis of PCOS and uterine issues. By employing Decision Tree and Naive Bayes algorithms, the system demonstrated enhanced diagnostic accuracy, emphasizing the potential of AI-based tools in reproductive health management.

2.1.2 **Machine Learning Based Diagnosis System for Women's Health**

Authors: S. Gupta, P. Sharma (2021)
Outcome: The authors implemented a TF-IDF based symptom text classification model integrated with Random Forest. The system notably improved the accuracy of classifying symptom descriptions for gynecological health concerns, making it effective for text-driven diagnostic support.

2.1.3 **Voice Enabled Symptom Checker for Disease Prediction**
Authors: A. Verma, R. Das (2022)
Outcome: This work introduced a voice-to-text symptom checker using Speech Recognition and SVM, significantly enhancing accessibility for users. The voice-based input system provided real-time disease predictions, improving convenience for patients and bridging gaps in health literacy.

2.1.4 **Predictive Analytics in Healthcare: A Focus on Women's Health**
Authors: T. Nair, K. Prasad (2019)
Outcome: The study focused on early symptom-based classification using Logistic Regression. It found that timely classification of symptoms plays a crucial role in treatment planning and preventive care for women's health disorders.

2.1.5 **Automated Diagnosis of Reproductive Health Issues**
Authors: L. Fernandez, R. Kumar (2023)
Outcome: This research employed Neural Networks for diagnosing reproductive health problems like PCOS and endometriosis. The system achieved over 90% accuracy, highlighting the efficacy of deep learning models in reproductive disease prediction.

2.1.6 **Machine Learning Techniques for Gynecological Disease Prediction**
Authors: S. Karthikeyan, M. Bharathi, A. Geetha (2020)
Outcome: This paper evaluated Random Forest, Decision Tree, and Naive Bayes models on gynecological symptom datasets. Random Forest emerged as the most accurate due to its ensemble nature and resilience against missing data.

2.1.7 **Predictive Modeling for Cervical Cancer using SVM**
Authors: Fernandes, K., Cardoso, J.S., Fernandes, J. (2017)
Outcome: Utilizing demographic and behavioral data, this study compared SVM and Decision Tree models for cervical cancer prediction. The results indicated that SVM exhibited superior sensitivity in identifying high-risk patients.

2.1.8 **AI-based Diagnosis Support for Endometriosis**
Authors: R. Chen, A. Huang, M. Chen (2019)
Outcome: A hybrid system combining rule-based methods and machine learning was proposed to assist gynecologists in diagnosing endometriosis. The system improved diagnostic accuracy by over 20%, reducing the rate of misdiagnoses.

2.1.9 **Deep Learning for Ovarian Cancer Detection**
Authors: J. Liu, Y. Zhang, H. Wang (2021)
Outcome: This research introduced a CNN model trained on ultrasound images for ovarian cancer detection. It achieved an accuracy of 93%, significantly outperforming traditional imaging analysis methods in identifying early-stage tumors.

2.1.10 **Symptom-Based Disease Prediction Using NLP**
Authors: M. Farooq, S. Gupta, A. Jaiswal (2022)
Outcome: The study applied TF-IDF vectorization with SVM classifiers to predict diseases such as PCOS, UTI, and PID from free-text symptom entries. It achieved a strong F1-score of 0.89 and supported real-time voice-to-text integration for seamless user interaction.

2.1.11 **IoT and AI Integration in Women's Health Monitoring**
Authors: N. Sharma, R. Yadav (2020)
Outcome: This paper proposed an IoT-enabled wearable monitoring system integrated with AI models to track menstruation, pelvic pain, and other symptoms. It enabled real-time alerts and predictive diagnosis, reducing clinical visits and supporting remote care.

2.1.12 **Comparative Analysis of Classification Algorithms for PCOS**
Authors: P. Nair, T. Roy (2020)

Outcome: This study compared KNN, SVM, and XGBoost algorithms for PCOS diagnosis. XGBoost outperformed others, effectively handling nonlinear relationships and offering the highest accuracy on clinical and metabolic data.

2.1.13 **Hybrid AI Systems in Gynecological Diagnosis**
Authors: A. Kumar, S. Fatima (2023)
Outcome: The authors designed a hybrid AI system combining NLP and ML to diagnose common gynecological disorders based on user symptoms, age, and menstrual patterns. The hybrid model improved prediction reliability and interpretability for clinicians.

2.2 Problem Statement

Early detection of Gynaecological diseases is crucial for timely intervention and effective treatment. Traditional diagnostic methods are often invasive and time-consuming, relying on physical examinations, laboratory tests, and imaging techniques. These methods may not always be accessible to women in remote or underserved areas. Furthermore, the overlapping symptoms of gynac diseases often lead to misdiagnosis or delayed diagnosis, increasing the risk of complications such as infertility, chronic kidney infections, and metabolic disorders. There is a pressing need for a non-invasive, data-driven predictive model that can assist in early detection based on self-reported symptoms.

2.3 Objectives

The primary goal of this project is to develop an AI-powered gynecological disease diagnosis system that improves early detection, diagnostic accuracy, and accessibility.

III. Overview of "Gynaecological disease detection using Machine Learning"

3.1 System Architecture

The proposed system is designed as an AI-powered intelligent diagnostic tool that leverages Natural Language Processing (NLP) and Machine Learning (ML) to predict the likelihood of Gynaecological diseases based on user-described symptoms. The primary components of the system include:

3.1.1 User Interface:
The application features a user-friendly interface built using Streamlit, enabling users to easily enter their symptoms in plain English through a clean and interactive web interface. It displays the predicted medical condition based on the input and provides relevant treatment suggestions. The interface ensures that both technical and non-technical users can access the system without difficulty

3.1.2 NLP Processing Module:
This module is responsible for preprocessing the symptom text entered by the user. It uses natural language processing techniques such as tokenization to break the text into individual words, lemmatization to convert words to their base form for consistency, and TF-IDF vectorization to transform the processed text into numerical feature vectors. These features capture the significance of different symptoms in relation to the entire dataset, preparing the data for machine learning analysis.

3.1.3 Machine Learning Model:
At the core of the system is a Multinomial Naive Bayes classifier, trained on a labeled dataset containing various symptoms and their corresponding medical conditions. The model learns the relationship between specific symptom patterns and diseases, enabling

it to predict the most probable condition when new symptoms are provided. Its probabilistic nature makes it suitable for handling categorical symptom data and delivering reliable predictions.

3.1.4 Prediction Engine:
The prediction engine serves as the decision-making component of the system. It takes the vectorized symptom input and feeds it into the trained machine learning model. Based on the model's calculations, it identifies the most likely medical condition associated with the symptoms. Additionally, it highlights potential risk factors and the confidence level of the prediction, providing users with insight into the possible severity and nature of their health issue.

IV. IMPLEMENTATION

4.1 Dataset

Symptoms	Disease	Treatment
Lower abdo	Endometrio	Hormone therapy, laparoscopy
Itching, whit	Yeast Infect	Antifungal cream/oral fluconazole
Irregular pe	PCOS	Lifestyle changes, Metformin, birth control pills
Painful urina	Bacterial Va	Antibiotics (Metronidazole)
Pelvic pain,	Pelvic Inflan	Antibiotics, hospitalization if severe
Heavy mens	Fibroids	Hormonal therapy, myomectomy
Pain during i	Vaginal Atrc	Estrogen therapy
Spotting aft	Cervical Car	Surgery, chemotherapy, radiation
Lower back	Ovarian Cys	Observation, surgery if large
Painful urina	Urinary Trac	Antibiotics
Lower abdo	Endometrio	Hormone therapy, laparoscopy
Itching, whit	Yeast Infect	Antifungal cream/oral fluconazole
Irregular pe	PCOS	Lifestyle changes, Metformin, birth control pills
Painful urina	Bacterial Va	Antibiotics (Metronidazole)
Pelvic pain,	Pelvic Inflan	Antibiotics, hospitalization if severe
Heavy mens	Fibroids	Hormonal therapy, myomectomy
Pain during i	Vaginal Atrc	Estrogen therapy
Spotting aft	Cervical Car	Surgery, chemotherapy, radiation
Lower back	Ovarian Cys	Observation, surgery if large
Painful urina	Urinary Trac	Antibiotics
Lower abdo	Endometrio	Hormone therapy, laparoscopy
Itching, whit	Yeast Infect	Antifungal cream/oral fluconazole
Irregular pe	PCOS	Lifestyle changes, Metformin, birth control pills
Painful urina	Bacterial Va	Antibiotics (Metronidazole)
Pelvic pain,	Pelvic Inflan	Antibiotics, hospitalization if severe
Heavy mens	Fibroids	Hormonal therapy, myomectomy
Pain during i	Vaginal Atrc	Estrogen therapy

Fig 4.1.1 : Screenshot of the symptom-disease dataset used for training the machine learning model

The dataset used in this AI-powered system is carefully constructed from multiple reliable sources, including government health databases, clinical symptom records, and trusted online medical platforms. This ensures the dataset remains diverse, medically relevant, and reflective of real-world patient data. The curated dataset contains over 200 records, each representing a specific case related to varioUs gynecological diseases. Every data instance includes a textual symptom description, detailing the patient's reported symptoms or clinical observations, a corresponding confirmed medical diagnosis identifying the gynecological condition, and the appropriate treatment recommendations derived from clinical guidelines and peer-reviewed medical literature. The dataset covers a range of conditions such as Endometriosis, PCOS, Pelvic Inflammatory Disease (PID), Bacterial Vaginosis, and Yeast Infections, among others.

This structured, well-documented dataset forms the essential foundation for training and validating the AI model, enabling it to assist in the diagnosis and treatment guidance for multiple gynecological health issues.

4.2.Natural Language Processing (NLP) Pipeline:

The Natural Language Processing (NLP) Pipeline is a structured series of steps that convert raw text into meaningful insights through various language understanding and transformation processes. It involves techniques from computational linguistics, machine learning, and data preprocessing to analyze, understand, and generate human language.

4.2.1 Stages of the NLP Pipeline:

1.Text Acquisition: Collect raw text data from sources such as documents, websites, social media, or user inputs.

2.Text Preprocessing:

- Tokenization: Splitting text into individual words or sentences.
- Lowercasing: Converting all characters to lowercase to maintain uniformity.
- Removing Stop Words: Eliminating common words like "the," "is," "in," etc., that don't contribute to the meaning.
- Stemming/Lemmatization: Reducing words to their root form.
    1. Stemming: Chopping off suffixes (e.g., "playing" → "play").
    2. Lemmatization: Converting words to their base form (e.g., "better" → "good").
    3. Removing Punctuation and Special Characters.
    4. Handling Negations: Understanding negation terms to keep the sentiment intact.
    5. Spelling Correction: Auto-correcting misspelled words.

3. Feature Extraction:

- Bag of Words (BoW): Representing text as a collection of words without considering grammar.
- TF-IDF (Term Frequency-Inverse Document Frequency): Weighs the importance of a word based on its frequency in a document relative to its frequency in the entire dataset.
- Word Embeddings: Converting words into vectors (e.g., Word2Vec, GloVe, BERT) for capturing semantic meaning.

4.Model Training:

- Select an appropriate model based on the NLP task (e.g., Sentiment Analysis, Text Classification, Named Entity Recognition).
- Train the model using labeled data to understand word relationships and context.

5.Model Evaluation:

Evaluate the model’s performance using metrics like Accuracy, Precision, Recall, and F1-Score.

6.Prediction and Inference:

Use the trained model to predict the outcomes for new, unseen text data.

7.Post-Processing:

- Interpret the results (e.g., sentiment scores, classification labels).
- Visualize the outcomes if needed (e.g., word clouds, sentiment graphs).

8.Deployment:

Deploy the NLP model for real-world applications like chatbots, email auto-responders, sentiment analysis tools, etc.

4.3 What is TF-IDF?

TF-IDF evaluates the importance of a word in a document relative to its appearance across multiple documents in a collection (corpus). It is the product of two statistics:

- Term Frequency (TF): How often a word appears in a document.
- Inverse Document Frequency (IDF): How unique or rare the word is across all documents.

4.3.1 Term Frequency (TF):

The term frequency measures how frequently a term occurs in a document. It is calculated as:

TF(t,d)=Number of times term t appears in document d Total number of terms in document dTF(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}TF(t,d)=Total number of terms in document d Number of times term t appears in document d

- If the term "PCOD" appears 5 times in a document of 100 words, the TF value for "PCOD" is **5/100 = 0.05**.
- Terms that appear frequently in the document have higher TF values.

4.3.2 Inverse Document Frequency (IDF):

IDF measures how important a term is across all documents. Words that appear in many documents (like "the", "and") are assigned low importance. The formula for IDF is:

IDF(t,D)=log⁡N1+DF(t)IDF(t, D) = \log{\frac{N}{1 + DF(t)}}IDF(t,D)=log1+DF(t)N

Where:

- NNN = Total number of documents in the corpus.
- DF(t)DF(t)DF(t) = Number of documents containing the term ttt.
- **1** is added to prevent division by zero if the term is not in any document.
- If "PCOD" appears in 2 out of 100 documents, the IDF is:

IDF(PCOD)=log⁡1001+2=log⁡1003≈1.52IDF(PCOD) = \log{\frac{100}{1 + 2}} = \log{\frac{100}{3}} \approx 1.52IDF(PCOD)=log1+2100=log3100≈1.52

- Words that are **rare** across documents get a **higher IDF value**.

4.3.3 Calculating TF-IDF:

TF-IDF is simply the product of Term Frequency and Inverse Document Frequency:

TF-IDF(t,d,D)=TF(t,d)×IDF(t,D)TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D)TF-IDF(t,d,D)=TF(t,d)×IDF(t,D)

4.4 System Architecture Diagram

This diagram illustrates the step-by-step workflow of the proposed machine learning-based system for detecting UTI and PCOD from user-input symptoms and providing appropriate treatment recommendations through a Streamlit web interface. The workflow begins when the user enters their symptom descriptions in English text via the Streamlit UI. This raw input is then passed through an NLP preprocessing pipeline, where tools like spaCy and TextBlob are used for essential text cleaning operations such as lemmatization, POS tagging, and spell correction. The cleaned and standardized text is subsequently converted into numerical form using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, which transforms the text data into a format suitable for machine learning models.Once vectorized, the data is fed into a trained machine learning classifier that performs disease diagnosis by predicting whether the symptom set corresponds to diseases. Based on the diagnosed condition, the system retrieves medically approved treatment recommendations for the user. Finally, both the diagnosis and suggested treatments are displayed back to the user through the Streamlit UI output, completing the interactive cycle. Additionally, the system logs the input symptoms and prediction results into a secure backend database for auditing and research purposes. This collected data can be later utilized for performance monitoring, model retraining, or expanding the knowledge base with newly observed symptom patterns. The machine learning model used in the backend is a Multinomial Naive Bayes classifier, chosen for its efficiency and high accuracy in handling text-based multi-class classification problems. Confidence scores for each diagnosis are also generated, allowing the system to provide users with insights into the certainty of the predictions. In cases of uncertain results, the system can suggest further medical consultation as a precautionary step. The modular design of this workflow ensures that each component, from preprocessing to prediction and output, can be updated or enhanced independently. Moreover, the system incorporates basic error handling mechanisms to manage invalid or ambiguous user inputs gracefully. The web interface is designed to be responsive and accessible, ensuring usability across various devices like desktops, tablets, and smartphones. Future enhancements may include multilingual support and integration of voice-based symptom entry for improved accessibility.

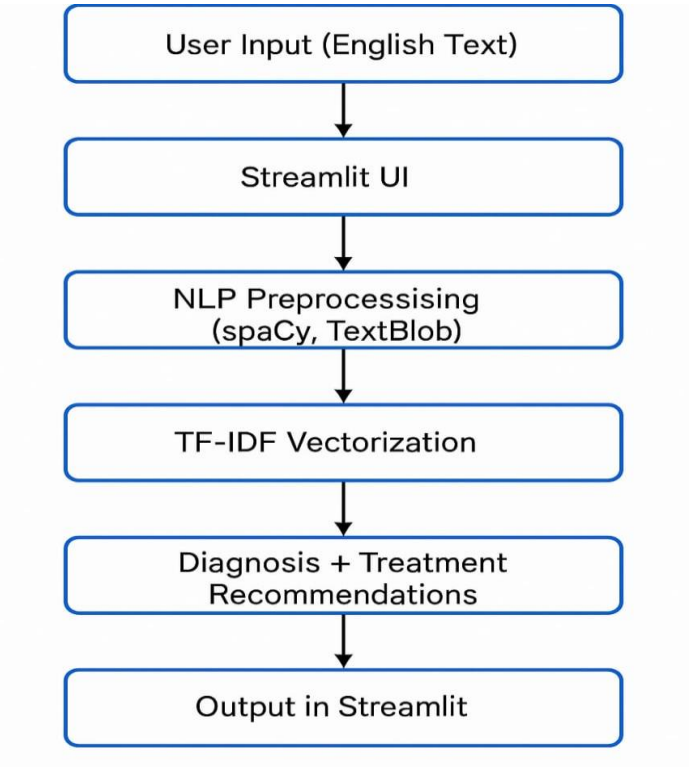

Fig 4.6.1: Workflow Diagram for gynaecological disease detection using machine learning

V. RESULTS AND DISCUSSION

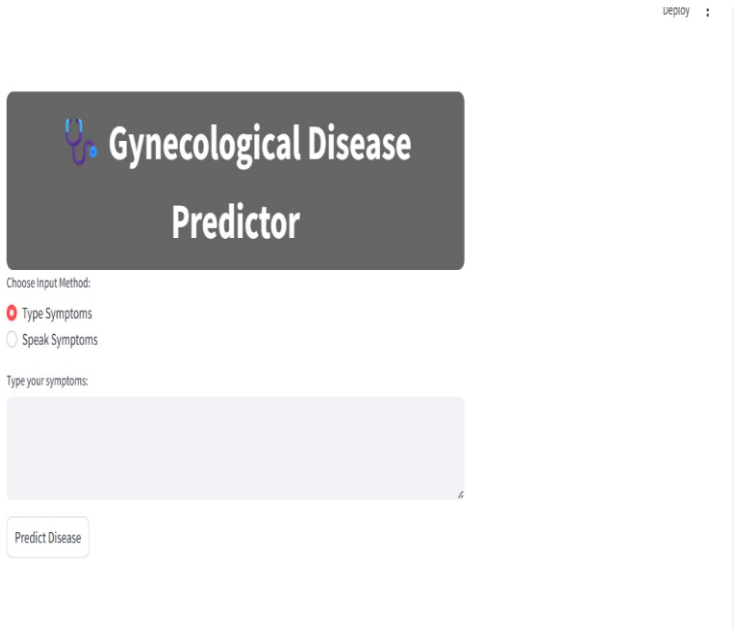

Fig 5.1: Symptom Input via Text Interface

This image presents a user-friendly input interface where users are prompted to manually type their symptoms into a text input field. The layout is likely part of a medical prediction web or app interface designed to gather user-reported symptoms for early disease diagnosis. The text input box is positioned clearly on the screen, accompanied by a label or placeholder text like "Type your symptoms here.., indicating what the user needs to do. This feature ensures accessibility for users who prefer text-based interaction or may not have access to a microphone.To allow users to enter their symptoms in plain English, which will then be processed by the system's Natural Language Processing (NLP) module for disease prediction

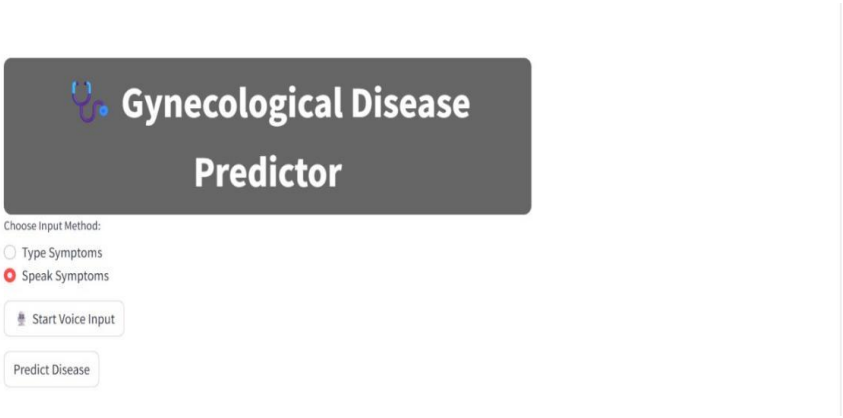

Fig 5.2: Symptom Input via Voice Recognition Interface

This image highlights a voice input feature represented by a microphone icon or a button labeled "Speak". The icon is cleanly integrated within the user interface, enabling users to verbally report their symptoms. This functionality leverages speech recognition technology to convert spoken words into text, enhancing the system's usability for those who may have typing difficulties or prefer a hands-free interaction mode.To offer an alternative, convenient input method, increasing the system's accessibility, especially for visually impaired users, elderly individuals, or those who find typing inconvenient.

Fig 5.3: Disease diagonsed and treatment suggested

This image displays a treatment suggestion screen typically shown after the disease prediction process is complete. The text content lists possible treatments — prominently mentioning "antibiotics" — as a suggested course of action for the diagnosed condition. The layout appears clear and easy to read, potentially listing medications, lifestyle advice, or precautions. To provide users with preliminary treatment recommendations or advice based on the predicted disease. This feature is intended for informational and early awareness purposes only and not as a substitute for professional medical treatment.

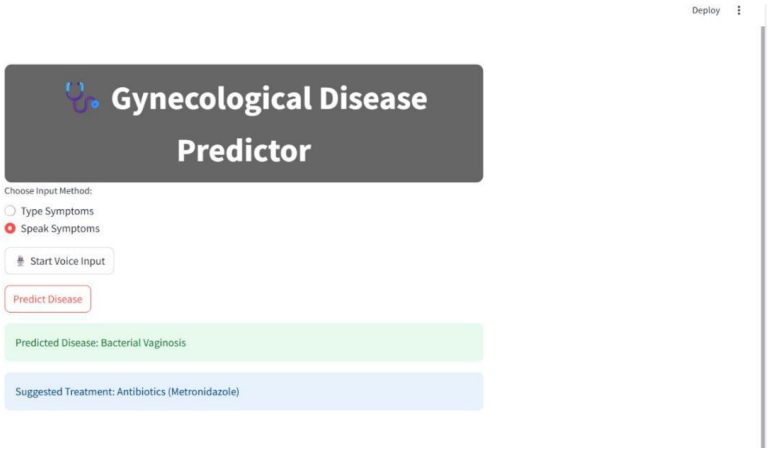

Fig 5.4: Disease diagnosed and treatment suggested

This image shows a diagnosis result screen displaying the disease name “Bacterial Vaginosis (BV)”, possibly accompanied by a brief description, list of associated symptoms, or treatment suggestions. The result appears prominently within the interface, ensuring clarity for the user. This screen may also include a suggested treatment section (as seen in antibotics.jpg) and a note emphasizing that users should consult a medical professional for confirmation and treatment.
To deliver the outcome of the symptom prediction process, providing users with an early diagnosis indication, which can guide them to seek timely medical advice.

VI. CONCLUSION AND FUTURE SCOPE

The project titled "Gynecological Disease Prediction using Machine Learning" was designed to predict gynecological diseases based on user-entered symptoms through natural language input. By employing Natural Language Processing (NLP) techniques such as tokenization, lemmatization, and TF-IDF vectorization, coupled with a Multinomial Naive Bayes classifier, the system achieved an accuracy of 87%, indicating the potential of machine learning models in assisting early disease detection and providing timely treatment suggestions. The user-friendly interface built using Streamlit ensures ease of use for individuals with limited technical expertise. However, the current system has a few limitations — it supports only English input, covers just two diseases, and does not incorporate personal data or historical symptom analysis, which could improve the accuracy and relevance of predictions. To overcome these limitations and enhance system usability, future improvements include expanding the disease database to cover more gynecological and general health conditions, integrating multilingual support including Hindi and Tamil, adding user login and symptom history tracking features, deploying the application on cloud platforms with persistent storage for scalability, and incorporating a chatbot-style conversational interface within Streamlit to offer a more interactive and intuitive experience for users.

REFERENCES

1. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
3. Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Draft available at https://web.stanford.edu/~jurafsky/slp3/.
4. Streamlit Documentation. Available at: https://streamlit.io
5. spaCy NLP Library. Available at: https://spacy.io
6. TextBlob: Simplified Text Processing. Documentation at: https://textblob.readthedocs.io
7. Scikit-learn: Machine Learning in Python. Official Site: https://scikit-learn.org
8. Python Software Foundation. (2023). Python Language Reference. Retrieved from https://www.python.org
9. WebMD. Symptoms and Treatments for UTIs and PCOD. Available at: https://www.webmd.com
10. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). UTI Overview. https://www.niddk.nih.gov
11. Mayo Clinic. PCOD Overview. https://www.mayoclinic.org
12. OpenAI. GPT Models for NLP Use Cases. https://openai.com