IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 March 2024, accepted 18 April 2024, date of publication 27 May 2024, date of current version 21 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406162

RESEARCH ARTICLE

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

SUMANA DE[1], PAROMITA GOSWAMI[2], NEETU FAUJDAR[2], AND GHANSHYAM SINGH[3]

[1]Department of Computer Science and Engineering, C. V. Raman Global University, Bhubaneswar, Odisha 752054, India
[2]Department of Computer Engineering and Application, GLA University, Mathura, Uttar Pradesh 281406, India
[3]Centre for Smart Information and Communication Systems, Department of Electrical and Electronic Engineering Science, University of Johannesburg, Auckland Park Campus, Johannesburg 2006, South Africa

Corresponding author: Neetu Faujdar (neetu.faujdar@gmail.com)

ABSTRACT In this paper, the Gynaecological Disease Diagnosis Expert System (GDDES) is a Graphical User Interface, developed with the Support Vector Classifier (Machine Learning Algorithm) and Natural Language Processing. It is language-independent, allowing women from any state in India to use the system in their own native tongue and have their disorders diagnosed in that language. The diagnosis process is divided into two steps: At first, the user selects their regional language and the system asks some queries in their selected language and submits the reply for each query, then the system uses the Support Vector Classifier (SVC) Model to predict the disease name; and secondly, the user is prompted to record their symptoms in their native tongue and GDDES uses Natural Language Processing to calculate cosine similarities and play the most similar voice recording of disease diagnosis, and displays the sentences of the recording in the user's native language. The system with the SVC Model provides 93% accuracy and precision and 92% recall and f1 score.

INDEX TERMS Disease diagnosis, support vector classifier, natural language processing, gynaecological disease diagnosis expert system (GDDES).

I. INTRODUCTION

This gynaecological issue will affect almost all women at some point in their lives. There are many different gynaecological issues that women deal with, but some of them are common. These issues include pelvic pain, irregular periods, difficulty getting pregnant, frequent urination, burning while urinating, ovarian cysts, and many others. In countries like India, where only a small percentage of underprivileged women have access to high-quality gynaecological care and some women refuse to visit doctors for gynaecological issues. Some women use online doctor consultants or online applications to learn about their disease but still, some of them can encounter difficulties due to a lack of finance or knowledge of English. In this paper, to help such women, we have designed the Graphical User Interface of the Gynaecological Disease Diagnosis Expert System (GDDES) based on the Support Vector Classifier (Machine Learning Algorithm) and Natural Language Processing. The unique feature of GDDES is that it is language-independent, allowing women from any state in India to use the system in their native tongue and have their disorders diagnosed in that language. The diagnosis process in the GDDES is divided into two steps.

In the first step, the system asks the user to select her local language, after selecting the language, the system will ask some queries in the selected language and the user submits the reply for each query. According to the user's reply, the system provides the exact disease name to the user. For this diagnosis, the system uses a Disease Dataset that contains different symptoms as input features, and according to the symptoms, the disease name is the output or target feature. The dataset is

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Ni.

built from different datasets, taken from the Kaggle machine learning database. The detail about the dataset is discussed in section III. In order to forecast the disease name based on the newly submitted query by the user, the system builds the Support Vector Classifier (SVC) Model using the dataset.

In the second step, the user is prompted to record their symptoms in their native tongue. The system will fetch this recording and also the stored recordings from the particular language Sub-Database and translate both recordings into English texts after that use Natural Language Processing both texts to calculate the cosine similarities and according to the smaller cosine similarity, the most similar recording is fetched from Sub-Database and after the user clicks on the play button, GDDES plays that recording. And also displays the sentences of the recordings in the user's native language.

The system with the SVC Model provides 93% accuracy and precision and 92% recall and f1 score.

The remaining portions of the work are structured as follows: The emphasis of Section II is on related earlier works; Section III demonstrates the suggested GDDES architectural model in depth; The implementation and methodology is covered in Section IV; The Proposed Algorithm is displayed in Section V, GDDES performance analysis is shown in Section VI, simulation results are shown in Section VII, and Section VIII shows the limitations and Section IX concludes with a conclusion and recommendations for future work.

II. RELATED WORKS

Based on ultrasound diagnostic pictures, the suggested HHO-DQN approach in [1] performs better than current active learning algorithms for the categorization of seven distinct kinds of ovarian cysts.

Zhou B et. al. examined the history, uses, and ethical issues of artificial intelligence and NLP-driven smart healthcare, and discussed limitations and future work directions in [2].

Seed words were inserted to construct representations for UTI symptoms using four-word embedding approaches and phrase detection methods. Prototype embedding can capture semantic information about UTI symptoms, resulting in more symptom words [3].

Researchers described a chatbot service for the Covenant University Doctor telemedicine system in their research paper [4]. This chatbot provides a personalized diagnosis based on self-input.

Moller et al. developed two predictive models to forecast HA-UTI risk before it occurs, allowing healthcare workers to take preventive measures in the paper [5].

The study referred to in [6] shows the majority of polycystic ovarian syndrome classification is based on clinical data sets, so a hybrid methodology will be more precise.

In this paper, Subrato et al. [7] examined the history, applications, and ethical issues of artificial intelligence, which will lead the digital revolution in obstetrics and gynaecology as well as medical practice in general.

Data-driven polycystic ovarian syndrome (PCOS) identification in women was the focus of Subrato et al.'s study [8]. The Kaggle repository contains a dataset that can be used for free, and machine learn- ing methods are applied to it. This dataset includes 541 women with 43 characteristics, including 177 with PCOS. The best features for predicting PCOS are first found using a univariate feature selection technique.

An automated deep learning system for auxiliary PCOS detection that looks into the potential of scleral changes in PCOS identification was proposed by Lv and Ying et. al. [9]. The technique was applied to a collection of 721 Chinese women's full-eye pictures, 388 of whom have PCOS. An improved U-Net was used to separate scleral images from full-eye images, and a Resnet model was used to derive deep features from the scleral images.

A system for identifying and forecasting PCOS treatment based on an ideal and minimal collection of characteristics was described by Vaidehi et al. [10]. To identify whether a woman has PCOS, five different machine learning classifiers were used. Using the CHI SQUARE technique, the best 30 features from the dataset were chosen and included in the feature vector. The Random Forest Classifier has the highest and most reliable accuracy for predicting PCOS treatment.

A scenario was put up by the researchers [11] to assess the effectiveness of integrating machine learning and natural language processing within a disease prediction system. The authors thoroughly analyzed their supervised data utilizing clustering analysis, similarity, and the frequency of symptoms. As a result, they are able to view the forecast quite well, however, there are still certain issues to be fixed.

To assess the resources needed for the creation, deployment, and operation of an NLP-augmented system, researchers in [12] sought to determine whether manual chart review could be substituted in a practical clinical context.

A single uropathogenic (E. coli) RUTI may be more accurately predicted by RF, according to the paper [13]'s development of tested machine learning models. Both host and bacterial characteristics greatly influenced the development of RUTI in the prediction models in the two clinical settings. According to the results, doctors might take action to prevent the emergence of RUTI.

The researchers in [14] employed DSaaS to predict antibiotic resistance in the Clinical Laboratory Surgical Unit using 1486 hospitalized patients with nosocomial UTIs. Neural networks, Support Vector Machines, and Catboost are a few machine learning approaches that were used to develop predictive models.

ML and DL techniques are also used in other fields, such as in [15] to identify Breast cancer, in [16] Type 2 Diabetes, to identify plant disease in [17], and in [18] bone cancer detection.

III. THE PROPOSED ARCHITECTURAL MODEL

In this paper, the GDDES has mainly four components. That is the user interface, administrator module, inference engine, and explanation system. Following Figure 1 shows the architectural components of the system.

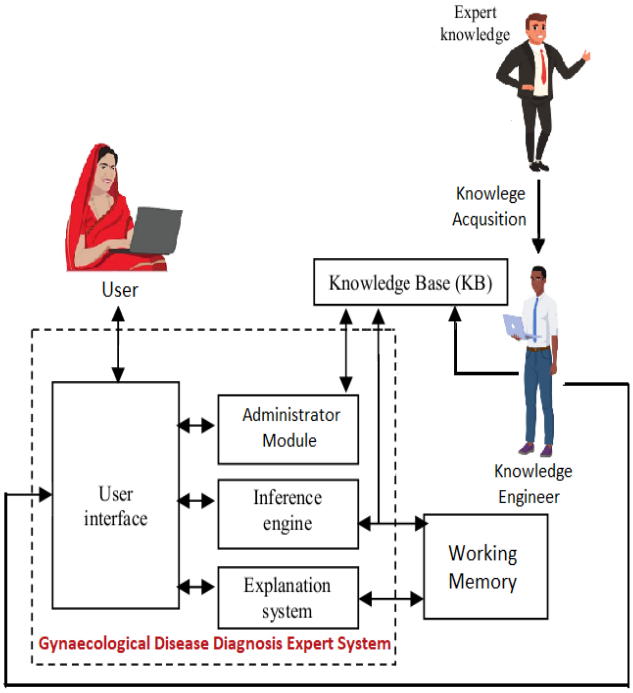


FIGURE 1. Architectural components of the GDDES.

The system's **user interface** is in charge of facilitating communication between users and administrators. The Graphical User Interface (GUI) for this system's administrator, knowledge expert, and knowledge engineer was created using the Python programming language.

A. KNOWLEDGE ENGINEER

Any knowledge management program that analyses sophisticated expert decision-making and turns it into guidelines and recommendations that less experienced individuals may utilize would benefit greatly from the expertise of the knowledge engineer. It is the responsibility of knowledge engineers to extract implicit information from subject-matter experts, convert it into explicit knowledge, and store documents with guidelines and data in the Knowledge Base. The supervisor function is performed by the knowledge engineer. Only he has the authority to alter the information base and gain access to it. He will also use the user interface to communicate with the computer.

The **expert** in the system is a gynaecologist who has the depth knowledge of gynaecological disease diagnosis.

Through the **administrator module**, an administrator can modify the knowledge base or access the knowledge base.

TABLE 1. Dataset contains 800 records with the following attributes.

Input Attributes	Values
Occurrence of nausea	Yes/No
Lumbar pain	Yes/No
Urine pushing (continuous need for urination)	Yes/No
Micturition pains	Yes/No
Burning of urethra	Yes/No
Itch	Yes/No
swelling of urethra outlet	Yes/No
Inflammation of urinary bladder	Yes/No
Nephritis of renal pelvis origin	Yes/No
irregular periods	Yes/No
no periods	Yes/No
excessive hair growth	Yes/No
buttocks weight gain	Yes/No
belly fat	Yes/No
hair loss	Yes/No
Acne	Yes/No
Output Attributes	**Disease**
Disease Name	UTI/PCOS

B. WORKING MEMORY

During an encounter, the user's input to the expert system is kept in working memory. Working memory values are used to assess antecedents in the knowledge base. The working memory may hold newly added, altered, or removed values as a result of knowledge base rules.

C. EXPLANATION SYSTEM

With the aid of this module, the expert system is better able to explain to the user how it arrived at a specific decision.

The **user** can be any person who wants to know about gynaecological disease diagnosis. The user will interact with the system through the GUI.

D. KNOWLEDGE BASE

The knowledge base stores the results of the gynaecological disease diagnosis according to some parameters. The whole knowledge is stored in the form of the dataset in the knowledge base. For the experimentation, to analyze the results of GDDES, the dataset is made by taking information from the websites and also by downloading the Urinary tract infection dataset from the Kaggle Machine Learning database. And also it stores the voice recordings of the gynaecological disease diagnosis in different regional languages.

Based on these various input parameters, and the output attribute is the "Result," which diagnoses the disease name. The dataset contains a total of 2 different classification results: Urine Infection and Polycystic ovary syndrome (PCOS).

IV. THE SYSTEM IMPLEMENTATION AND METHODOLOGIES

Gynaecological Disease Diagnosis Expert System (GDDES) is built with two steps and the working flow in it also has two steps.

Step-1

Figure 2 shows the implementation and diagnosis methodology of GDDES in phase-1

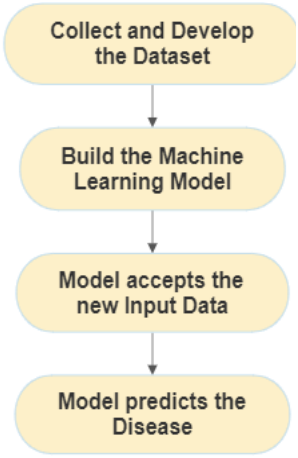


FIGURE 2. Implementation and diagnosis methodology of GDDES in phase-1.

(i) Collect and Develop the Dataset: For experimentation, The Dataset contains two different Gynaecological Diseases: Urinary tract infection, and Polycystic ovary syndrome. For our purpose, the dataset is made by taking information from the websites and also by downloading the Urinary tract infection dataset from the Kaggle Machine Learning database.

From Table 1, it is clear that the dataset contains 16 different Input Attributes and one Target Attributes. For example, Table 2 shows attributes with its values.

(ii) Build the Machine Learning Model: For the prediction of diseases based on the collected dataset in this system, different supervised machine learning algorithms are applied such as Decision Tree, Random Forest Classifier, Support Vector Classifier (SVC), Naïve Bayes, and K-Nearest Neighbor. Figure 7 shows the different accuracy, Precision, and Recall Results; from where it is found that the SVC Model performs well with 93% accuracy and precision and 92% recall and f1 score.

(iii) The model accepts the new input Data: After the development of the machine learning model, the user can submit new inputs with different symptoms to the model.

Model Predicts the Disease: The model then predicts the disease name according to the symptoms.

Step-2

Figure 3 shows the implementation and diagnosis methodology of GDDES in phase-2

TABLE 2. Example of data in dataset for disease diagnosis in Step-1.

Input Attributes	Values
Occurrence of nausea	No
Lumbar pain	No
Urine pushing (continuous need for urination)	Yes
Micturition pains	Yes
Burning of urethra	Yes
Itch	No
swelling of urethra outlet	Yes
Inflammation of urinary bladder	Yes
Nephritis of renal pelvis origin	No
irregular periods	No
no periods	No
excessive hair growth	No
buttocks weight gain	Yes
belly fat	No
hair loss	No
Acne	No
Output Attributes	**Disease**
Disease Name	UTI

(i) Collect Different gynaecological Disease Diagnosis Recordings in Different Languages: For the primary stage of the system development, for the experimentation, we collected disease reviews from different patients from different websites and recorded the reviews. As we all know for the same disease, different people can have different signs and symptoms. So, we collected different reviews that describe the disease in detail with different symptoms, different diagnosis results, and treatments.

For the examples, two following reviews are given:

review 1: "I have been diagnosed with UTI. My doctor gives me an antibiotic for 1 week. Is it normal if I have UTI that my urine smells something strange or fishy and burning sensation while urinating? The doctor prescribes me to drink water and take rest for 7 days".

Review 2: "I need to pee more often than usual, I have pain or discomfort when peeing and also sudden urges to pee, doctor gives me paracetamol and antibiotic for 3 days".

These two reviews are in English language but in Bengali language it will be like the following

Review 1 : আমার ইউটিআই ধরা পড়েছে। আমার ডাক্তার আমাকে 1 সপ্তাহের জন্য একটি অ্যান্টিবায়োটিক দেয়। আমার ইউটিআই থাকলে এটা কি স্বাভাবিক যে প্রস্রাব করার সময় আমার প্রস্রাবের অদ্ভুত বা মাছের গন্ধ এবং জ্বালাপোড়া হয়। ডাক্তার আমাকে জল পান করতে এবং ৭ দিন বিশ্রাম নিতে বলেছে।

Review 2: আমার স্বাভাবিকের চেয়ে বেশি ঘনঘন প্রস্রাব করা দরকার, প্রস্রাব করার সময় আমার ব্যথা বা অস্বস্তি হয় এবং হঠাৎ প্রস্রাব করার তাগিদ হয়, ডাক্তার আমাকে 3 দিনের জন্য প্যারাসিটামল এবং অ্যান্টিবায়োটিক দেন

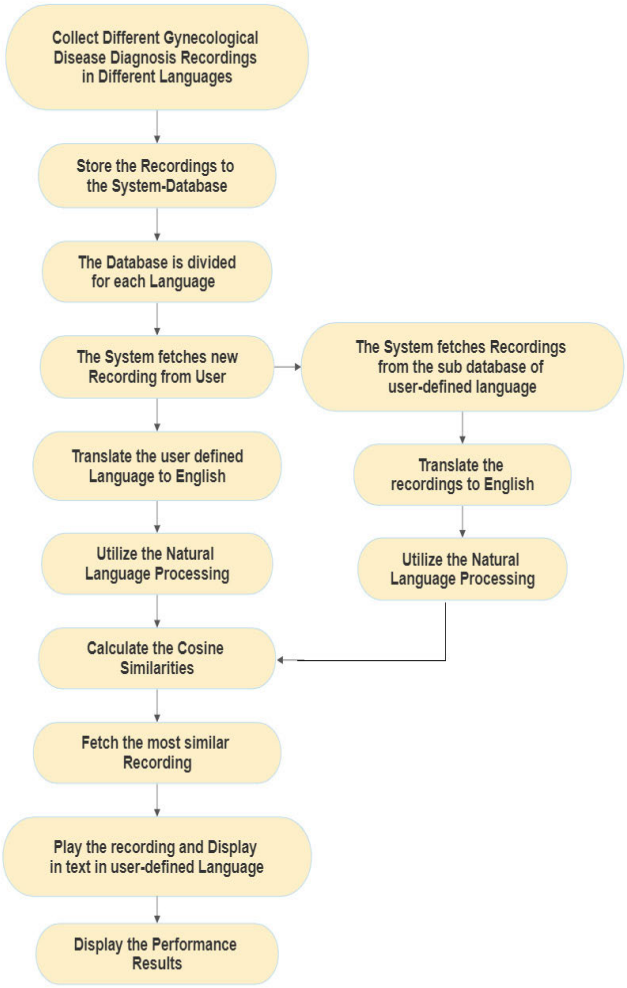


FIGURE 3. Implementation and diagnosis methodology of GDDES in phase-2.

These same sentences two are in Hindi language:

Review 1 : मुझे यूटीआई का पता चला है। मेरा डॉक्टर मुझे 1 सप्ताह के लिए एंटीबायोटिक देता है। क्या यह सामान्य है अगर मुझे यूटीआई है कि मेरे पेशाब से कुछ अजीब या गड़बड़ गंध आती है और पेशाब करते समय जलन होती है। डॉक्टर ने मुझे पानी पीने और 7 दिन आराम करने की सलाह दी है।

Review 2: मुझे सामान्य से अधिक बार पेशाब करने की आवश्यकता होती है, मुझे पेशाब करते समय दर्द या परेशानी होती है और अचानक पेशाब करने की इच्छा होती है, डॉक्टर मुझे 3 दिनों के लिए पेरासिटामोल और एंटीबायोटिक देते हैं

In the future at the time of system upgradation, for real-life use, the reviews can be recorded directly from the patients.

(ii) Store the recordings in the System's Database: Each of the review recordings for each disease will be stored in the Database.

(iii) The Database is divided for each language: Each language of recording needs separate storage space. So, The Database is divided into sub-databases for each language. Different language reviews have different storage spaces in the database.

(iv) The system fetches new recordings from the user: After the system is built successfully whenever the user accesses it at the second stage of diagnosis, the system asks to record her symptoms by clicking the record button in the Graphical user interface (GUI). The user records her problems or the disease symptoms in her native language and submits it to the system.

(v) The system fetches recordings from the sub-database of user-defined language: As at the first step of the diagnosis, the system identifies the native language of the user so whenever the user records her symptoms and submits them to the system then the system simultaneously fetches the recordings that belong to the identified language from the sub-database.

(vi) Translate the user-defined language to English: After getting the recording the system translates the voice into English text.

(vii) Translate recordings to English: the system simultaneously does the same process for the stored recordings it will translate each of the recordings into English texts

(viii) Utilize natural language Processing: after that, the system applies natural language processing on the translated recordings text and also on the stored recordings text.

Here, Natural Language Processing includes the steps:

- **Tokenizing by word:** Tokenization in Python essentially means breaking up a long piece of text into smaller lines, or words, or even making up words for languages other than English. Tokenizing content word by word enables the discovery of especially frequent words.

  For example: Tokenizing the sentence ''She has Disease'' would result in three different tokens: ''She'', ''has'', and ''Disease''.
- **Filtering Stop Words:** Search engines have already been designed to ignore stop words like ''the,'' ''a,'' ''an,'' and ''in'' when indexing items for searching and when retrieving them as the outcome of a search query.
- **Stemming:** Reducing a word to its stem is called stemming, and it attaches to suffixes, prefixes, or the roots of words that are referred to as ''lemmas.'' Stemming (NLP) is beneficial for both natural language understanding and natural language processing.

  For Example, the stem of the words ''suffering,'' ''suffered,'' is ''suffer.''
- **Tagging Parts of Speech:** Assigning each word in a sentence to its appropriate part of speech is known as POS tagging. We already know that the constituents of speech are nouns, verbs, adverbs, adjectives, pronouns, conjunctions, and their subcategories. Following POS tagging, we deleted the proper noun singular ('NNP'), proper noun plural ('NNPS'),

personal pronoun ('PRP'), verb, present tense not third person singular ('VBP'), coordinating conjunction ('CC'), adverb ('RB'), preposition/subordinating conjunction ('IN'), and verb, present tense with third person singular ('VBZ'), and after that we joined the words to form a new sentence. For Example,
sentence = "I have fever and also burning urine and nausea and low back pain"
After performing the above operations,
New Sentence = "I fever also burning urine nausea low back pain"

- **Utilize the cosine similarity:** The cosine of the angle that separates any two non-zero variables in an inner space of products is used to compare them.
Similarity = (A.B) / (||A||. ||B||) where A and B are vectors. To find the Cosine similarity, we have to first count how many times each word comes in each manuscript. The 'CountVectorizer' and 'TfidfVectorizer' procedures from the Scikit-Learn library can be used to count the instances of a word in each document. Then, using a formula called Term Frequency Inverse Document Frequency (TFIDF), we may assess how each one compares and contrasts with the others.
(ix) Fetch the most similar Recording: The stored recording that has the smaller cosine similarity value is selected from the Database.
(x) Play the recording and display in the text in user-defined language: In this step, the user can listen to the recording of one review, stored in the database, that is most similar to her symptoms, and also, she can see the sentences of this recording on the graphical user interface of the system.
(xi) Display the performance results: after the diagnosis process, the system displays the performance results as accuracy Precision recall, and f1 score.

The Data Flow Diagram of the Implementation and diagnosis methodology of GDDES

To develop the gynaecological disease diagnosis expert system, we first imported the libraries and configured the Graphical User Interface (GUI) of the system. We used Python programming language. The GUI contains labels, textboxes, and buttons.

It also contains a drop-down menu with the parameters of different regional languages in India. At the time the user selects his favourite language from the drop-down menu in the GUI then the system fetches the user-selected language and also processes the further calculation with this language. So we have added different regional languages as the parameter in the drop-down menu and also added the labels according to the selected parameter. If the selected language is Bengali, then the system shows all of the options and diagnosis results in Bengali. The system will display different symptoms with the help of labels and ask the user to answer each of the questions using the drop-down menu with the parameter yes or no in Bengali. Now the user selects the options provides the answers to each of the questions and submits it to the system. Now the system uses machine learning techniques by fetching the user-selected options from the GUI and converting the texts into numerical values.

After that, the system reads the data set, trains and tests the dataset using a machine learning model that is a support vector classifier applies the new user numerical input data to predict the disease name, and lastly displays the disease name in the GUI in English and Bengali. In the same way, the system works with other regional languages. Figure 4 shows the details for phase 1.

Now in the second stage, shows in Figure 5, we have used the depth analysis of the disease. We have added a voice recorder in the GUI. Whenever the user clicks on the record button the GUI will open one new window for this record button and the user will be allowed to record her symptoms for this particular disease in her own regional language. After that, the system gets the user's recording and also fetches the particular disease recordings from the Bengali sub-database. The database here contains different disease diagnosis recordings in different languages and according to the languages, the database is divided into sub-databases.

V. THE PROPOSED ALGORITHM

Step 1: Import Libraries –
Example: tkinter, import pandas, sklearn.model, speech_recognition, nltk, math,
import speech_recognition.
Step 2: Configure the Graphical User Interface –
Example:

```
window=Tk()
mywin=MyWindow(window)
window.title('Gynaecological Disease Diagnosis Expert System (GDDES)')
menu.set("Select Any Language")
drop= OptionMenu(win, menu,"Hindi", "Bengali","Odiya","punjabi")
```

Step 3: Read the Dataset –
Example: disease = pd.read_csv('UTI.csv')
Step 4: Train and Test the Dataset –
Example:

```
rf = RandomForestClassifier(n_estimators=100, random_state=0)
rf.fit(X_train, y_train)
T=(rf.score(X_test, y_test))
T=T*100
```

Step 5: Get the new input –
Example:

```
a = int(self.t1.get())
b = int(self.t2.get())
c = float(self.t3.get())
d = int(self.t4.get())
e = int(self.t5.get())
f= int(self.t6.get())
```

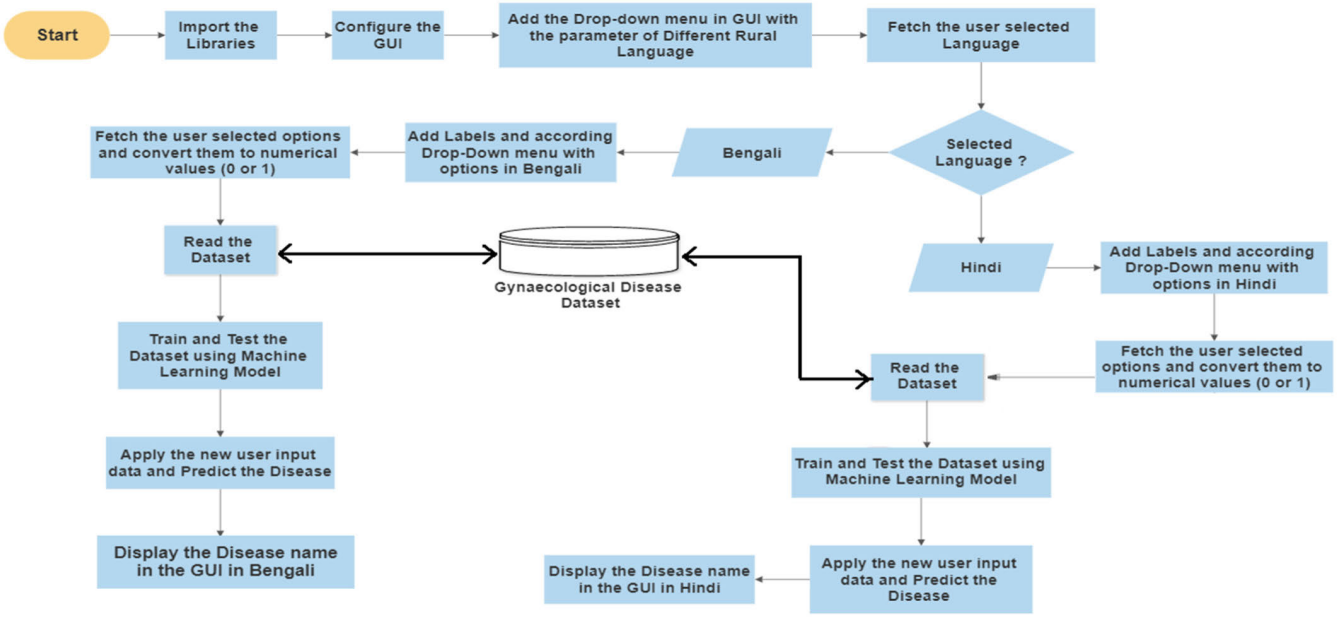


FIGURE 4. Phase 1 data flow diagram of GDDES.

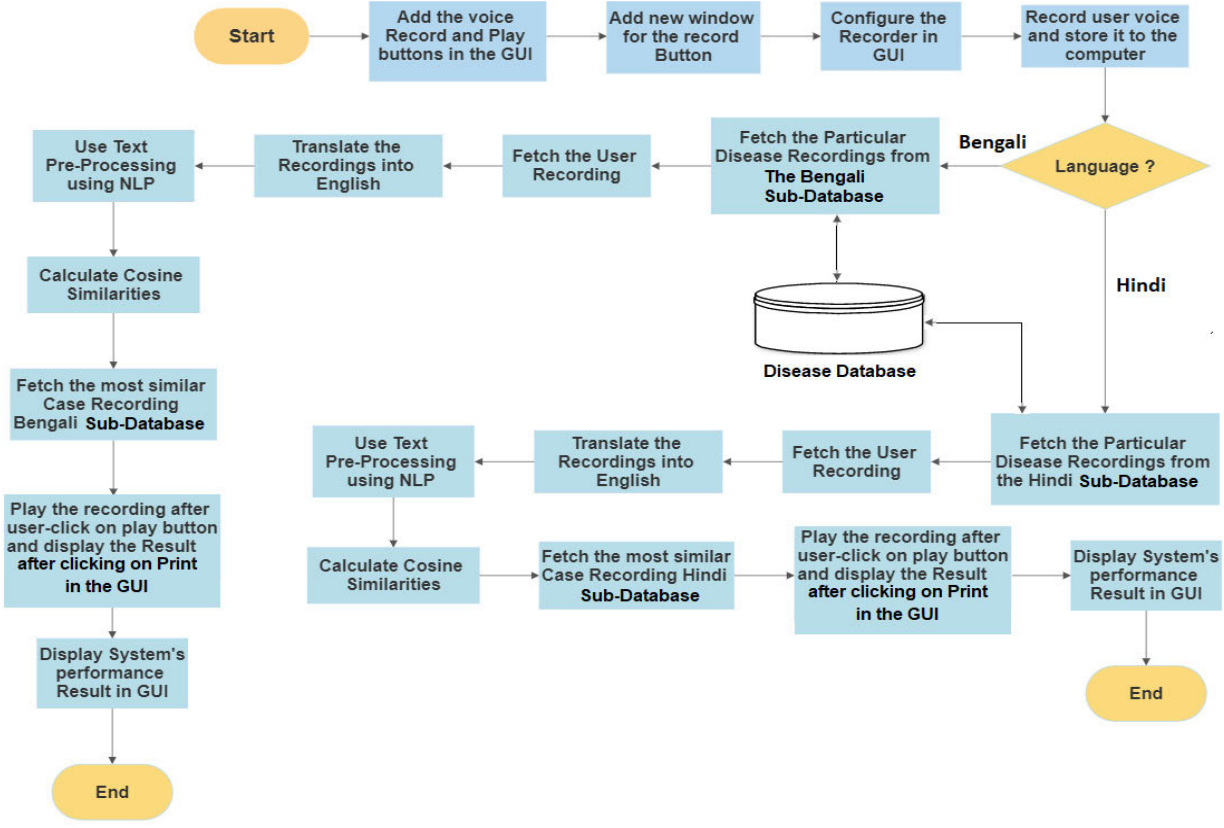


FIGURE 5. Phase 2 data flow diagram of GDDES.

```
g = int(self.t7.get())
h = int(self.t8.get())
i = int(self.t9.get())
j= int(self.t10.get())
k = int(self.t11.get())
l = float(self.t12.get())
m = int(self.t13.get())
n= float(self.t14.get())
```

Step 6: Store it to the new input in 'Xnew'

```
Xnew = [a,b,c,d,e,f,g,h,i,j,k,l,m,n]
```

Step 7: Predict the classification-

```
ynew = rf.predict(Xnew)
```

Step 8: Fetch the recording –

Example:

```
# getting the recording duration from the entry
duration = int(duration_entry.get())
# calling the recorder via the rec() function
recording = sounddevice.rec(duration*freq,
samplerate=freq, channels=2)
```

Step 9: Store the recording –

Example:

```
# looping through all the files in the current folder
for file in os.listdir():
# checking if the file name is recording.wav
if file == 'recording.wav':
# spliting the base and the extension
base, ext = os.path.splitext(file)
# getting current time
current_time = datetime.now()
# creating a new name for the recorded file
new_name = 'recording-' +
str(current_time.hour) + '.' +
str(current_time.minute) + '.' +
str(current_time.second) + ext
# renaming the file
os.rename(file, new_name)
# display a message when recording is done
showinfo('Recording complete', 'Your recording
is complete')
```

Step 10: Convert the new recording and stored recordings in English

Example:

```
pip install translate
from translate import Translator
translator= Translator(to_lang="English")
translation = translator.translate(text)
```

Step 11: Use NLTK for text processing

Example:

```
sentence= translation
tagged_sentence =
nltk.tag.pos_tag(sentence.split())
edited_sentence = [word for word,tag in
tagged_sentence if tag != 'NNP' and tag !=
'NNPS' and tag != 'PRP' and tag != 'VBP' and
tag != 'CC' and tag != 'IN' and tag != 'VBZ']
sentence= (' '.join(edited_sentence))
```

Step 12: Use Cosine Similarity to calculate the similarities

```
X_list = word_tokenize(sentence)
Y_list = word_tokenize(Stored_1)

# remove stop words from the string
X_set = {w for w in X_list if not w in sw}
Y_set = {w for w in Y_list if not w in sw}

# cosine formula
for i in range(len(rvector)):
```

TABLE 3. Comparison of previous works.

Previous Paper	Conclusion
Referred to in [1]	The suggested model outperformed the ANN, CNN, and AlexNet models in terms of precision (96%), recall (96.5%), accuracy (97%), and IoU (0.65).
Referred to in [3]	With a mean accuracy of between 0.51 and 0.86, 142 additional UTI symptom terms were discovered.
Referred to in [5]	At the time of admittance, 72% (185/256) of patients who developed a HA-UTI had a risk score of less than 0.15.
Referred to in [8]	When 40-fold cross validation is applied to the ten most important criteria, RFLR has the highest testing accuracy (91.01%) and recall value (90%) according to the results. As a result, PCOS individuals can be accurately classified using RFLR.
Referred to in [9]	Findings demonstrate the high ability of deep learning in the diagnosis of PCOS, with our approach achieving an overall AUC of 0.979 and an accuracy of 0.929.
Referred to in [10]	To identify whether a woman has PCOS, five different machine learning models were used, with the Random Forest Classifier being the best and most accurate.
Referred to in [12]	In acute care units, the system with NLP enhancements discovered more permanent urinary catheter days, and in intensive care units. The total positive predictive value was 54.2%, while the sensitivity was 65% (90.9%).
Referred to in [13]	The majority of NCKUH patients experienced RUTI caused by E. coli caused by re - infection, to 74% experiencing 2 cases of UTI inside of six years and 63% experiencing 3 cases inside of twelve months.
Referred to in [15]	With the highest value in each measure, Catboost showed the best predictive results.

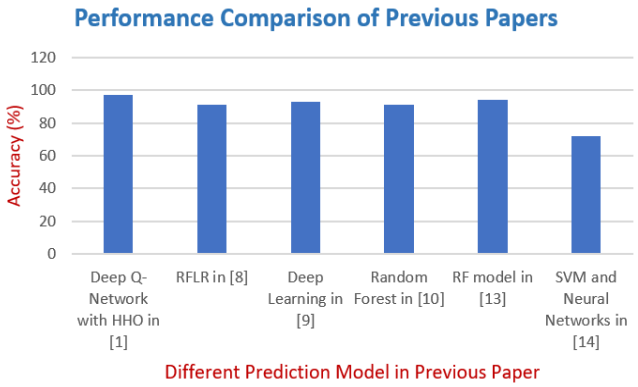


FIGURE 6. Performance comparison of previous works.

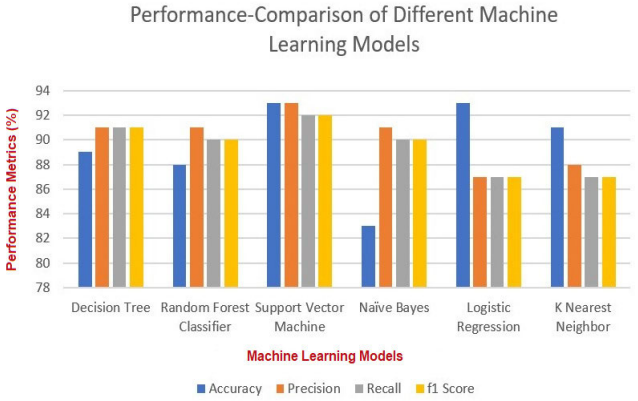


FIGURE 7. Performance comparison of different machine learning model in the proposed work.

FIGURE 8. Voice recorder.

```
c+ = l1[i]*l2[i]
cosine = c / float((sum(l1)*sum(l2))**0.5)
print("similarity:", cosine)
```

Step 13: Get the recording with smaller cosine similarity from Database
Step 14: Play the recording

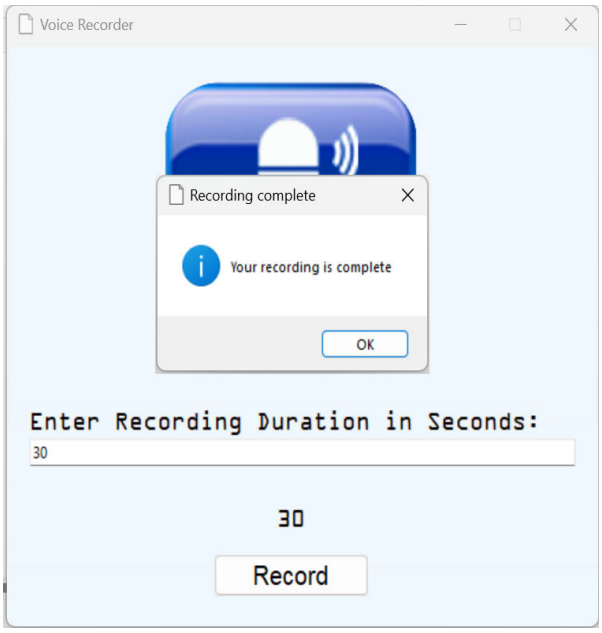


FIGURE 9. Voice recorder.

Step 15: Display the stored recording text to the user-defined language
Step 16: Display the performance Result

VI. THE PERFORMANCE ANALYSIS

As we can see, numerous attempts have been made in the past to classify gynaecological disorders based on different criteria. Each piece of work offers a varying level of prediction accuracy. Table 3 shows one comparison of previous works and Figure 6 compares various machine learning methods used in various projects with varying degrees of accuracy.

Some well-known machine learning methods from earlier studies are used in this article and are applied to the dataset we downloaded. Figure 7 demonstrates that the Support Vector Machine (SVM) outperforms other machine learning methods in this case.

VII. THE SIMULATION RESULT

To help women to know about their gynaecological diseases GDDES is developed. Figure 8 and Figure 9 show the voice recorder that will open after clicking the record button on the GUI. In this paper, one expert system is contracted with the Graphical User Interface using Python programming language.

Figure 10 shows the diagnosis is done in Bengali language, and Figure 11 shows the diagnosis done in the Hindi language.

VIII. THE LIMITATIONS

The system is built based on very few parameters. The Disease should be diagnosed by analyzing vast data and images. There is a total of 22 regional languages in India.

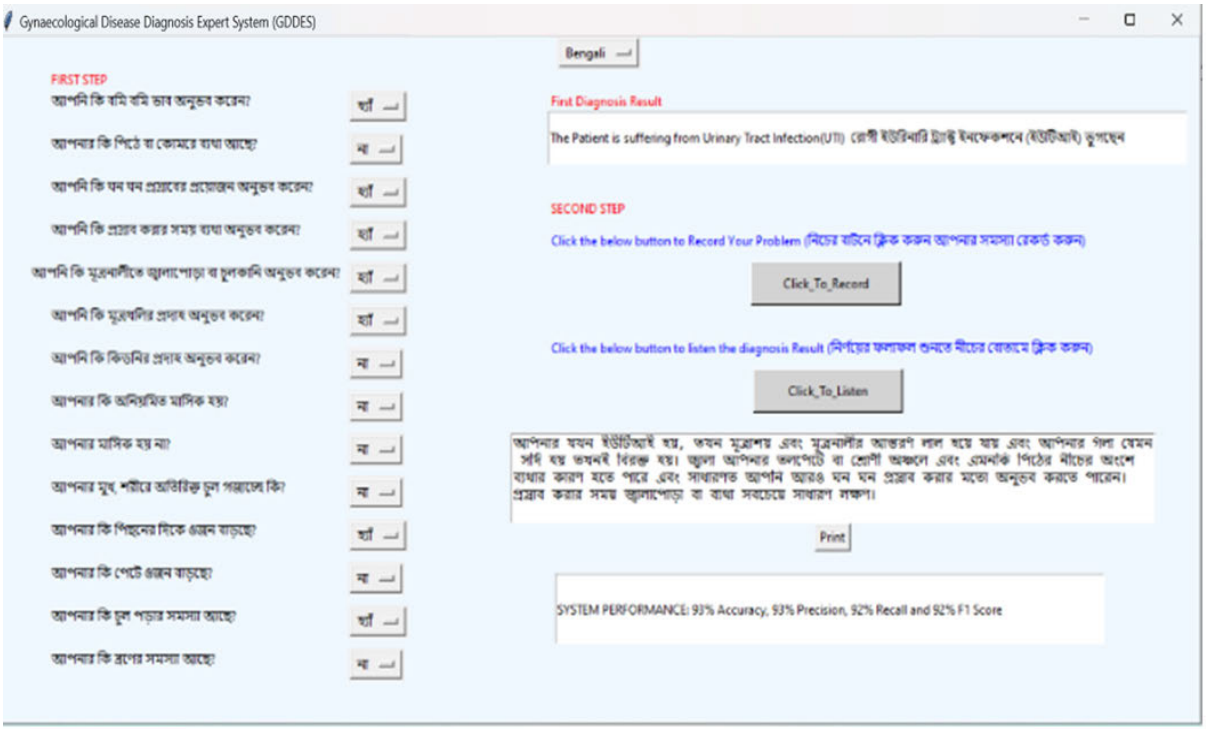


FIGURE 10. The disease diagnosis is in Bengali language.

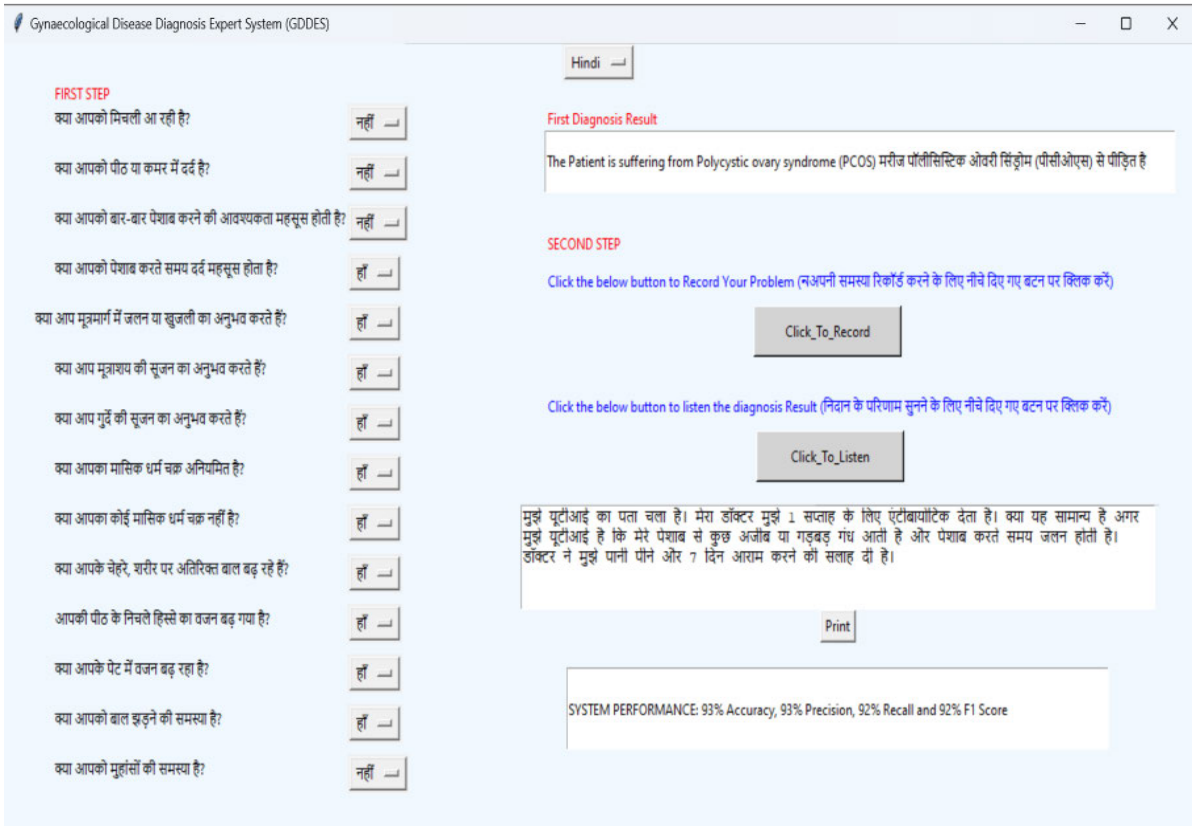


FIGURE 11. The disease diagnosis is in Hindi language.

We have added only four of them in the GUI. And we have diagnosed only two particular diseases that are UTI and PCOS.

IX. THE CONCLUSION AND FUTURE WORK

Women face a variety of gynaecological issues, including pelvic pain, irregular periods, difficulty getting pregnant, frequent urination, burning while urinating, ovarian cysts, and infertility. Women in underprivileged countries especially villagers often lack access to high-quality gynaecological care due to a lack of finance or English proficiency. In this paper, to help women, Gynaecological Disease Diagnosis Expert System (GDDES) based on a Support Vector Machine (ML Algorithm) and NLP is developed. GDDES is language-independent, allowing women from any state to use it to diagnose their disorders in their language. The GDDES has mainly four main components, user interface, administrator module, inference engine, and explanation system. The knowledge base of the system stores the results of the gynaecological disease diagnosis in a dataset and voice recordings in different regional languages.

The diagnosis process in the GDDES is divided into two phases, (i) with the user selecting their local language and submitting a reply to the system, which provides the exact disease name using the Support Vector Classifier Machine learning model and (ii) GDDES uses Natural Language Processing to calculate cosine similarities and play the most similar voice recording of disease diagnosis, displaying the sentences of the recording in the user's native language. The system with the SVC Model provides 93% accuracy and precision and 92% recall and f1 score.

In future work, we will overcome each of the limitations of the system by enhancing the knowledge base, increasing the parameters in the dataset, adding more regional languages, diagnosing more diseases and upgrading the GUI.

ACKNOWLEDGMENT

The academic assistance for this research study is gratefully acknowledged by Sumana De from the CSE Department, C. V. Raman Global University, Bhubaneswar; and Paromita Goswami and Neetu Faujdar from GLA University, Mathura.

REFERENCES

[1] C. Narmatha, P. Manimegalai, J. Krishnadass, P. Valsalan, S. Manimurugan, and M. Mustafa, "Ovarian cysts classification using novel deep reinforcement learning with Harris Hawks optimization method," *J. Supercomputing*, vol. 79, no. 2, pp. 1–11, 2021.

[2] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Natural language processing for smart healthcare," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 1–17, 2022, doi: 10.1109/RBME.2022.3210270.

[3] M. U. Alam, A. Henriksson, T. Hideyuki, T. Emil, N. Pontus, and H. Dalianis, "Terminology expansion with prototype embeddings: Extracting symptoms of urinary tract infection from clinical text," in *Proc. 14th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2021, pp. 47–57.

[4] N. A. I. Omoregbe, I. O. Ndaman, S. Misra, O. O. Abayomi-Alli, and R. Damaševičius, "Text messaging-based medical diagnosis using natural language processing and fuzzy logic," *J. Healthcare Eng.*, vol. 2020, pp. 1–14, Sep. 2020.

[5] J. K. Møller, M. Sørensen, and C. Hardahl, "Prediction of risk of acquiring urinary tract infection during hospital stay based on machine-learning: A retrospective cohort study," *PLoS ONE*, vol. 16, no. 3, Mar. 2021, Art. no. e0248636.

[6] S. J. Pulluparambil and S. Bhat, "Medical image processing: Detection and prediction of PCOS—A systematic literature review," *Int. J. Health Sci. Pharmacy*, vol. 5, pp. 80–98, Dec. 2021.

[7] S. A. Shazly, E. C. Trabuco, C. G. Ngufor, and A. O. Famuyide, "Introduction to machine learning in obstetrics and gynecology," *Obstetrics Gynecology*, vol. 139, no. 4, pp. 669–679, Apr. 2022.

[8] S. Bharati, P. Podder, and M. R. Hossain Mondal, "Diagnosis of polycystic ovary syndrome using machine learning algorithms," in *Proc. IEEE Region 10 Symp.*, Jun. 2020, pp. 1486–1489.

[9] W. Lv, Y. Song, R. Fu, X. Lin, Y. Su, X. Jin, H. Yang, X. Shan, W. Du, Q. Huang, H. Zhong, K. Jiang, Z. Zhang, L. Wang, and G. Huang, "Deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images," *Frontiers Endocrinol.*, vol. 12, pp. 1–8, Jan. 2022.

[10] V. Thakre, "PCOcare: PCOS detection and prediction using machine learning algorithms," *Bioscience Biotechnol. Res. Commun.*, vol. 13, no. 14, pp. 240–244, Dec. 2020.

[11] S. Akila, J. Prakash, and S. Uma, "Disease identification using machine learning and NLP," *J. Sci. Technol. Res.*, vol. 3, pp. 78–92, 2022.

[12] W. Branch-Elliman, J. Strymish, V. Kudesia, A. K. Rosen, and K. Gupta, "Natural language processing for real-time catheter-associated urinary tract infection surveillance: Results of a pilot implementation trial," *Infection Control Hospital Epidemiology*, vol. 36, no. 9, pp. 1004–1010, Sep. 2015.

[13] S.-L. Jeng, Z.-J. Huang, D.-C. Yang, C.-H. Teng, and M.-C. Wang, "Machine learning to predict the development of recurrent urinary tract infection related to single uropathogen, Escherichia Coli," *Sci. Rep.*, vol. 12, no. 1, p. 17216, Oct. 2022.

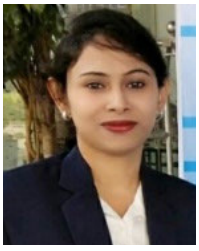
[14] A. Mancini, L. Vito, E. Marcelli, M. Piangerelli, R. De Leone, S. Pucciarelli, and E. Merelli, "Machine learning models predicting multidrug resistant urinary tract infections using 'DsaaS,'" *BMC Bioinf.*, vol. 21, no. S10, pp. 1–12, Aug. 2020.

[15] S. Aditya, "Comparison of two deep learning methods for classification of dataset of breast ultrasound images," in *Proc. IOP Conf. Series, Mater. Sci. Eng.*, 2022, pp. 1–7.

[16] A. Saxena, N. Mathur, P. Pathak, P. Tiwari, and S. K. Mathur, "Machine learning model based on insulin resistance metagenes underpins genetic basis of type 2 diabetes," *Biomolecules*, vol. 13, no. 3, p. 432, Feb. 2023.

[17] A. Verma, S. Shekhar, and H. Garg, "Plant disease classification using deep learning framework," in *Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solutions*, 2022, pp. 512–518.

[18] A. Sharma, D. P. Yadav, H. Garg, M. Kumar, B. Sharma, and D. Koundal, "Bone cancer detection using feature extraction based machine learning model," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–13, Dec. 2021.



SUMANA DE received the B.Tech. and M.Tech. degrees in computer science and engineering from West Bengal University of Technology (WBUT), Kolkata, West Bengal, India, in 2009 and 2012, respectively, and the Ph.D. degree in computer science and engineering from NIT Durgapur, West Bengal, in 2022.

She is currently an Assistant Professor with the CSE Department, C. V. Raman Global University, Bhubaneswar, Odisha, India. She has seven years of teaching experience in the academic field and has eight international publications. Her research interests include developing expert systems/knowledge management systems for different problem detection and diagnosis using artificial intelligence techniques/machine learning techniques/natural language processing/reasoning methodologies.

PAROMITA GOSWAMI received the M.C.A. degree from the University of North Bengal, West Bengal, in 2014, and the M.Tech. degree in multimedia and software systems from NITTTR, Kolkata, in 2016. Currently, she is an Assistant Professor with the Department of Computer Engineering and Application, GLA University, Mathura, India. She is also a Research Scholar with Mizoram University, Aizwal, India. Her research interests include cloud computing and big data.

NEETU FAUJDAR received the B.E. degree in information technology from VTU, Karnataka, in 2011, the M.Tech. degree in computer science engineering from Invertis University, Bareilly, in 2013, and the Ph.D. degree in computer science engineering from Jaypee University, Solan, India, in 2017. Currently, she is an Assistant Professor with the Department of Computer Engineering and Application, GLA University, Mathura, India. She has published more than 60 research articles. Her research interests include HPCA, GPU computing, the IoT, networking, and cloud computing. She has memberships in professional societies, such as the International Association of Engineers (IAENG) and the Institute of Research Engineers and Doctors (IRED).

GHANSHYAM SINGH received the Ph.D. degree in electronics engineering from Indian Institute of Technology, Banaras Hindu University, Varanasi, India, in 2000. He was a Visiting Researcher with Seoul National University, Seoul, South Korea. He was a Professor with the Department of Electronics and Communication Engineering, Jaypee University of Information Technology, Wakanaghat, Solan, India. He is currently a Professor of electronics and communication engineering and the Director of the Centre for Smart Information and Communication Systems, Department of Electrical and Electronic Engineering Science, University of Johannesburg, South Africa. He is the author/coauthor of more than 370 scientific research papers in peer-reviewed journals and international conferences. He has more than 23 years of teaching and research experience from the Academia and Research and Development Institutions. He has supervised various Ph.D. theses and master's dissertations. He has executed several sponsored research projects of ISRO and DRDO and currently working on 5G enabling technology localization resource allocation sponsored by SENTECH. He is the author of several books and book chapters published by Springer, Elsevier, IET, Wiley, and CRC. His research and teaching interests include a broad range of spectrum (RF/microwave, millimeter/THz wave, free-space optics, and visible light) technologies with it's emerging applications, such as in next-generation communication systems (5G/6G), sustainable smart cities, industry 4.0/5.0, healthcare 4.0, intelligent transport systems, energy management (IoE), and digital farming.