# A MACHINE LEARNING APPROACH TO CLINICAL DIAGNOSIS OF PEPTIC ULCER

**BY**

**IVARA, DAVID FREDRICK**
**GOU/U22/CSC/772**

**DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS, FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY,**
**GODFREY OKOYE UNIVERSITY, ENUGU STATE.** **IN PARTIAL FULFILLMENT OF THE AWARD OF BACHELOR OF SCIENCE (B.SC), DEGREE IN COMPUTER SCIENCE.**

**SUPERVISOR: DR. FRANK OKEBANAMA**

**JULY, 2026**

## APPROVAL PAGE

This research, titled **"A MACHINE LEARNING APPROACH TO CLINICAL DIAGNOSIS AND MEDICATION,"** has been assessed and approved by the Department of Computer Science and Mathematics Committees in the Faculty of Computing and Information Technology, Godfrey Okoye University, Enugu, Nigeria.

Supervisor

Signature

17/07/2026

Date

**External Examiner**

External Examiner

Signature

16/02/2026

Date

**Dr. Frank Okebanama**

HOD, Department of Computer Science and Mathematics

Signature

22/7/26

Date

**Prof. I.A Ajah**

Dean, Faculty of Computing And Information Technology.

Signature

22/7/26

Date

## CERTIFICATION

I certify that I completed the research included therein and that it was primarily based on my observations and investigation. Consequently, I will fully accept the University's decision if I am found to have cheated on the assignment.

**IVARA, DAVID FREDRICK**

**GOU/U22/CSC/772**

## DEDICATION

This project is dedicated to the trinity the source of my being.

## ACKNOWLEDGMENTS

I acknowledge the guidance dedicated by my supervisor Dr. Frank Okebanama towards this project, the HOD computer science and all scholars/lecturers which through their output of knowledge and inspiration cumulated and gives precedence to the output of this project.

Finally, I acknowledge and give gratitude to my mother: Eld. Esther F. Ivara, for her overall guidance and support in/beyond this project.

**ABSTRACT**

The rate of Peptic ulcer disease (PUD) has been a common gastrointestinal ailment most times usually caused by H. pylori infection or regular usage of NSAIDs. Due to high costs and inaccessibility of existing diagnostic procedures, such as endoscopy and laboratory tests, PepticSense—an aided machine learning-based clinical decision support web application (CDSS) for preliminary detection of presence or absence of gastric or duodenal ulcers—is developed in this paper. The data was synthesized from 6,000 patients' records in Nigeria with 42 features representing demographic, social and lifestyle information, patient's health condition, medical history and lab results. In order to develop the solution in a structured way using the Agile Software Development method is discussed alongside the Object-Oriented Analysis and Design approach and Decision Tree, Support Vector Machine, Naive Bayes, and Random Forest classifiers, CRISP-DM process, data was cleaned, prepared, feature-engineered, encoded, selected, scaled, and split into train and test sets. Comparisons of Accuracy, Precision, Recall, F1 Score, and AUC-ROC have been evaluated based on their performance, models showed that the latter is the most accurate classifier. The classifier was hyperparameter tuned through GridSearchCV and incorporated into Django REST Framework application with Angular frontend, PostgreSQL database, JWT authentication and role-based access control system. After evaluation of its performance on 1,200 records from the same distribution, but not seen during development of the model, high metrics of accuracy, precision, recall, F1-score, and AUC-ROC and low misclassifications between similar ulcer types were achieved. Thus, PepticSense became an inexpensive solution that could be used in any healthcare environment.

## Table of Contents

**Chapter Three: System Analysis and Design**



**Chapter Four: System Implementation**

**List of Tables**

## List of Figures

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background of the Study

Computers are automatic machines which takes input, perform a function to returns an output. With computers task and actions are performed faster and efficiently because of the breakthrough in advancement of computing. Computers originated and evolved from the analytical engine which was designed by Charles Babbage together with Ada Lovelace. As time evolves, in 1950, an expansion to the capabilities/how computer works was led by Alan Turing which gave rise to the question: “Can machines think?” in his seminar paper "Computing Machinery and Intelligence". Intelligence relates to the action of interacting (learning), reasoning, planning and adapting. In 1956, through John McCarthy artificial intelligence was adopted. Artificial intelligence is concerned with the autonomous mimicry of human intelligence through a computational syllogism, this intelligence has a variety of different types and in one of the types it’s projected to be beyond human intelligence, and this is known as artificial super intelligence (ASI) which is (not yet implemented).

Machine learning is a subset of artificial intelligence which uses an approach of computer machines been trained from data to act and react intelligently either narrowly or generally or in a super manner. A different approach is the traditional programming where instructions are explicitly defined, the output from an input is deterministic and solely dependent. Computationally trained machines, poses some of these attributes:

1. Heuristic and factual nature of problem solving
2. Knowledge base/inference engine where interaction communicates with rational thinking
3. These machines can be simple reflexive, model based, utility based, and goal based. These are used ranging from machines been able to precept input using sensors and acting using actuators. Most often, model based are widely used/deployed and this approach is also used in healthcare and will further be seen in this study.
4. The environment which the machine communicate with can be fully or partially observable.

Machine learning is described as a sub-area in the umbrella of artificial intelligence and a branch in the field of computer science. With regard to developing machine learning models for making predictions, the focus is often on algorithms, programming, and results. Machine learning entails much more than creating predictive models as illustrated in the definition given in Tom Mitchell's book from 1997 (Brownlee, Statistical Methods for Machine Learning, 2018).

Numerous methods been developed exists in the field of artificial intelligence and machine learning, tentative sometimes by statisticians, that prove very useful for the task which involves predictive modeling. A classical good example is the classification and regression trees that produce no resemblance to classical methods in statistics (Brownlee, Statistical Methods for Machine Learning, 2018).

Machine learning spans a huge application in many fields from science to non-science, an example is machine learning applied in healthcare systems which finds a space in predictive problems.

As predictive machines (e.g., ML models) are becoming more effective, more and more problems are being reframed as predictive problems. Whatever question you might have, you can always frame it as: "What would the answer to this question be?" regardless of whether this question is about something in the future, the present, or even the past (Huyen, 2022).

Popularly, machine learning is classified into three techniques: supervised learning, unsupervised learning and reinforcement learning.

Supervised learning, here, the instructor is aware of the correct answer and provides the algorithm with a correction when it makes a mistake (Krishna Kant Singh M. E., 2021). Unsupervised learning is described as the process through which the algorithm learns through the dataset that does not have any labeled data. The system has the responsibility of detecting the underlying pattern in the dataset (Krishna Kant Singh M. E., 2021). Reinforcement learning, in this type of learning process, the system is fed back its past action along with a "good" or "bad" score. If every good score adds +1 to the total and a bad score subtracts -1 from it, the goal is to keep the score as high as possible (Krishna Kant Singh M. E., 2021).

With these types of approach various algorithms are used for the implementation of this intelligent software/system.

With machine learning, agents can be trained and deployed. An agent can be in a form of an intelligent systems having the capability to autonomously interact with its environment and carrying out actions and reactions based on an advanced condition action rule. An example is the use of agents in the clinical decision support system (CDSS).

With the evolution of computers been able to intelligently and explicitly carried out operations, this integrated into various applications ranging from scientific and non-scientific fields. Specifically, software and systems are created which are applied in the health sector and this software are focused to handle task and operations intelligently which elevate and when compared with the traditional method of how computer software were programmed or built which follows a deterministic and un scale approach of solving problems.

With this evolving improvement and breakthrough in computers, this gave rise and a reason for me to undergo an exploration and understanding of how through data collected and put into computational use can follow an intelligent clinical approach of diagnosing the presence of peptic ulcer which can be either gastric or duodenal which respectively occurs in the stomach or duodenum.

Over the last few years, there have been many people worldwide. Therefore, collecting, managing, and processing huge amounts of medical data in order to give useful information to the doctors to make accurate diagnoses and select the right treatment option is a difficult job for the health sector. In undeveloped and developing nations, there may be a lack of specialists for treating some diseases, less availability of the medical devices, and fewer treatment centers. In such situations, machine learning (ML) has the capacity to help in the detection and treatment of specific diseases and also in ensuring personalized treatment for individuals. ( (Krishna Kant Singh M. E., 2021).

For over past few years, machine learning (ML) has exhibited remarkable abilities in the healthcare sector, owing to the fact that it has the ability to process large amounts of medical data that cannot be handled by humans. Through this, it provides valuable information that helps clinicians in effective planning and making sound decisions about the treatment. This means that doctors are now in a position to save money in terms of the cost of treatment without affecting the level of patient satisfaction. Besides, machine learning is used in drug discovery. (Krishna Kant Singh M. E., 2021).

**1.2 Statement of the Problem**

To explore and understand the peptic ulcer disease with an approach of using machine learning models which can detect and predict the presence of this disease also before a state or condition which is high or worsened.

Using machine learning to predict/diagnose the presence of the two types of peptic ulcer which are:

- Duodenal ulcer
- Gastric ulcer

About two-thirds of the world's population is harboring Helicobacter pylori. Prevalence of H. pylori infection is greater in older people, blacks, and Hispanics, as well as those of low socioeconomic standing in the United States (National Center for Infectious Diseases (U.S.). Division of Bacterial and Mycotic Diseases. Health Communications Activity, 1998/07/01)

From in-depth review of theoretical review of related literature, this study will solve and fundamentally explore the following:

- Compared with traditional clinical diagnosis which is time consuming and sometimes expensive for the lower socioeconomic group, a model which can handle this at no cost is aimed.
- With the availability of this model, swift diagnosis of this disease can be conducted at a comfort location or anywhere without primarily visiting a health facility.
- With this model early detection and prediction of this disease can be achieved.

### 1.3 Aim and Objectives of the Study

This study titled **"a machine learning approach to clinical diagnosis of peptic ulcer**" aims at through the understanding and exploration of the pathology of peptic ulcer, build a model using machine learning algorithms which can diagnose and predict duodenal and gastric types of ulcers.

To achieve this model, the objectives are as follows:

- To Use a synthetic dataset, perform exploratory data analysis (EDA) and create features defined by the disease pathology.
- To develop an intelligent/dynamic web application for patients and clinicians/physicians.
- To deploy and test this model in real time to evaluate its interoperability.
- To make diagnosis and prediction of peptic ulcer seamless through computer software.

### 1.4 Significance of the Study

- This study will not just build machine learning models but explore the pathology of peptic ulcers which will then conjunctively be used with the various required machine learning algorithms.
- The model to be created can be deployed within Godfrey Okoye University which can give assistance to the rapid diagnosis of peptic ulcers.
- Diagnosis can be conducted from anywhere and from a comfort zone at any time through the model/software built from this study.

## 1.5 Scope of the Study

The extent of these programs falls within the preview of Godfrey Okoye University; there are still concerns about how the model will fit beyond.

This study focuses on using synthetic datasets to pipeline the creation of building a model capable of diagnosing duodenal and gastric ulcers.

This synthetic data will follow a defined pathology which will derive the features that will be encoded into this model.

## 1.6 Limitation of the Study

With the close nature for the integrity and confidentiality of patient's records, the availability and gaining access to the data which would have been used to train the model on diagnosing duodenal and gastric ulcers were difficult, hence, an approach of substituting synthetic data to tackle this challenge is adopted. However, I would have applied the use of already defined data.

## 1.7 Operational Definition of Terms

Intelligence: The ability to learn, interact and adjust.

Syllogism: A precedence thinking approach to derive a conclusion.

Synthetic: Something created artificially to mimic/replicate an existing original form.

Heuristic: Actions/reaction for the current state is determined by the past.

Pathology: The study of disease.

Reflexive: For X and Y, both are independent.

Exploratory: Breaking down components to understand and derive something.

Socioeconomic: a ranking based on one's economic and social category which influences their scale in society.

## CHAPTER TWO

## LITERATURE REVIEW

### 2.1 Introduction

Moving forward, the chapter presents aa in-depth and comprehensive review of existing literature relevant to the research topic. The purpose of this review is to acknowledge the works done by some scholars and find out the gap in literature. In the cause of the study, an abstract review of existing literature was carried out inorder to explore and determine what should/can be addressed.

Through analysis of scholarly articles, research papers, and previous project reports, this section identifies the strengths and shortcomings of existing solutions, and highlights the knowledge gaps that remain:

Classical diagnostic techniques are characterized by costly nature, prolonged duration, and dependency on human interference. On the contrary, ML systems are not subject to human limitations, and computation does not get fatigued. Hence, it is possible that the ML-based approach can be employed in order to solve the diagnostic issues related to the presence of an exceptionally high number of patients or unusual patients' characteristics in the sphere of health care. In the development of the ML-based approach to diagnosing diseases, various types of data in the sphere of health care are used (Ahsan M. M., 2020)

In medical domains, artificial intelligence (AI) primarily focuses on developing the algorithms and techniques to determine whether a system's behavior is correct in disease diagnosis. Medical diagnosis identifies the disease or conditions that explain a person's symptoms and signs. Typically, diagnostic information is gathered from the patient's history and physical examination (McPhee & Papadakis, 2010)

It is frequently difficult due to the fact that many indications and symptoms are ambiguous and can only be diagnosed by trained health experts. Therefore, countries that lack enough health professionals for their populations, such as developing countries like Bangladesh and India, face difficulty providing proper diagnostic procedures for their maximum population of patients (Ahsan M. A., 2021). Moreover, diagnosis procedures often require medical tests, which low-income people often find expensive and difficult to afford (Ahsan M. L., 2022).

(Ahsan M. L., 2022) Several aspects could lead to misdiagnosis, including:

- Inadequate symptoms presentation, which are usually quite subtle

- Uncommonness of the disease
- Accidental non-consideration of the disease during diagnosis

Looking at the classification and regression problems, one of the notable machine learning techniques is support vector machines (SVM). SVM is credited to Vapnik and dates back to the end of the twentieth century (Drucker, 1999). SVMs have found wide application in various fields, including facial expression recognition, protein folding, distant homology search, speech recognition, and text classification. Unsupervised machine learning cannot be used for processing unlabeled data. SVM solves this problem by finding a hyperplane that can separate different data clusters and thus make it possible to classify unlabeled objects. However, SVM output is not always linearly separable. To prevent this disadvantage, choosing a proper kernel function and parameters is essential when applying SVMs to the data analysis.(Houssein, 2021)

### 2.2 Theoretical Background

This section discusses the core theories, principles, and technical concepts that support and inform this research.

The following subsections break down the major areas of knowledge relevant to this research:

#### 2.2.1 Overview of Peptic Ulcer Disease

According to CDC.gov, Gastric ulcers refer to those ulcers that occur in the stomach while the duodenal ulcers occur in the duodenum. Prevalence rate is about 90% in duodenal ulcers and 80% in gastric ulcers come from H.Pylori. Peptic ulcer disease is widely abbreviated as PUD. Most people attribute the cause of ulcer or peptic ulcer to be from stress or lack of food but with in-depth study, this does not lead to the primary cause of peptic ulcer. Peptic ulcer is an abdominal pain which occurs in the belly (stomach). This disease has been a world phenomenon.

In respect to the mis-understanding and wrong primary cause of peptic ulcer, the primary cause is stated as follows:

1. Peptic ulcer implies as the penetration of acid into the stomach mucosa thereby breaking the protective layer of the stomach lining which leads to a positive (+) increase of acid in the stomach and a negative (-) defense of the stomach. The process of peptic ulcer formation is illustrated in Figure 1

Acid penetrates the mucosa

↓

Histamine release

↓

Parietal cells release more hydrochloric acid

Figure *1*: Peptic ulcer formation

Within the mucosa there exist the following layers:

- Submucosa
- Muscularis Externa

The submucosa connects the body tissues, nerves and vessels while the muscularis externa which is of three (3) layers of the muscles uses peristalsis to push food to the digestive system. Also within the mucosa is the Gastric pits (special cells), parietal cells, chief cells and the G-cells. The parietal cells secrete hydrochloric acid, the chief cells release pepsinogen, and the G-cells release Gastrin.

Within the mucosa, the Bicarbonate is responsible for the protection of the stomach lining from acid penetration. With that, the prostaglandins play a huge role in the lining of the stomach by regulating the mucosa, controlling acid amount secreted by the parietal cells and makes us feel pain and fever.

The gastric acid digests, liquefies food and sends it to the intestine then food passes through the Esophagus to the stomach and finally the stomach breaks down the food.

Peptic ulcers are commonly classified into the following Two major types and there include:

- Gastric ulcer
- Duodenal ulcer

Gastric ulcers occur in the stomach, while duodenal ulcers occur in the duodenum these are the areas where these ulcers can be found to take effect.

- The etiologies and Symptoms of peptic ulcer are widely attributed by these two:

1. H.Pylori (Helicobacter pylori) and
2. NSAID (non-steroidal anti-inflammatory drugs) usage.

Helicobacter pylori are bacteria that comes into the body system through oral-oral (mouth to mouth) or fecal-oral (faces to mouth) caused by poor sanitation or hygiene which affects the mucosa of the stomach lining. This produces urea's then breaks the urea's to produce:

- Ammonia: neutralizes acid and breaks mucosa
- Carbon dioxide

NSAID are drugs which when taken reduces the production of prostaglandins.

- Symptoms of peptic ulcer:

One of the main symptoms of peptic ulcer is indigestion. Epigastric pain occurs which is (burning, gnawing or dull) and this takes place at the breastbone down to the belly bottom.

Some of the symptoms of Gastric and duodenal ulcers are tabularly delineated as follows:

| **PEPTIC ULCER** | |
|---|---|
| **GASTRIC ULCER** | **DUODENAL ULCER** |
| Food meals worsen pain after one (1) to two (2) hours. | Pain gets relief (better) after three (3) to four (4) hours of meals. |
| A dull pain (aching) is experienced. | A gnawing pain is experienced at the mid of the night. |
| Weight loss. | Normal weight loss. |
| Severe vomit: "coffee ground" or bright blood (red). | Severe tarry or dark stool. |

Burning > gnawing > dull in respect to pains which occur.

The early diagnosis and prediction of this disease is important to prevent severe, abdominal and disturbing pain which occurs in the stomach and can also lead to stooling.

### 2.2.2 Conventional Methods of Peptic Ulcer Disease Diagnosis

Traditional diagnostic methods include scope of the stomach (EGD) which uses a visual scan (endoscope) to examine the stomach, and this takes a process of:

- Looks from the esophagus, then
- The stomach and then
- The duodenum

An alternative procedure includes H. pylori tests that include the use of blood, stool, and the breath test. In the breath test, the patient consumes a pill containing urea; if the H. pylori is detected in the body, then the bacteria breaks down urea to ammonia and carbon dioxide. Breath samples are collected in order to measure the amount of carbon dioxide. High amounts of carbon dioxide indicate infection by H. pylori. In synopsis the two methods are Esophagogastroduodenoscopy (EGD) and H.Pylori test.

- Challenges of manual diagnosis: manual diagnosis can be time consuming, prone to human errors and can also lead to misdiagnosis.

### 2.2.3 Machine Learning and Its Applications in Healthcare

Machine learning refers to using algorithmic approaches—commonly referred to as learners, predictive models, or estimators—to categorize data or predict an outcome (Sanjay Basu, 2020).

Machine learning is an approach to (1) learn (2) complex patterns from (3) existing data and use these patterns to make (4) predictions on (5) unseen data (Huyen, 2022).

(Huyen, 2022): To summarize, Machine Learning is great for:

• Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better.

• Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.

• Fluctuating environments: a Machine Learning system can adapt to new data.

• Getting insights about complex problems and large amounts of data.

Finally, A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E (Mitchell, 1997).

Machine learning can be categories into these three (3) types which are:

- Supervised learning

- Unsupervised learning
- Reinforcement learning

Supervised learning is where you have input variables (X) and an output variable (Y ) and you use an algorithm to learn the mapping function from the input to the output (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016). In terms of predictive models, it involves approximating the mapping function to make predictions regarding new input data, whereby for the new input data (X), we would want to predict the output variables (Y). In this case, it involves what is referred to as supervised learning since the teacher supervises the training process. The algorithm learns from a data set, where the answers are already known. The algorithm makes predictions and gets corrected by the teacher. The training process ends once the algorithm achieves sufficient accuracy. (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016).

(Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016) The concept of unsupervised machine learning refers to learning where the algorithm is left alone to find structures present in the data. Unsupervised learning problems can further be divided into two types: clustering and associations. (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016).

Reinforcement learning addresses the question of how an autonomous agent that senses and acts in its environment can learn to choose optimal actions to achieve its goals (Mitchell, 1997). Reinforcement learning algorithms are related to dynamic programming algorithms frequently used to solve optimization problems (Mitchell, 1997).

- Applications in healthcare: With data from health records, in this present century, machine learning algorithms have the capability to train computational models which can efficiently diagnose diseases either visually or through strings of input. Medical imaging techniques are also applied, Algorithms for machine learning have played a vital role in detecting diseases through the use of various methods of medical imaging, such as MRI and CT scans, due to the patterns which can be recognized only through algorithms and not through naked eyes drug discovery (Shivahare, et al., 2024).
- Benefits: automation, speed, accuracy, and scalability.

### 2.2.4 Machine Learning for Peptic Ulcer Disease Detection

In the diagnosing of diseases, with the evolution, a range of algorithms used in machine learning have been applied and some of these algorithms reviewed are support vector machine, Naïve Bayes Algorithm, Decision Tree algorithm, Logistic Regression and Random Forest Classification.

As earlier stated in this study, synthetic datasets will be primarily used to implement this research.

An overview of algorithms applied are reviewed as follows:

- **support vector machine:** Optimal separating hyperplane is determined by counting the number of nodes present at the boundary of the two categories. Margin represents the difference in values between the two categories. Better performance of the classifier can be achieved by having a higher margin of safety. Boundary data is represented by the support vectors. Support Vector Machines are useful for both regression and classification problems (K. Rajendran, 2020).

  The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016).
- **Naïve Bayes Classifier:** Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016). The Naive Bayes classifier is used to classify instances both in binary and multiclass classification problems. The Naive Bayes classifier is easiest to describe using binary and categorical values of features. It is called "Naive Bayes" or "idiot Bayes" since the calculation of the probability of each hypothesis is made simple enough to enable their computations. Rather than calculating the joint values of $P(d1, d2, d3 \mid h)$ for all features, these probabilities are considered to be conditionally independent from one another given the value of the target and are calculated as $P(di \mid h) \times P(d2|h)$ and so on (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016)
- **Decision Tree algorithm:** Classification and Regression Trees (CART) refer to a term coined by Breiman to refer to decision tree methods useful for classification and regression models (Brownlee, Master Machine Learning Algorithms: Discover How

They Work and Implement Them From Scratch, 2016). The decision tree algorithm represented as CART can be represented in a model as follows: The algorithm takes a binary tree representation. Each node in the tree represents one input variable (x) and a cut-off point for that particular variable when it happens to be numerical. Terminal nodes (leaves) contain the output variable (y) which is used for predictions. Greedy algorithm such as recursive binary splitting controls the splitting of the input space (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016). The output variable (y), which is used for making predictions, is found at the leaf nodes (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016). The greedy approach called recursive binary splitting is used to segment the input space (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016). The decision tree algorithm has a stopping criterion.

- **Logistic Regression:** Logistic regression models the relation through an equation, similar to linear regression. The input values (x) are weighted or multiplied by some coefficients and then summed up to make a prediction of an output value (y).. The logistic regression uses a logistic function which can be written as:

$$\frac{1}{1 + e^{-value}}$$

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform (Brownlee, Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch, 2016).

- **Random Forest Classification:** The random Forest classification works another technique consists of generating many Decision Tree learners using random samples of the features and then taking the average of their predictions. The random forest algorithm can be implemented using a sci-kit (SK learn) library of the "RandomForestClassifier" class algorithm and this is a more optimized approach than the decision tree.

- In data science and machine learning, data is explored to understand and generate insight which will further be condensed or expanded to extract features. An approach to this is exploratory data analysis and feature engineering. In the case/cause event of this research, exploration of the records contained in the synthetic dataset will be applied as it will give climax to the extraction of features through a concatenation of feature engineering.

  Exploratory data analysis and feature engineering will be performed to handle missing data in line with this approach, data cleaning, feature scaling, feature selection, transformation is performed to juxtaposed data set in order not to get an in-balance dataset, biased data which can lead to an overfitting or underfitting dilemma and can affect the performance metrics, accuracy of the trained model. Encoding

  Each phase of preprocessing of the dataset is critically essential as the development of the model depends on the use of synthetic dataset else can lead to a biased outcome of prediction. Graphically, the preprocessing pipeline is expressed in Figure 2 below:

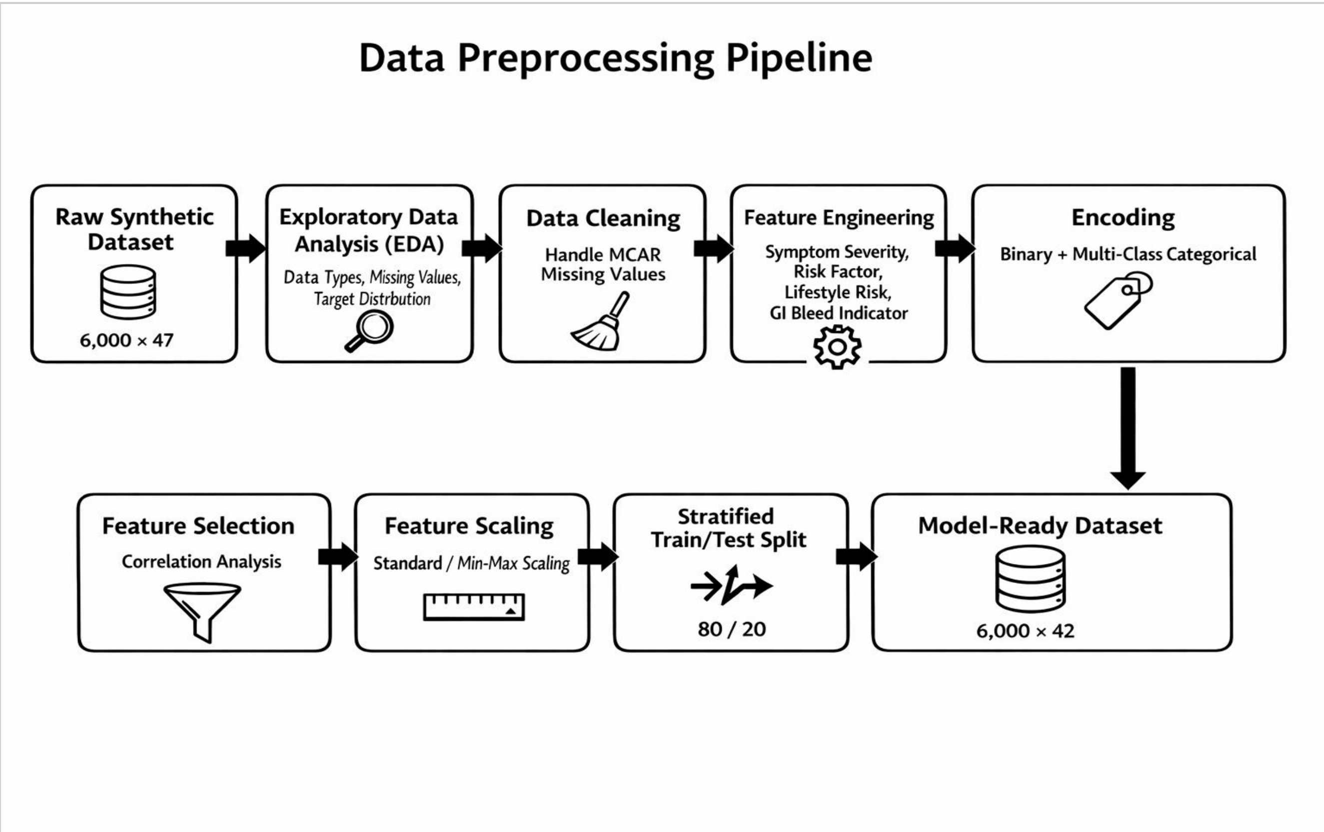


Figure 2: Data preprocessing pipeline

Data preprocessing techniques such as standardization and min max scaling will be performed on the data set. This was considered necessary as it aids it in giving a unique scale to get a balanced feature which is free from biasness. This will play a major role in uniformly distributing data which is dispersed and has different levels of variability. Assuming there exist N columns and for each N-1 column selected to extract a feature these columns contain different range of values, it can affect the training process of the peptic ulcer model since computer machines (machine learning) are heavily dependent on data which are numeric data type.

Preprocessing of the synthetic data was necessary cause of the following:

- The data obtained must be prepared as it can partially be classified as information
- Machin learning algorithms are dependent requires an input of digits or numeric data type.
- The performance and performance metrics of the model depend on the data.
- The problem that has been solved is predictive modeling.

### 2.3 Review of Related Works

Non-steroidal anti-inflammatory drugs (NSAIDs) are frequently prescribed for musculoskeletal disease management; however, NSAID use has been associated with the development of peptic ulcers (PUs). PU prediction in NSAID users is critical for the prevention of adverse events, such as bleeding or perforation. Here, we introduce and validate a deep learning model for PU prediction within 180 days after NSAID exposure based on longitudinal EHRs. Our population consists of 125,930 NSAID users who received at least 7 days of NSAID prescriptions. Various data, including laboratory test results, drug exposures, and demographic factors, were leveraged for training various ML and DL algorithms, such as random forests, gradient boosting machines (GBM), recurrent neural networks (RNN), long short-term memory networks (LSTM), gated recurrent units (GRU), and transformers. Endoscopy report free-text was used to determine PU occurrence rate. Among the algorithms, the GRU outperformed others, with an AUROC of 0.941 in internal validation and 0.964 in external validation. Hemoglobin level, NSAID exposure duration, and aspirin were found to be significant predictors. Risk scores revealed a dramatic increase in risks starting from about two months before PU diagnosis. We provided highly accurate models for NSAID-induced PU prediction using longitudinal EHRs. Further studies can concentrate on improving existing models and applying them to other databases (Joo Seong Kim, 2024).

Machine learning (ML) methods have found widespread application among researchers and practitioners for the diagnosis of a variety of diseases. In this section, a number of prominent ML-based disease diagnostic approaches will be reviewed and discussed that have found importance and significance. These include some key diseases such as heart diseases, kidney diseases, breast cancer, diabetes, Parkinson's disease, Alzheimer's disease, and COVID-19 (Song Y, 2021).

Peptic ulcer disease (PUD) is one of the most prevalent ailments of the stomach and duodenum caused by H. pylori infection that is also associated with multiple complications and co-morbidities. Early prediction of peptic ulcers is vital for the prevention of such adverse results. So far, the application of machine learning methods has shown success in the efficient prediction of various health problems in the medical environment. Therefore, this study will attempt to create a model using machine learning techniques and risk factors of PUD (Nopour, 2025).

Recently, many automated analysis models based on different supervised learning approaches have been created by researchers. The early detection of disease can help to decrease the mortality rate of the disease. In this project, a new recognition system is being introduced, which is an automated disease diagnosis model based on machine learning approach. Here, in this study, patient records having 41 different diseases have been taken into account, out of which, 95 symptoms of 132 symptoms highly related to the diseases selected have been identified and optimized. In the proposed recognition model, the data set will be arranged within Tkinter graphical user interface; the analysis will be done with the help of which disease prediction will be done. Prognosis will be made by applying decision tree and Naive Bayes algorithms. Precise analysis of the medical database helps to predict the disease for early detection and better service delivery (A. Divya, 2022).

Peptic ulcer disease refers to abdominal pain and discomfort. The main causes include infection with H. Pylori and consumption of medicines like aspirin. Peptic ulcers can cause severe complications if not treated in time. The second major cause of cancer deaths is gastric cancer. Fast diagnosis is required to treat peptic ulcers as soon as possible. Combining artificial intelligence with medical expertise gives us the chance to make fast and accurate diagnosis. The main objective of this research is to create a novel, cost effective and fast diagnosis of peptic ulcers using fuzzy expert system. (Saeedreza Arab, 2021).

A computational approach based on a Fuzzy Inference System (FIS) is suggested in this study for the evaluation of peptic ulcer. The Fuzzy Inference System was produced with Fuzzy C-Means and tuned using the Adaptive Neuro-Fuzzy Inference System model (ANFIS) (Saeedreza Arab, 2021).

## 2.4 Summary of Literature Review

This section brings a condensed/detailed summary of the key findings from the literature. Presented in tabular format

| S/N | AUTHOR(S) | CONTRIBUTION OR (WORK DONE) | TECHNIQUE OR METHODOLOGY | LIMITATIONS |
|---|---|---|---|---|
| 1. | Joo Seong Kim (2025 | Prediction of NSAID-induced peptic Ulcers using longitudinal health records. | Used GRU, RNN, LSTM, Random forests, and transformers. | The GRU model was able to achieve the highest performance metrics with an AUROC of 0.941. |
| 2. | Nopour 2025 | Developed a prediction model which is based on factors to achieve preventive goals. | Five (5) ensemble Machine learning algorithms were utilized and are as follows: XG-Boost (Extreme Gradient Boosting), Random Forest, Ada-Boost, LightGBM, Cat-Boost. | Demonstrated machine learning promises for an efficient prediction in healthcare settings. |
| 3. | A. Divya (2022) | Automated diagnosis model for 41 different diseases using patient symptoms. | Decision tee and Naïve Bayes algorithm with an interface integration of Tkinter GUI. | Precise analysis of medical database aids in providing better patient care. |

| | | | | |
|---|---|---|---|---|
| 4. | Saeedreza Arab (2021) | An inexpensive and quick diagnostic system for peptic Ulcer disease was designed. | Used a Fuzzy Inference system (FIS) which was turned with ANFIS (Adaptive Neuro-Fuzzy Inference System). | With a combination of AI and medical knowledge faster and more accurate diagnosis is achieved. |

## 2.5 Research Gap

There is a cumulative significance which has been made in the application of machine learning to health diagnosis. After qualitative reviews of related literature, this study reveals the specific limitations and gaps that this study aims to address.

Joo Seong Kim (2025), successfully utilized complex deep learning architecture like the GRU And LSTM to predict NASID type of ulcers using longitudinal electronic health records. In this study, data sources are placed on massive, private hospital datasets, which poses a significant challenge with respect to data accessibility and patient confidentiality. Also, Nopour (2025) and A. Divya (2022) focused on general predictive models which spans across 41 different diseases and lacks a required in-depth differentiability to explicitly diagnose/predict between Gastric and Duodenal ulcers which relies on a distinct pathological feature.

This study addresses these gaps through the following:

- Alternative to handle data scarcity: A challenge faced due to data policy/confidentiality is resolved through the adoption of a synthetic dataset.
- Pathological focus: as contained and reviewed from Nopour (2025), this study will have specific targets for Duodenal and Gastric Ulcers which will occur during the feature engineering phase of data preprocessing of model training.
- Accessibility to all socio-economic groups: the proposed model aims to provide cost-effective and diagnostic software which aids early detection without a primary criterion of clinical visits and will have a wider span of demographic outreach.

## CHAPTER THREE
## SYSTEM ANALYSIS AND DESIGN

### 3.0 Introduction

The peptic ulcer system, codename “PepticSense” for the web application went through a stage of analysis and design before been implemented. This chapter spans the requirements, design methodology using the Agile life Cycle of Software Development (SDLC) and Object-Oriented Analysis and Design (OOAD), and the analysis of the existing system, that is, the traditional (conventional) methods currently used to diagnose peptic ulcer disease whose limitations gave rise to “PepticSense”. The system behavior is later described in this chapter using an activity diagram.

### 3.1 System Analysis

This section involves understanding the existing system, the traditional (conventional) methods currently used to diagnose peptic ulcer disease and their limitations. It delineates what PepticSense does and outlines the issues to be solved.

### 3.1.1 Analysis of the Existing System

In the context of this study, the existing system refers to the traditional (conventional) methods that are presently used to diagnose peptic ulcer disease in clinical practice. Before a patient is confirmed to have a gastric or duodenal ulcer, he or she typically passes through a manual, facility-based diagnostic pathway that relies heavily on the availability of medical specialists, laboratories, and specialized equipment. Drawing on observations from clinical practice and the literature reviewed in Chapter Two, the traditional diagnosis of peptic ulcer is carried out through the following methods:

1. **Clinical Consultation and History Taking:** The patient reports to a health facility and pays to create or retrieve a case file from the Health Record and Bio-Statistics Unit before meeting a consultant or physician. The physician arrives at a provisional diagnosis through a series of pathological questions covering the patient’s medical history and symptoms (such as epigastric pain and its relationship to meals), together with a physical examination, and may then refer the patient to the laboratory for further tests.

2. **Endoscopy (Esophagogastroduodenoscopy, EGD):** This is the primary confirmatory method, in which a flexible tube fitted with a camera (an endoscope) is passed through the mouth to visually examine the esophagus, then the stomach, and finally the duodenum for the presence of ulcers. A tissue sample (biopsy) may also be taken during the procedure for further analysis.
3. **Helicobacter pylori (H. pylori) Testing:** Because H. pylori infection is a leading cause of peptic ulcers, laboratory tests are used to detect it. These include blood (serology) tests, stool antigen tests, and the urea breath test, in which the patient consumes a urea preparation and the carbon dioxide released when the bacteria break down the urea is measured from breath samples, with high levels indicating infection.
4. **Radiological Investigation:** Imaging techniques such as the barium swallow (upper gastrointestinal series) are also used, where the patient ingests a contrast material that coats the lining of the digestive tract so that ulcers can be visualised on X-ray images.

These traditional methods are largely manual and must be carried out by trained medical personnel within hospitals, clinics, and laboratories. They therefore represent the current standard of care against which the proposed PepticSense system is analysed and compared.

### 3.1.2 Limitations of the Existing System

Although the traditional methods of diagnosing peptic ulcer disease described above remain the clinical standard, they present a number of limitations, including the following:

1. **Invasiveness and discomfort:** Procedures such as endoscopy are invasive and are often uncomfortable or distressing to patients, which discourages many people from seeking early diagnosis.
2. **High cost:** Endoscopy, laboratory tests, and radiological investigations are expensive, placing them out of reach for many patients, particularly those in the lower socioeconomic group.
3. **Time-consuming pathway:** Registering at a facility, consulting a physician, waiting for laboratory or imaging results, and returning for interpretation is a lengthy process that can delay diagnosis and treatment.
4. **Dependence on specialists and equipment:** These methods require trained gastroenterologists, endoscopy machines, and well-equipped laboratories, which are scarce or entirely unavailable in many rural and resource-constrained areas.

5. **Susceptibility to human error:** Manual examination and interpretation are prone to human error and can lead to misdiagnosis, especially where symptoms are subtle or overlap between gastric and duodenal ulcers.
6. **Limited accessibility:** Because diagnosis must take place within a health facility, patients in remote locations have limited access, and early or preventive screening from a comfortable, non-clinical location is not possible.

These limitations justify the need for a new system that is non-invasive, affordable, fast, and accessible from any location.

### 3.1.3 Analysis of the Proposed System

- ❖ This application is designed to run on the web and can be used by the public and verified clinicians/physicians. Intelligently interact with the user during the process of diagnosis. It will intersect a kind of human in a loop to actively collect data to make inference. Connectivity through the internet thereby providing access from anywhere and at any time.
- ❖ A well-defined navigation/experience that improves on the traditional manual diagnostic pathway, which is invasive, costly, time-consuming, and dependent on specialists and equipment. The software will not only be limited to desktop computers.
- ❖

## 3.2 System Requirements Specification (SRS)

To meet the requirements and specifications of this software, the functional and nonfunctional requirements are specified providing guidance for the software inspection which provides verification and verifiability.

### 3.2.1 Functional Requirements

- ❖ FR-1: User Authentication & Role Management

FR-1.1: The system shall support three distinct user roles: General User (Patient), Physician/Clinician, and Administrator.

FR-1.2: Users shall be able to authenticate via standard Email/Password credentials or Google Sign-In (OAuth2).

FR-1.3: The system shall use JWT (JSON Web Tokens) for secure, stateless session management.

FR-1.4: Administrators shall be able to provision Physician accounts and generate secure initial setup links.

- ❖ FR-2: Interactive Diagnostic Assessment

FR-2.1: The system shall provide a multi-step wizard to collect patient data, including age, symptom profile, and clinical history.

FR-2.2: The system shall maintain a conversation log/session to provide real-time feedback and progress tracking during the diagnostic process.

FR-2.3: The interface shall provide clinical context for each symptom category (Pain, Digestive, Alarm features) to educate the user.

- ❖ FR-3: ML-Driven Inference & Diagnosis

FR-3.1: The system shall process clinical inputs through a trained Random Forest model to classify the risk of Gastric Ulcers, Duodenal Ulcers, or No Ulcer detected.

FR-3.2: The system shall calculate a raw probability score (confidence) for each prediction.

FR-3.3: The backend shall calculate feature importance weights to identify which clinical factors most heavily influence the result.

- ❖ FR-4: Explainable AI (XAI) & Reporting

FR-4.1: The system shall generate role-appropriate explanations. General users receive non-technical summaries, while Physicians receive technical clinical assessments.

FR-4.2: Results shall display a "Clinical Terms Map" to correlate model features with human-readable medical terms.

FR-4.3: The system shall flag "Alarm Features" (like: GI bleeding, vomiting blood) as high-priority medical warnings.

- ❖ FR-5: Physician Collaboration & Data Sharing

FR-5.1: General users shall be able to request a specific Physician to review their diagnostic report.

FR-5.2: Physicians shall have a dashboard to approve or reject access requests.

FR-5.3: Upon approval, the system shall grant the Physician secure access to the detailed diagnostic report.

FR-5.4: The system shall provide real-time notifications for request statuses (Received, Approved, Rejected).

- FR-6: Administrative Oversight

FR-6.1: Administrators shall have a dashboard to monitor system-wide statistics.

FR-6.2: The system shall log administrative actions (user creation, modification of physician email, update of medical license number, activation of accounts, deactivation of accounts) for audit purposes.

### 3.2.2 Non-Functional Requirements

- NFR-1: Performance & Efficiency

NFR-1.1: The ML model must be loaded into memory once during the Django apps.ready() lifecycle to ensure sub-second inference latency.

NFR-1.2: Frontend components must utilize Angular's Signal-based reactivity to ensure a highly responsive UI during the diagnostic wizard.

- NFR-2: Security & Data Privacy

NFR-2.1: All API communication must be protected by CORS (Cross-Origin Resource Sharing) policies.

NFR-2.2: Physicia’s sensitive credentials need to be stored in symmetric encryption (Fernet), not in plain text.

NFR-2.3: There needs to be strict implementation of Role-Based Access Control (RBAC) in the API layer.

- NFR-3: Usability & User Experience

NFR-3.1: The UI and UX should adhere to the Eight Golden Rules of Interface Design by Ben Shneiderman.

- NFR-4: Reliability & Clinical Guardrails

NFR-4.1: The output from each diagnostic must come with a disclaimer stating that this application is for decision-making purposes and not a replacement for medical advice from professionals.

NFR-4.2: The system should efficiently cope with the lack of data (for example, the unknown Helicobacter pylori status) by making an inference about the patient's risk based on other clinical data.

❖ NFR-5: Scalability & Portability

NFR-5.1: The software product must be deployed in Docker containers.

NFR-5.2: The architecture must remain strictly decoupled (Django REST Framework for logic, Angular for presentation).

### 3.3 System Design Methodology

Two design approaches are followed. There exists a myriad of design methodology like waterfall, prototyping methodology, throwaway prototyping methodology, v model methodology, spiral methodology, extreme programming methodology, and lots more. In the cause of this study, the agile process methodology of SDLC is selected for the project which brings the work breakdown structure of the project as follows:

i. This project was planned on different levels which in one of them contains the lower bound of how data will be obtained.
ii. Staged through a modeling of analysis, design, and then implementation before the peptic ulcer system can be delivered.
**iii.** Continuous development, integration and deployment are pipe-lined after software inspections of specifications and requirements.

Also, object-oriented analysis and design (OOAD) is used to follow structure implementation principles and model the system behaviors using the Unified Modeling Language.

In addition to Agile SDLC and OOAD, which respectively governs the software development cycle (SDLC) and structural construct of the system, the **Cross-Industry Standard Process for Data Mining (CRISP-DM)** was adopted to guide the machine learning lifecycle of the PepticSense engine.

The CRISP-DM methodology provides an organized process which involves/includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The six steps above correspond directly to what was done in this research where business understanding relates to the problem definition provided in Chapter 1, data understanding and data preparation relate to the processes of exploratory data analysis (EdA), data cleaning, feature engineering, feature encoding, feature selection, and scaling as discussed in Chapter 4; modeling and evaluation relate to the training and hyperparameter tuning of the Random Forest classifier and its performance evaluation; and deployment involves incorporating the trained classifier in the Django back-end of the live web application. By applying the CRISP-DM methodology together with Agile development methods and Object-Oriented Analysis and Design (OOAD), the data science process is able to achieve the same level of rigour and repeatability like that of software engineering.

### 3.3.1 Justification for Methodology

- An iterative approach, adaptability, and user-centered design are the key factors that the Agile methodology highlights. As far as the described project concerns, the Agile process model can be applied to it due to the following criteria:
  - Iterative Integration of Machine Learning Model: The core of the application is the machine learning model (UlcerModelEngine). Machine learning model development usually lacks linearity and involves continuous tuning and validation. Thus, Agile will allow the updating and re-integration of the machine learning model.
  - User-Centered Design: Following the 8 Golden Rules of Ben Shneiderman, it is important to collect user feedback. With the help of Agile, one can improve Diagnostics Wizard and Physician Dashboard basing on user interactions and feedback.
  - Mitigation of Risks: Agile helps to discover technical risks (CORS, Google Auth integration, model load time, etc.) at the earlier stages of development. It is especially important for the system working with medical data.
  - Management of Decoupled Complexity: Working on a decoupled architecture (with Django REST Framework (DRF) backend and Angular frontend) is the process that requires careful coordination of efforts between various system layers. Therefore, the Agile phased approach, which allows to maintain the synchronization between different layers through iterative updating of API contract, is quite suitable in this case. By specifying JSON interfaces at the beginning of every iteration, one can prevent any problems with integration because the updated interfaces will always match the current backend logic and machine learning model results.

Moreover, application of Object-Oriented Analysis and Design (OOAD) allows the system to have a solid and well-organized logical basis. By means of OOAD, one can represent real-world medical entities in software objects:

**1. Modularity**

OOAD helps to split the system into several discrete, self-contained modules. According to the proposed architecture, each module operates separately from the others. The accounts module controls the UserAccount, PhysicianProfile, and Role objects and works separately from the diagnosis logic. The diagnosis module controls the DiagnosticInput and PredictionRecord objects. The physician requests module controls the whole process of the request and its approval and operates separately as a standalone module.

**2. Reuse**

Given well-defined objects, code reuse between various components within the application is possible:

i. Service Reuse: The AuthService is used by the DiagnosisComponent, the NavbarComponent, and the PhysicianSelectorComponent.

ii. Logic Reuse: The UlcerModelEngine within the backend includes all logic for machine learning and can be called from the MakePredictionView.

**3. Clarity and Mapping**

i. Class Responsibility: Prediction Record holds information about its own outcome and the connection it has with the Diagnostic Input. Prediction Record does not need to have an idea of how the AuthUser has been authenticated.

ii. Inheritance and Composition: The system shows good use of composition where in the system, PhysicianProfile is part of User Account.

4. **Improved Maintenance (Encapsulation)**

### 3.3.2 Method of Data Collection

Data and diagnostic rationale for the system were developed using secondary research and knowledge engineering techniques rather than primary patient data collection due to the need for confidentiality described in Section 1.6. Clinical criteria, risk factors, and symptoms were extracted from a wide variety of authoritative sources, which include peer-reviewed clinical guidelines on the management of peptic ulcer disease, standard textbooks of gastroenterology on the pathophysiology of gastric and duodenal ulcers, Helicobacter pylori infection, and NSAID ulcers, and public health sources from the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), with emphasis on information on the epidemiology and prevention of Helicobacter pylori infection. This information has been used to determine the feature set, labeling classes, and decision criteria used in the system.

### 3.4 System Modeling Using UML

Unified Modeling Language (UML) is used to depict the diagrams showing the functionalities and structures of the system.

#### 3.4.1 Use Case Diagram

The use case diagram for the PepticSense web application, as shown in Figure 3, describes the main interaction between the system and its three main actors: the General User (Patient), the Physician/Clinician, and the Administrator. The General User starts the process of diagnosis by either registering himself/herself or logging into the system; completes the assessment; views the prediction results; and, in case of necessity, asks for the review by a physician. The Physician/Clinician interacts with the system by analyzing the information provided in the report, dealing with access requests, and making decisions based on diagnostic results. The Administrator controls the work of the platform by working with the users and the physicians and observing the actions in the system. Thus, it can be concluded that the system under consideration is an intelligent clinical support system with role-based approach.

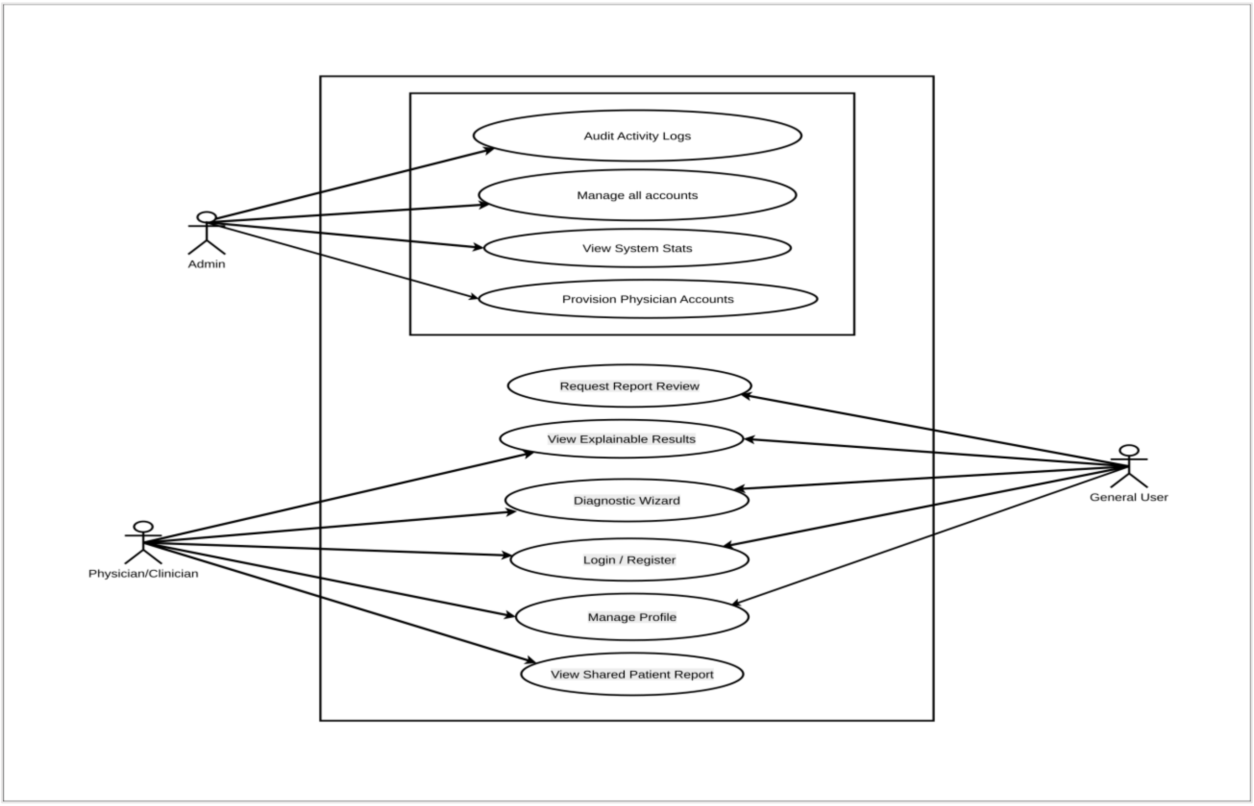


Figure 3: Pepticsense use case diagram

### 3.4.2 Activity Diagram

The end-to-end process of diagnosing in the PepticSense system is illustrated by the diagnostic activity flow depicted in Figure 4. It starts once the user is logged in or registered and started the assessment process. Afterward, the user is guided through a multi-stage diagnostic wizard where he/she should provide demographics information, symptoms, and clinical history. All inputs provided are validated at client and server levels before sending to the machine learning inference engine, where feature vector is processed and a prediction of ulcer type (Gastric, Duodenal, or None) along with its probability and explanation of feature importance is produced. Next, depending on the result and the role of the user, there will be further actions: general users will get the report understandable to them with the recommendations about their lifestyle and an opportunity to ask for the doctor's opinion, whereas physicians and administrators can access the actual technical clinical assessment straight from their dashboard. The decision nodes on the way control alarm features (such as GI bleeding or hematemesis) and manage the approval or rejection of report-sharing requests, ensuring the diagnostic workflow remains secure, role-aware, and clinically meaningful from start to finish.

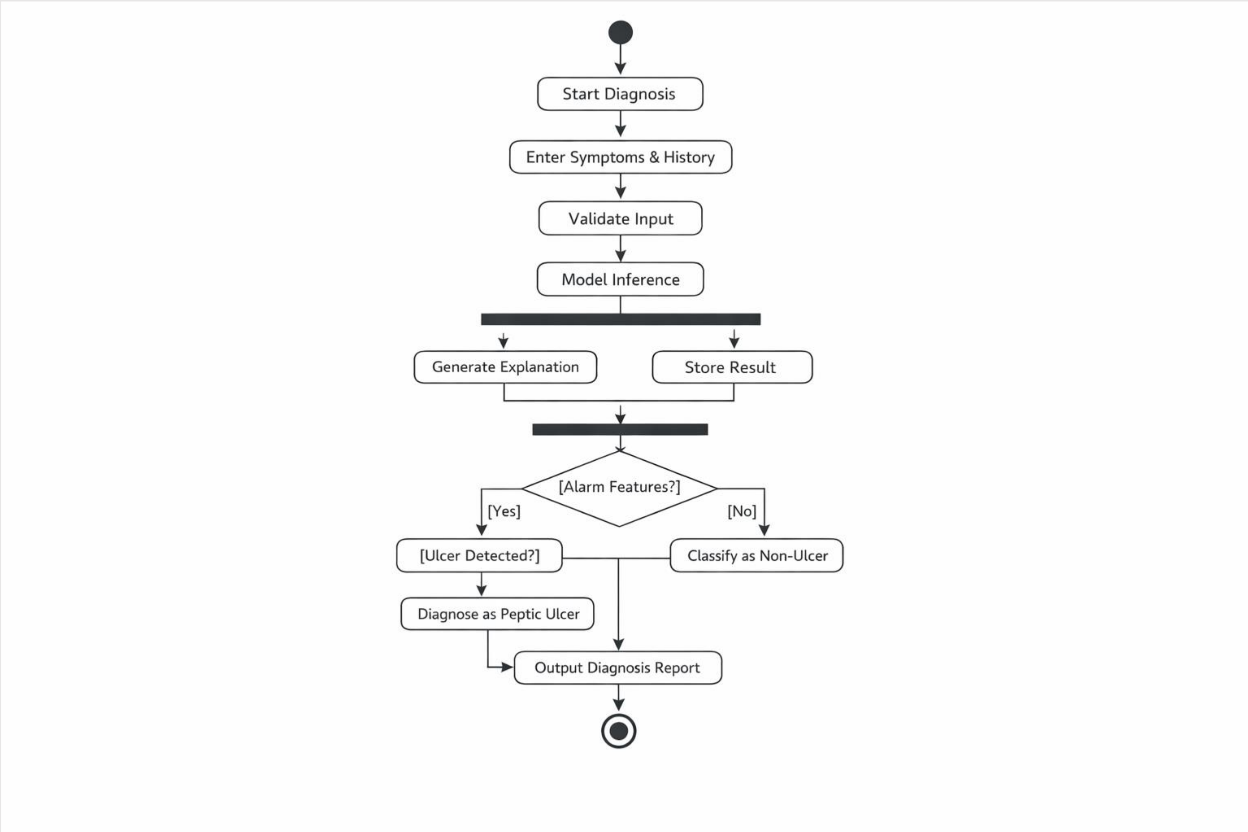


Figure 4: Pepticsense diagnosis Activity diagram

### 3.4.3 Class Diagram

Figure 5 describes the static, object-oriented architecture of the PepticSense system. The UserAccount object includes identity data, credentials, and the Role attribute (values GENERAL, PHYSICIAN, ADMIN), which controls the access level. It inherits a PhysicianProfile object that holds additional information about the physician's credentials and license. The diagnostic workflow is described by the DiagnosticInput object (holding patient demographic information, symptoms, and medical history) that has a one-to-one relationship with a PredictionRecord object (including the predicted ulcer class, confidence value, and explanation). At the same time, the PredictionRecord object is connected to an ExplainabilityMetrics object holding feature importance weights and a mapping between medical terms. Collaboration is implemented via two objects: PhysicianRequest that handles requests for sharing reports and ReportAccess object controlling row-level access to a particular patient report. Moreover, ConversationLog captures the conversation logs of each diagnostic process, while the UlcerModelEngine holds the trained Random Forest model used for prediction.

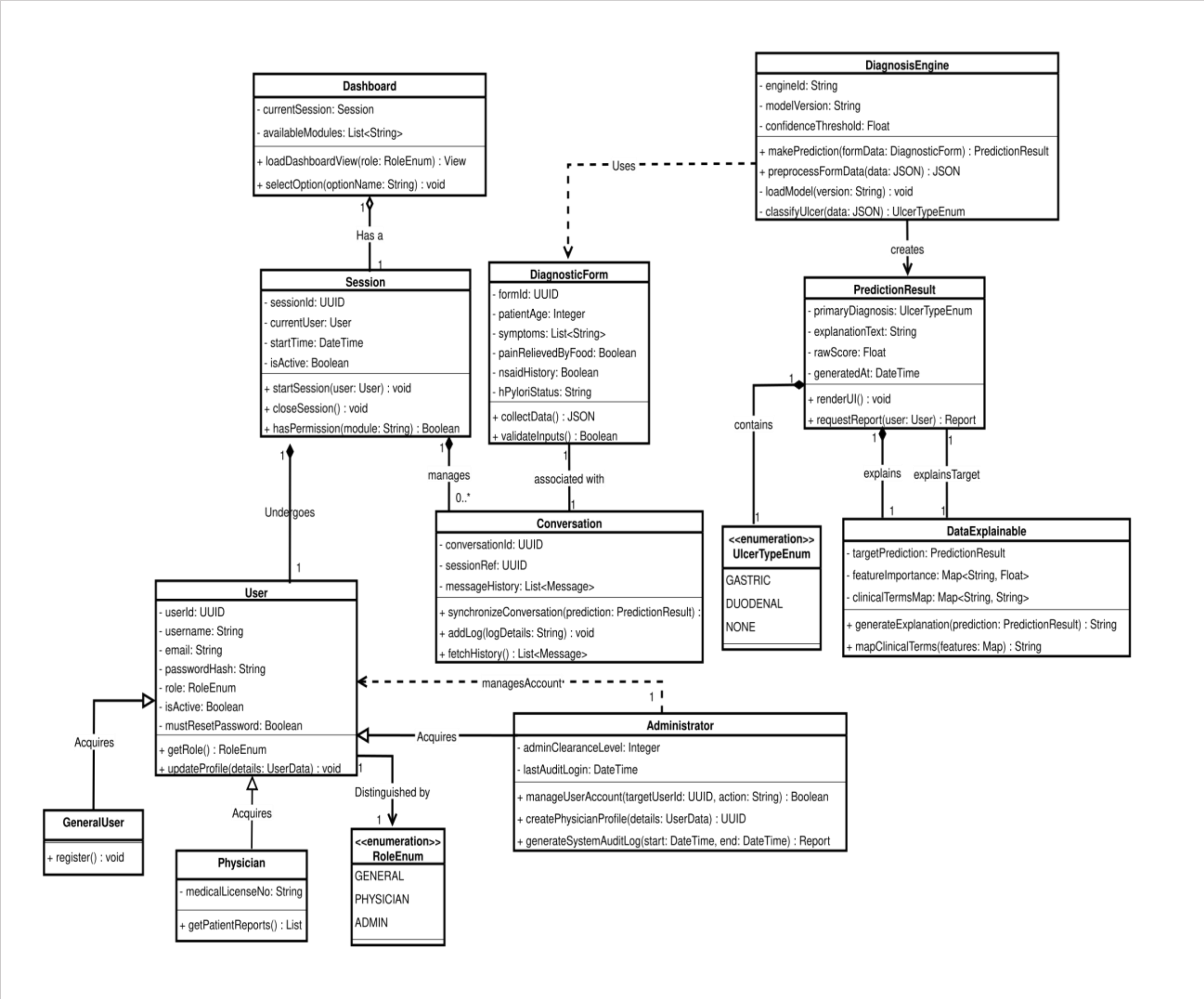


Figure 5: Pepticsense class diagram

### 3.5 System Design

The architecture of the system supports several layers based on the user roles, which are discussed in detail in the upcoming sections.

#### 3.5.1 Input Design

1. Diagnostic Assessment Form

   Demographic Data: Age of the patient.

   Symptom Checklist: Multiple select option for symptoms such as Epigastric Pain, Nausea, Vomiting, Melena, Hematemesis, Bloating and more.

History Data: Boolean flags indicating history of using NSAIDs, pain relief after eating (which is the characteristic symptom of Duodenal Ulcer) and Helicobacter Pylori infection.

2. Physician Account Creation

   Administrative Form meant for use by authorized personnel in collecting professional information, which include Full Name, Email, Username, and Medical License Number.

3. Physician Response

   Physician Response Form: A modal form that allows the physician to give consent or deny access to reports, and optionally includes a text box for clinical comments.

❖ **VALIDATION TECHNIQUES**

1. Client-Side (Angular): Uses Reactive Forms for immediate validation of required fields and valid email format.
2. Server-Side (Django REST Framework Serializers): PredictionRequestSerializer is used to ensure data accuracy before interaction with the machine learning algorithm. PhysicianCreateSerializer performs uniqueness validation for email and medical license number in the database. Data Collection Format: Using JSON (JavaScript Object Notation).

**Data Collection Format:** Using JSON (JavaScript Object Notation).

### 3.5.2 Output Design

This system will give a unique output layer depending on the role of the user, following the rule of informative feedback.

❖ **Primary Diagnosis Report**

1. Predictions Summary: This gives the predicted classification (GASTRIC, DUODENAL, or NONE) together with the percentage of confidence level (raw score).
2. Explanation Metrics: Gives an overview of the feature importance highlighting the symptoms and history elements that have impacted the prediction most. These results, as presented to the general user, are shown in Figure 6.

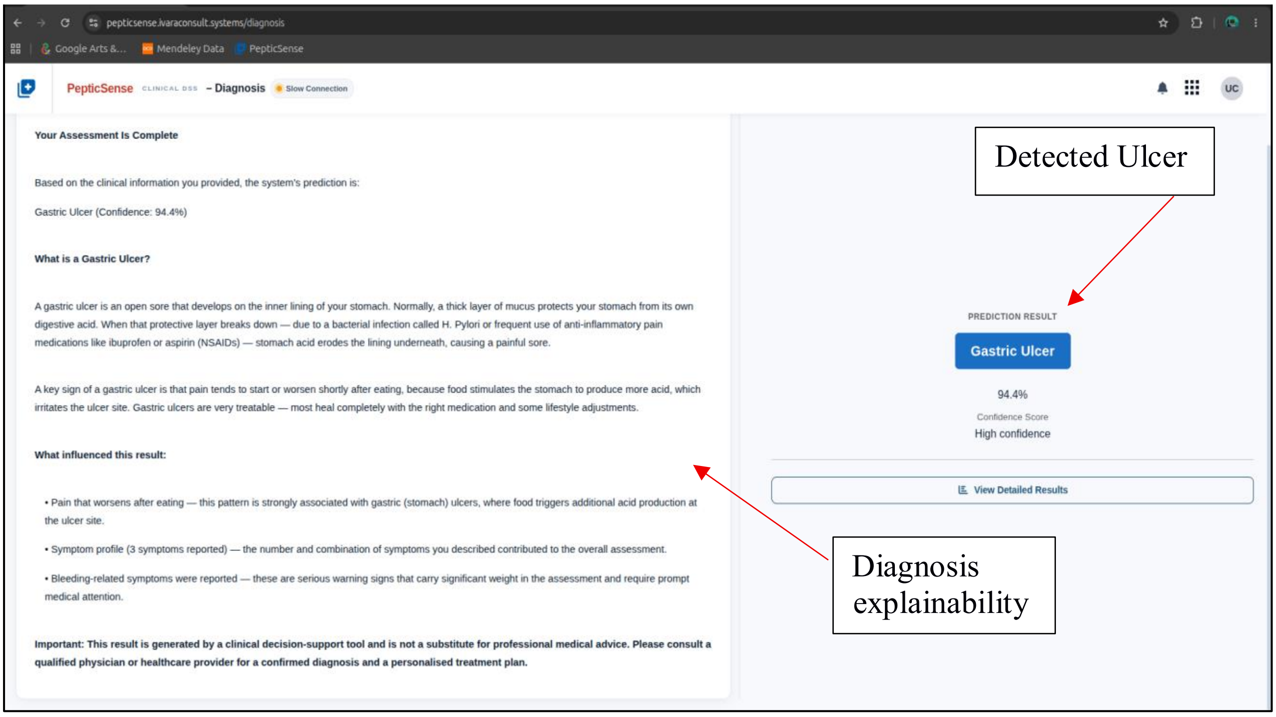


Figure 6: General User Diagnosis Result user interface - Desktop View

❖ **Role-Based Explanations**

1. General User View: Gives a general explanation about the disease to the patient, with some lifestyle suggestions and a clear disclaimer stating that the tool is a screening tool only and should not be used for a definite diagnosis.
2. Physician / Admin View: Provides technical analysis of the patient's condition, with suggestions for further action and relevant weights of the models according to clinical features, as shown in Figure 7.

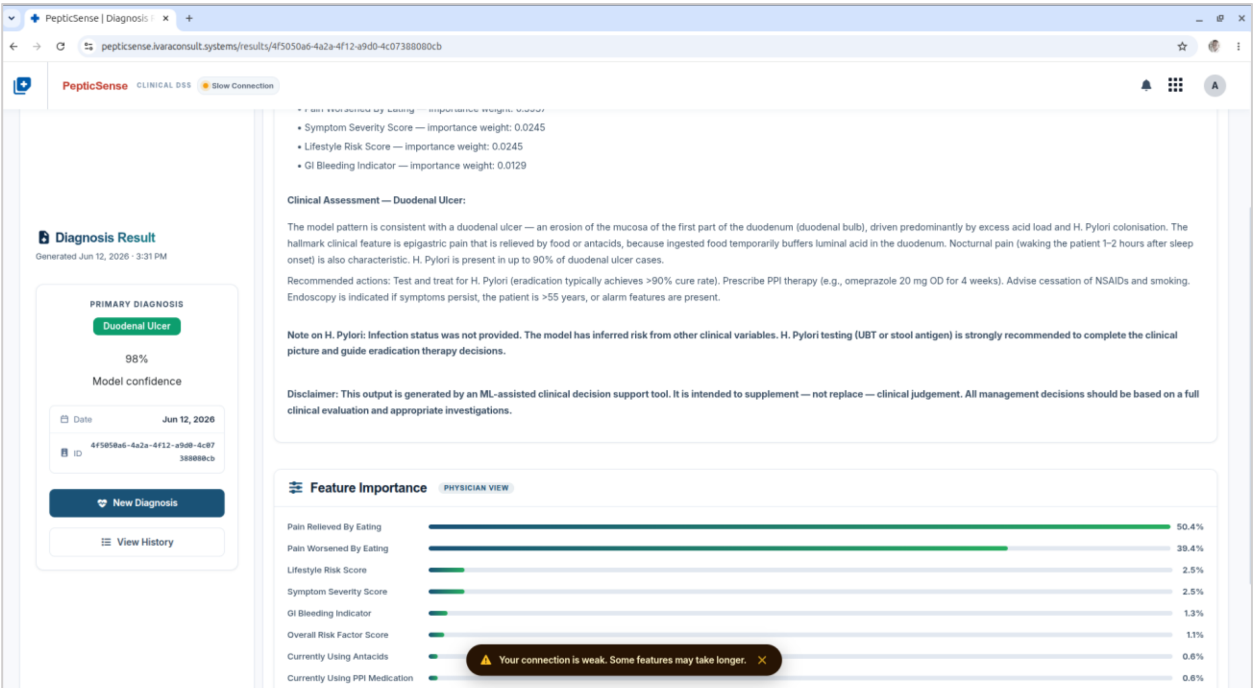


Figure 7: Admin View of Prediction Results - Desktop View

- **Administrative Dashboard**

  Stats View: Shows up-to-date statistics in terms of number of users, total number of diagnoses made, and type of ulcers. The administrative console is illustrated in Figure 8.

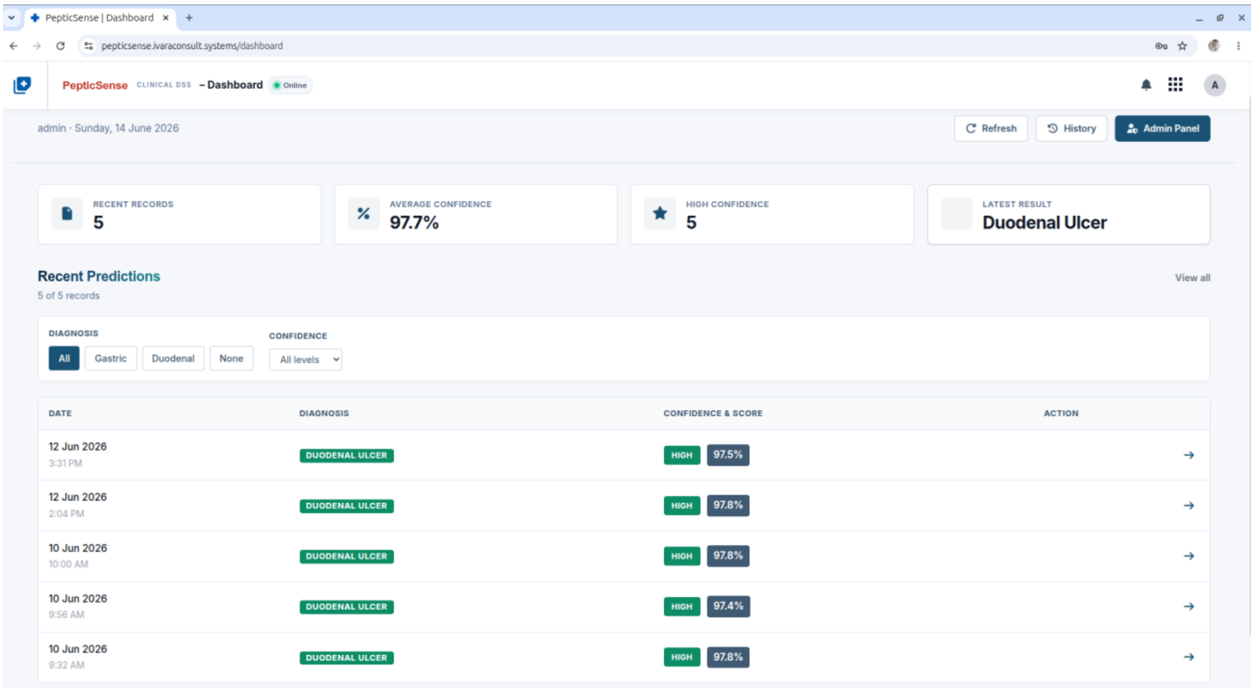


Figure 8. Admin console – Desktop view

### 3.5.3 Database Design

❖ Data Storage Structure (Key Entities)

The schema is segregated into the following functional layers, with the corresponding entity-relationship diagram (ERD) presented in Figure 9:

1. Identity Layer (accounts_useraccount): This layer contains details about user credentials, user roles (GENERAL, PHYSICIAN, ADMIN), and profile linking.
2. Input Layer (diagnosis_diagnosticinput): This layer holds the details about raw input entered during the diagnosis, associated with particular user sessions.
3. Inference Layer (diagnosis_predictionrecord & diagnosis_explainabilitymetrics): This layer contains details about the inference made using the machine learning model, including the primary diagnosis and confidence value. Additionally, there will be a JSON mapping of features with their importance in a particular prediction.
4. Access Control Layer (physician_requests_physicianrequest & reportaccess): This layer controls the sharing functionality between general users and Physician/Clinician, implementing row level security such that Physicians can only view reports where permission is given to them.

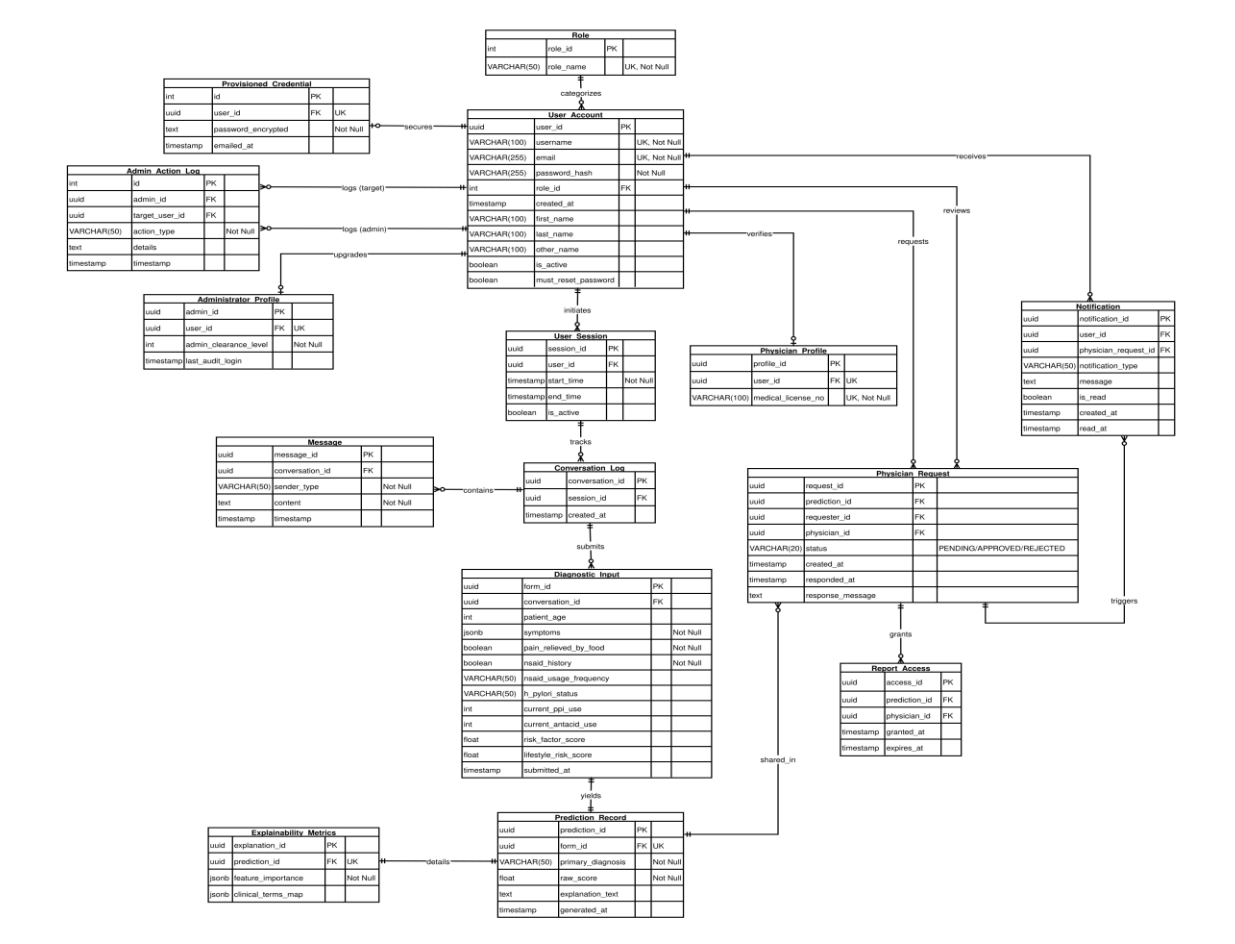


Figure 9. Pepticsense Database Design

❖ Access Method

1. Django ORM: All CRUD operations are performed using the Object-Relational Mapper, which protects the system from SQL injection attacks and standardizes query construction.
2. JWT authorization: Access control is performed using JSON Web Tokens. Every API call must contain the Authorization: Bearer header, which is used by the backend to filter out querysets.
3. Indexing: The primary key in the database is based on UUIDs (prediction_id, request_id).

### 2.5.4 System Architecture Design

**1. Presentation Layer (Client-Side)**

The client-facing part of the system is developed using the Angular framework, which serves as a reliable platform for developing SPAs. The presentation layer acts as the user interface (UI) and allows users including patients, doctors, and administrators to interact with each other. It uses a multi-stage diagnostic wizard to collect clinical symptoms and demographic data. Through client-side validation and management of the application state, the presentation layer helps ensure that clean data reach the server while providing instant feedback to the user.

**2. Application Logic Layer (Server-Side)**

However, the key business logic is found in the backend, which is built using the Django REST Framework (DRF). The client-server architecture is implemented in this layer. This is because the backend acts as a central communication point that uses RESTful APIs to communicate with the frontend. It is responsible for performing crucial functions such as user authentication using Token/JWT security, session management, RBAC, and coordinating review requests by the physicians.

**3. Intelligence Layer (Machine Learning Engine)**

One of the unique characteristics of the system is the inclusion of the Machine Learning Inference Engine. This component is part of the Django system and employs pre-trained classification models (Random Forest & SVM).

**4. Data Persistence Layer (Database)**

In terms of database design for storing data, PostgreSQL database is used, which is one of the strongest relational databases available on the market. The database structure is normalized so that data integrity can be achieved. The database holds important information such as encrypted passwords of users, patient's medical history, input diagnoses and predictions. The database is queried using Django ORM.

**5. Deployment and Infrastructure**

The complete application is set up within a Virtual Private Server (VPS) that uses an operating system based on Linux. In order to keep all data confidential and establish a secure communication channel between the client and the server, the application is protected using SSL/TLS certificates from Let's Encrypt. The complete web-based system architecture is depicted in Figure 10.

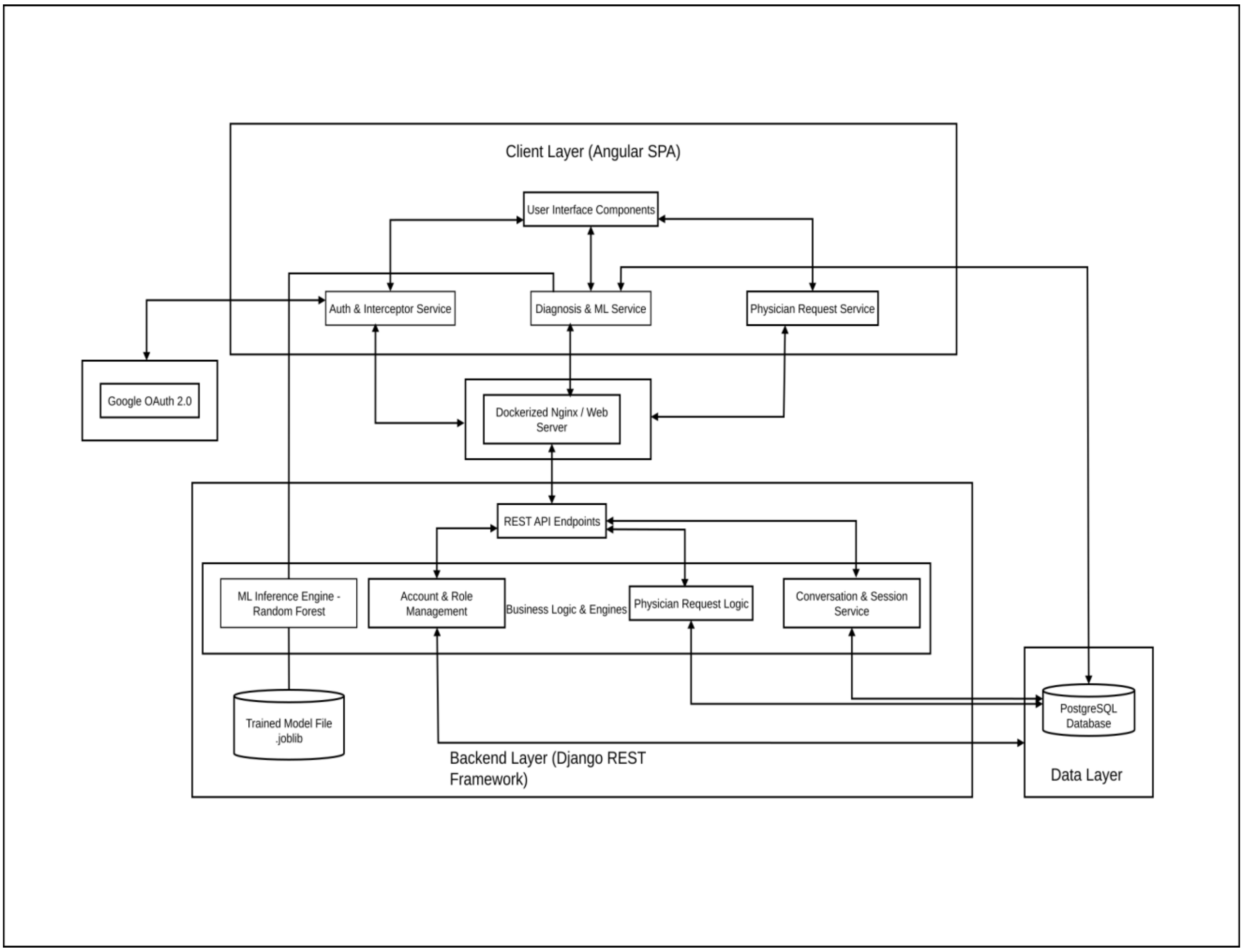


Figure 10: Pepticsense System Architecture

# CHAPTER FOUR
# SYSTEM IMPLEMENTATION

## 4.0 Introduction

This chapter explains the process of implementation of the PUD Prediction System which is referred to internally as "Peptic-Sense". This includes the development environment, the tools used, preprocessing of data for the machine learning model, and the overall architecture of the full-stack system. In addition, the testing procedures used and the results obtained will be explained.

## 4.1 Development Environment and Tools

The architecture was implemented with a decoupled design, thus guaranteeing the clear distinction between data processing functionality and the user interface (UI).

### 4.1.1 Programming Language and Libraries

- ❖ Python (Backend): Used because of rich machine learning and powerful web frameworks.
- ❖ Django & Django REST Framework (DRF): Was used for building scalable and secured API.
- ❖ Scikit-Learn: Was used for implementing the Random Forest classifier and pre-processing pipelines.
- ❖ Pandas, Matplotlib, Seaborn, NumPy: Was used for data manipulation and vectorization of features.
- ❖ Faker (Synthetic data generator): Used for generation and manipulation of data by statistical approaches.
- ❖ Joblib: Was used for serialization and deserialization of models.
- ❖ TypeScript (Frontend): Was used inside Angular framework for building type-safe and reactive single-page application (SPA).
- ❖ PostgreSQL: Chosen as the relational database because of reliability and support of such complicated data types as Universally Unique Identifier (UUID).

### 4.1.2 Development Platform and Hardware Requirements

- ❖ IDEs: Visual Studio Code (VS Code) is used for general development purposes, and Jupyter Notebooks have been used for the first stages of data exploration and modeling.
- ❖ Platform: The solution uses Docker containers along with Docker Compose in Ubuntu-based environments.
- ❖ Hardware Requirements: the development ran on an Intel Core i7 (Quad-core) processor.
- ❖ RAM: 16 GB (in order to run multiple Docker containers and ML training).
- ❖ Storage: 512GB SSD.

### 4.2 Dataset Description and Preprocessing

The dataset used for the training had a size of 6,000*47 but will have a final size of 6,000*42 patients in Nigeria, deliberately created to represent the epidemiology of the H. pylori prevalence (~75%). Dimensions of these columns are shown below:

1. **Demographic (6)**

   Age, Gender, Blood_Group, State_of_Residence, Occupation, Income_Class

2. **Lifestyle (10)**

   Smoker_Status, Alcohol_Frequency, NSAID_Usage_Freq, Spicy_Food_Intake, Fasting_Habits, Stress_Level, Coffee_Consumption, Water_Intake_Liters, Physical_Activity, BMI

3. **Symptoms (15 binary)**

   Epigastric_Pain, Pain_Relieved_By_Eating, Pain_Worsened_By_Eating, Nausea, Vomiting, Melena_Black_Stools, Hematemesis_Vomiting_Blood, Heartburn, Bloating, Belching, Unexplained_Weight_Loss, Loss_of_Appetite, Fatigue, Chest_Pain, Dizziness

4. **Medical History (10)**

   H_Pylori_Infection_History, Family_Ulcer_History, Hypertension, Diabetes, Previous_GI_Bleed, Current_Antacid_Use, Current_PPI_Use, Recent_Typhoid_Treatment, Chronic_Pain_Condition, Asthma

5. **Lab/Clinical (1):** Hemoglobin_Level

6. **Target (1):** Target_Ulcer_Type

- ❖ The preprocessing steps undertaken are as follows:

1. Exploratory data analysis (EDA): analyzed data types, missing value, and distribution of target variable

2. Cleaning data having missing character at random MCAR

3. Feature engineering for 4 features (Symptom_Severity_Score, Risk_Factor_Score, Lifestyle_Risk_Score, and GI_Bleed_Indicator), encoded the categorical variable in binary column and multiclass categorical column Correlation based feature selection

5. Scaling: standard scaling was done for ensuring that features like BMI and Age were not dominating the model due to numerical values. Split ratio: Stratified 80/20 train test split for keeping up the class ratio (Gastric, Duodenal, and None).

- ❖ Split Ratio: A stratified 80/20 train-test split was used to maintain class proportions (Gastric, Duodenal, and None).

### 4.3 Model Architecture and Training

- ❖ The Random Forest classifier is the main algorithm that controls the system and is selected after being benchmarked with Decision Trees, Support Vector Machines, and Naïve Bayes classifier.
- ❖ Algorithm training parameters include the use of 100 estimators, 15 maximum depth to prevent overfitting, and grid search cross-validation with Weighted F1-Score as the metric for hyperparameter optimization.

Figure 11 is a box-plot illustration of the engineered clinical risk variables. From a statistical point of view, the graph gives information about the median, interquartile range (IQR), whisker width, and any outlying observations, giving a statistical representation of the central tendency, spread, and distributional shape of each of the engineered features. Narrow IQRs and the presence of a small number of extreme values imply that most of the observations are centered around some relatively stable intervals and that extreme values correspond to some high-risk clinical cases but not to the instability of the sample as such. From the perspective of preprocessing, the chart was helpful in the identification of skewed distributions, different spreads of the distributions, and outlier observations which could affect the sensitivity of the learning algorithm when applied to the model training. Considering the high robustness of the Random Forest to slight deviations from normality and the presence of outliers, the presented distributions make it possible to use the engineered variables without distorting the learning process.

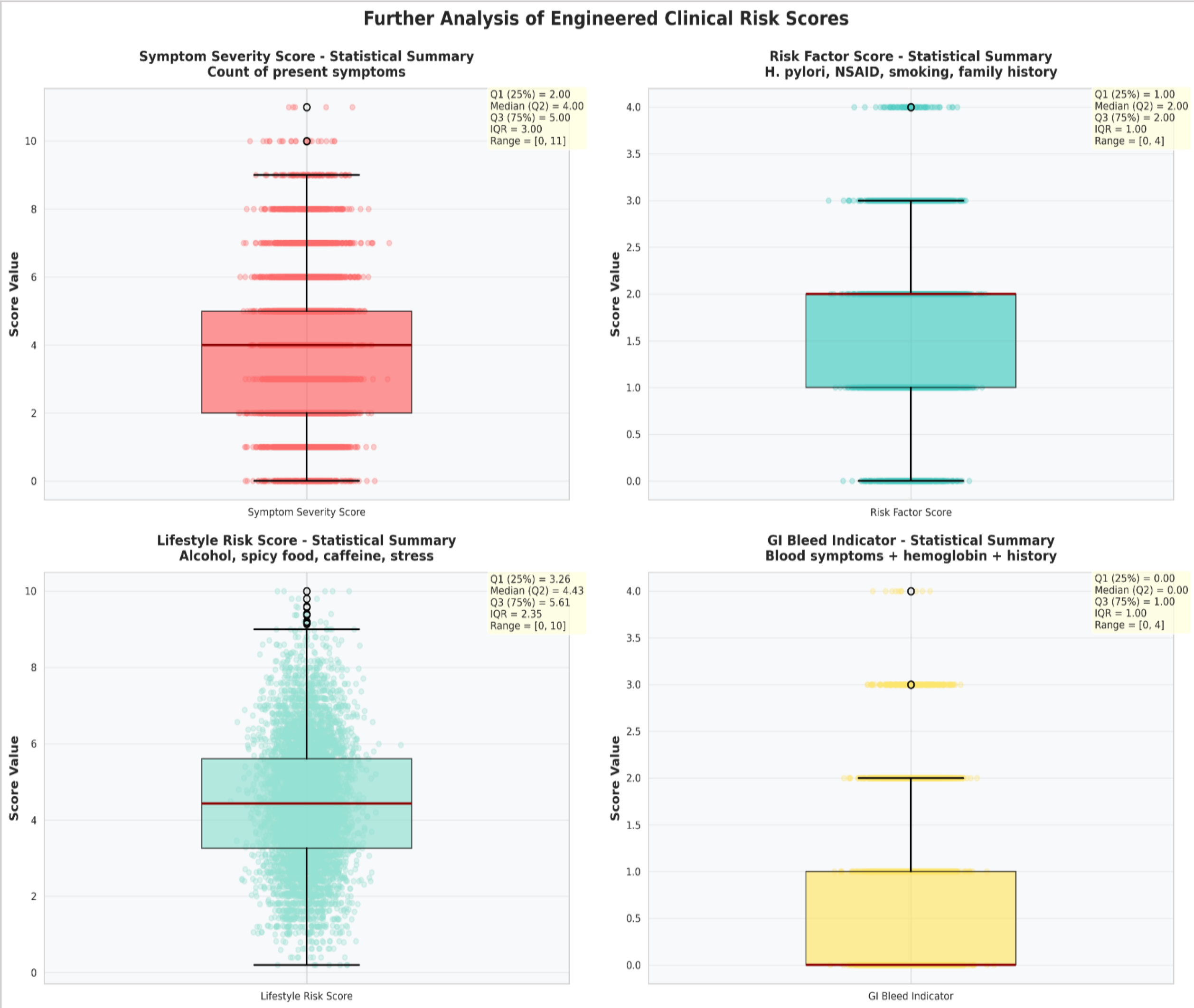

Figure 11: Box plot of Engineered Clinical Risk

Figure 12 demonstrates the correlation matrix of the engineered clinical risk variables. The purpose of the correlation matrix is to display the level of relationships between the variables. In each cell there is the correlation coefficient varying from -1 to +1. When the coefficient is close to +1, the increase in both variables is simultaneous; when it is close to -1, the inverse relationship exists; when the coefficient is close to 0, it implies the lack of any linear relationship. Correlation values greater than 0.70 mean that there is a strong relationship, values in the range of 0.30-0.69 denote the moderate relationship, while values lower than 0.30 indicate the weak relationship. The correlation matrix in this study was used to investigate the presence of multicollinearity, which means that two or more variables have a high level of relation and carry redundant information. The analysis shown in Figure 12 shows that the engineered variables have some kind of relation but still are not so dependent on each other as to consider them duplicates. Therefore, each of them makes its own contribution to the model. Even though the Random Forest classifier is more resistant to multicollinearity compared to models like logistic regression, it is necessary to check the correlation matrix in order to verify the features' meaningfulness and avoid redundancy. Hence, Figure 12 confirms the effectiveness of feature engineering by showing that the variables have some relationship but still have their uniqueness.

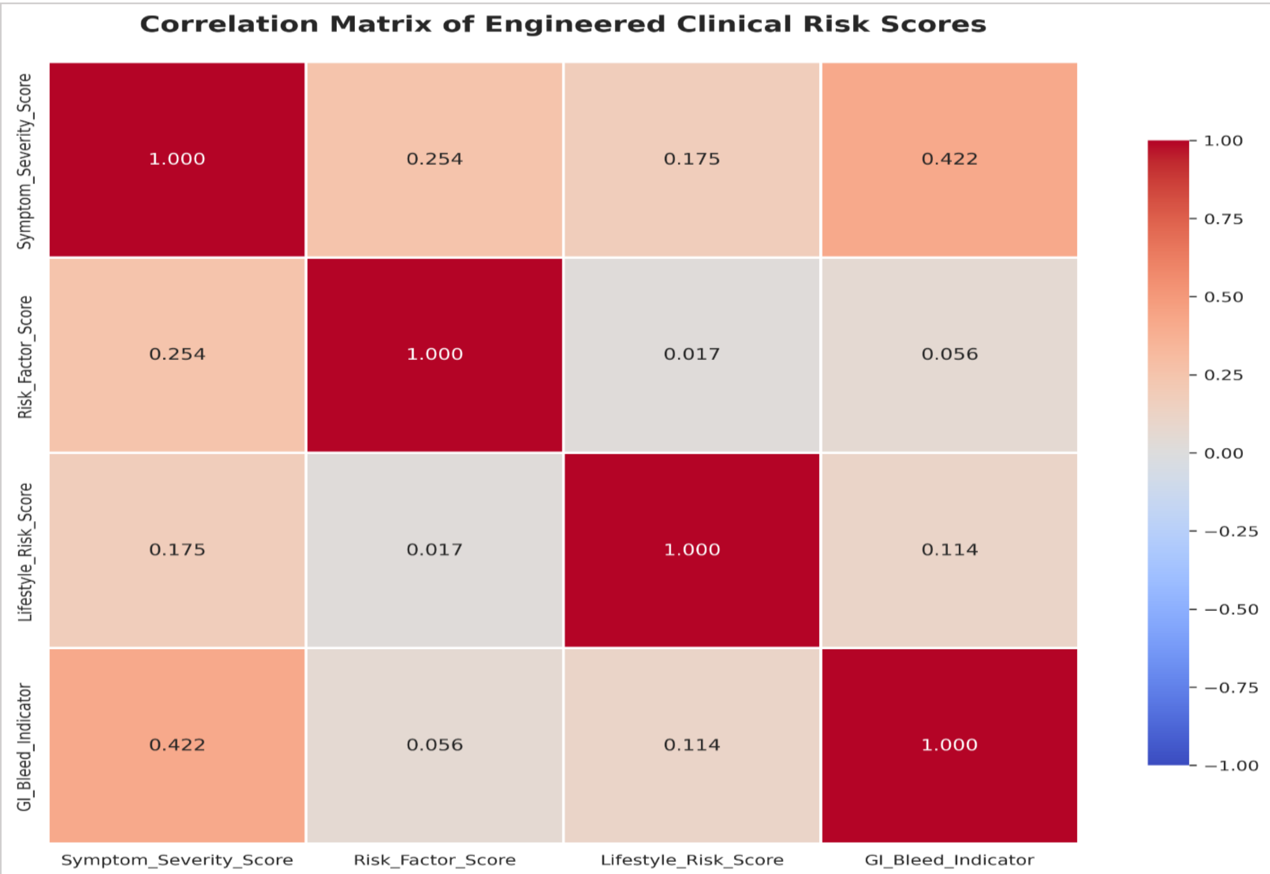


Figure 12: Correlation matrix of the Engineered Clinical Risk Score

The purpose of this figure is to show how often the distribution of the patient is low-risk, medium-risk, and high-risk according to each variable. In each chart, the horizontal axis refers to the value of the score (from low-risk on the left side to high-risk on the right side). The vertical axis represents the number of patients who achieved that score. The higher area on the chart shows that there are more patients with that score, and the lower area is associated with fewer patients. From a statistical point of view, the graph allows immediately assessing three parameters: the central tendency (how many patients), the dispersion (how the score is distributed) and the skewness (asymmetry of the distribution).

All the shown distributions are right-skewed that means that most of the patients have low values of each score, while some other patients have high values (more severe) that cause the long tail of the distribution on the right side. In practical sense, the majority of the patients from the dataset have the mild or medium clinical risk, while the small number of the patients have severe clinical risks, such as GI bleeding, hematemesis or the combination of NSAID usage with the presence of H. pylori infection. It matches the results, which are mentioned in the recent literature about the clinical research of peptic ulcer disease when there is the small number of patients with the severe disease but it is really important for the timely identification.

It can be also noted that the figure shows the distinct separation between the three groups of patients – None, Gastric and Duodenal. The "None" patients' scores are mostly distributed on the lower part of the Symptom_Severity_Score and Risk_Factor_Score that means the absence of symptoms. At the same time, the "Gastric" and "Duodenal" patients' scores are distributed on the higher values of the scores. Particularly, the "Duodenal" patients appear more often in the areas where the pain was reduced by eating and the NSAID risk was high.

In terms of modeling, such distribution works well for the Random Forest classifier. Variables are neither constant nor completely random but have some level of risk distribution that can be divided by the decision trees of Random Forest to obtain reliable results. The existence of skewness and relatively few outlier observations confirms two previous modeling assumptions: the use of Standard Scaling (Z-score normalization) that maintains similar numeric intervals for all variables and using tree-based models, which are known to be tolerant to such type of distribution and outliers. In other words, Figure 13 provides statistical proof that created features are good and clinically relevant enough to be used in the Random Forest model.

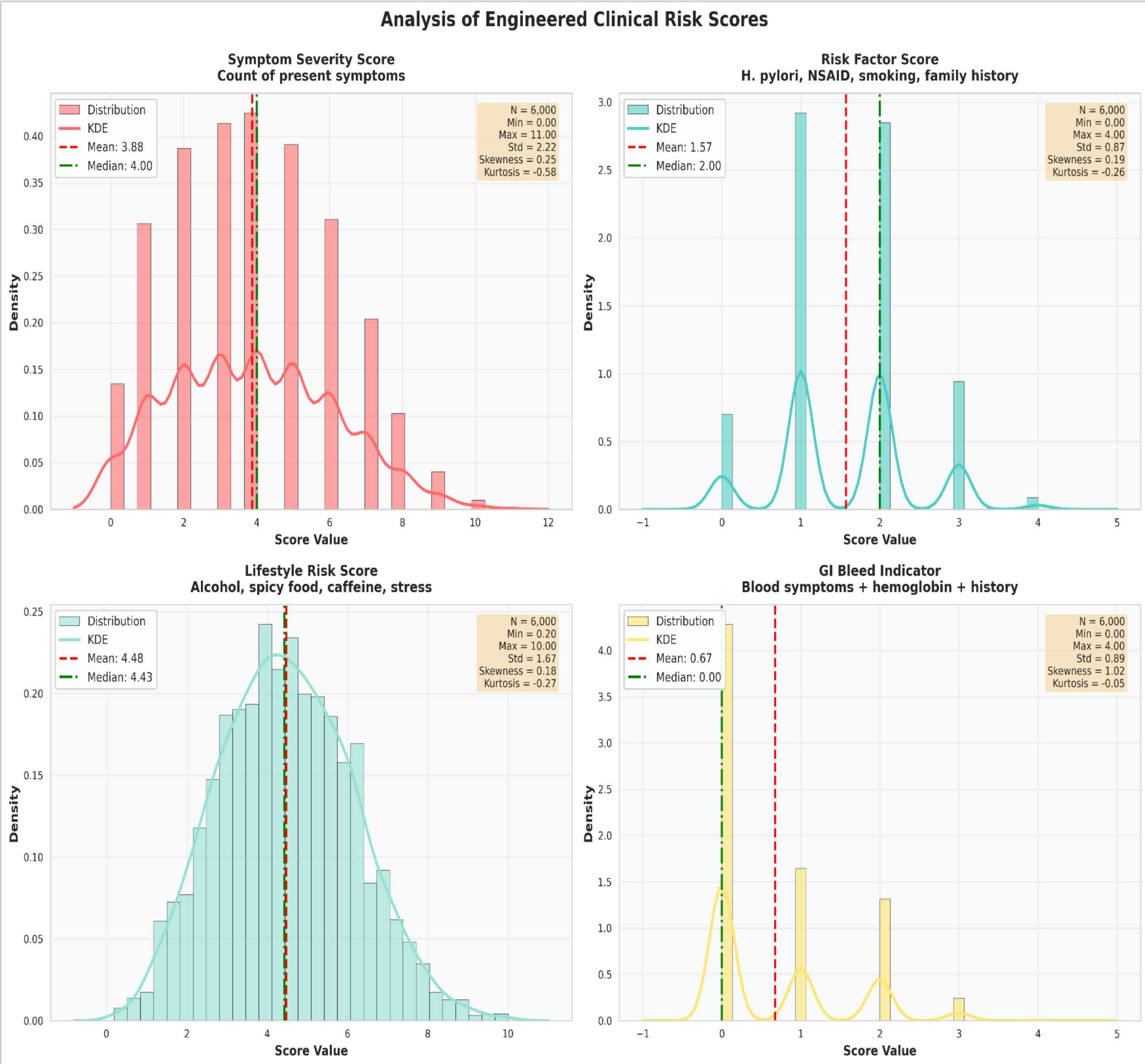


Figure 13: Engineered Clinical Risk Distribution

Figure 14 presents the distribution of the target variable across the three ulcer classes, Gastric, Duodenal, and None. illustrating the class balance that was maintained during the stratified train-test split.

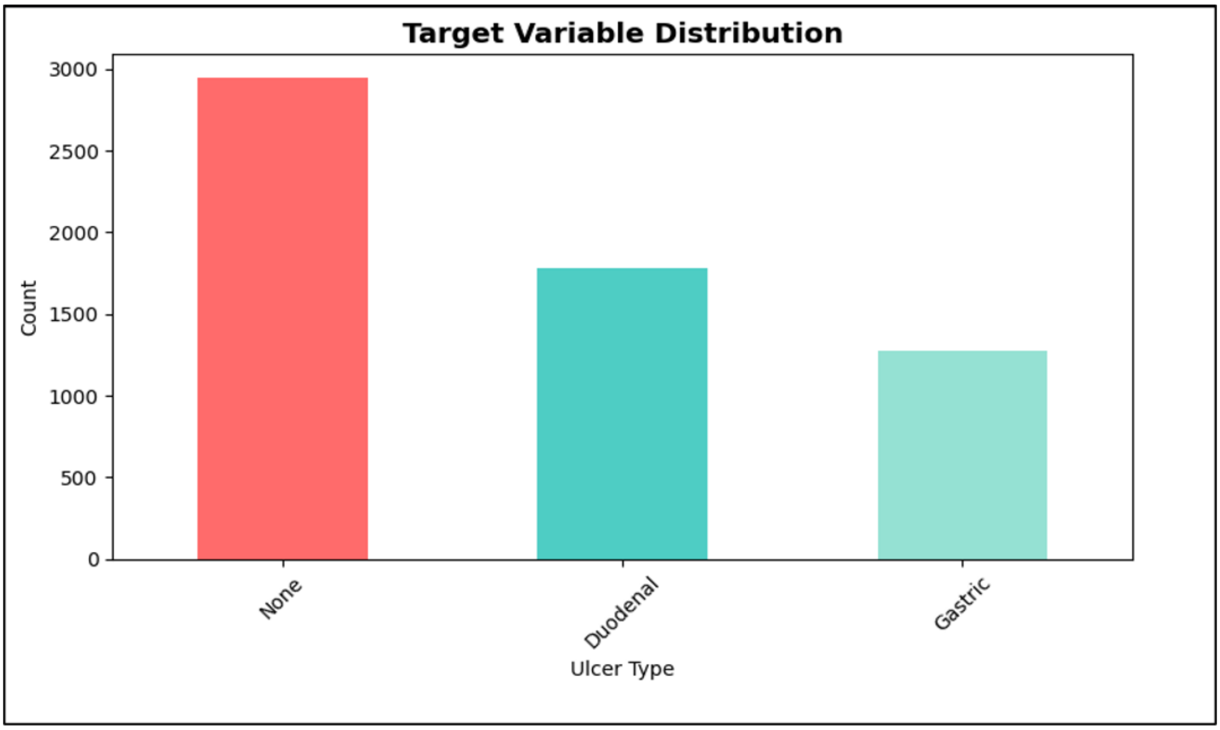


Figure 14: Target variable distribution of ulcer types

**4.4 System Implementation and Modules**

The architecture of the system includes the following interlinked modules:

- ❖ Authentication module: This module is responsible for security through JWT-based techniques and Google Sign-In integration.
- ❖ Diagnostic Wizard (frontend): This five-steps wizard form gathers patients' information and symptoms along with feedback in real-time.
- ❖ ML inference engine (backend): This module preprocesses raw JSON inputs into feature vectors and gives out the prediction, the confidence, and the clinical explanation.
- ❖ Physicians request and notification module: This module allows general users to send some selected diagnostic reports to the physicians, including real-time notification module to alert physicians about their pending requests.
- ❖ Admin Dashboard: Provides tools for user management and audit logs

**4.5 Testing and Evaluation**

**4.5.1 Evaluation Metrics**

- ❖ Accuracy, precision, recall, F1-score, the confusion matrix, and AUC-ROC (Area Under the Receiver Operating Characteristic Curve) can be used to evaluate the system.

Figure 15 shows the report card for the model after its evaluation on 1,200 unique patient records, which were never seen before by the model. There are only three possible values: GASTRIC (gastric ulcer), DUODENAL (duodenal ulcer), and NONE (no ulcer) produced by the model for every patient.

- **Precision** – when the model says "ulcer", how often is it right?
- **Recall** – out of the patients who really have an ulcer, how many did the model catch?
- **F1-score** – one combined score that is high only when precision *and* recall are both high.
- **Support** – how many real patients in the test set belong to that class.

The bottom of the report also shows three overall scores: **accuracy** (total correct predictions out of 1,200), **macro average** (the plain average across the three classes), and **weighted average** (the average adjusted for how many patients fall into each class).

Figure 15 shows that the proposed model possesses sufficient levels of accuracy, balance, and reliability that may be used for screening of stomach and duodenal ulcers.

**Overall Model Performance Metrics**

| Metric | Score | Percentage |
|---|---|---|
| Accuracy | 0.8358 | 83.58% |
| Precision (Weighted) | 0.8346 | 83.46% |
| Recall (Weighted) | 0.8358 | 83.58% |
| F1-Score (Weighted) | 0.8350 | 83.50% |
| AUC-ROC | 0.9513 | 95.13% |

**Per-Class Performance Metrics**

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Duodenal | 0.7830 | 0.7479 | 0.7650 |
| Gastric | 0.7812 | 0.7874 | 0.7843 |
| None | 0.8889 | 0.9100 | 0.8993 |

Figure 15: Classification Report

Figure 16 shows the confusion matrix of the Random Forest classifier, assessed on 1,200 test patients. In the confusion matrix, each row represents the true class (GASTRIC, DUODENAL, or NONE), while each column represents the predicted class by the model. Cells on the main diagonal stand for correct predictions and cells not on the main diagonal correspond to incorrect predictions. Clearly, diagonal values are significantly greater than off-diagonal values, meaning that the model is able to make the right prediction most of the time. Looking at rows of the matrix gives the information about recall and looking at columns provides information about precision. It appears that the model makes mistakes when trying to discriminate between GASTRIC and DUODENAL ulcers that seems to be natural due to clinical similarity of these illnesses (symptoms such as epigastric pain, infection with H. pylori, and NSAIDs). On the other hand, ulcer patients are rarely classified to the NONE class – this is important as missing the illness in the screening procedure means the most dangerous mistake; Figure 16 proves that such mistakes are very rare. The results shown in Figure 16 match the high F1-scores from Figure 15 and other metrics in Figure 17.

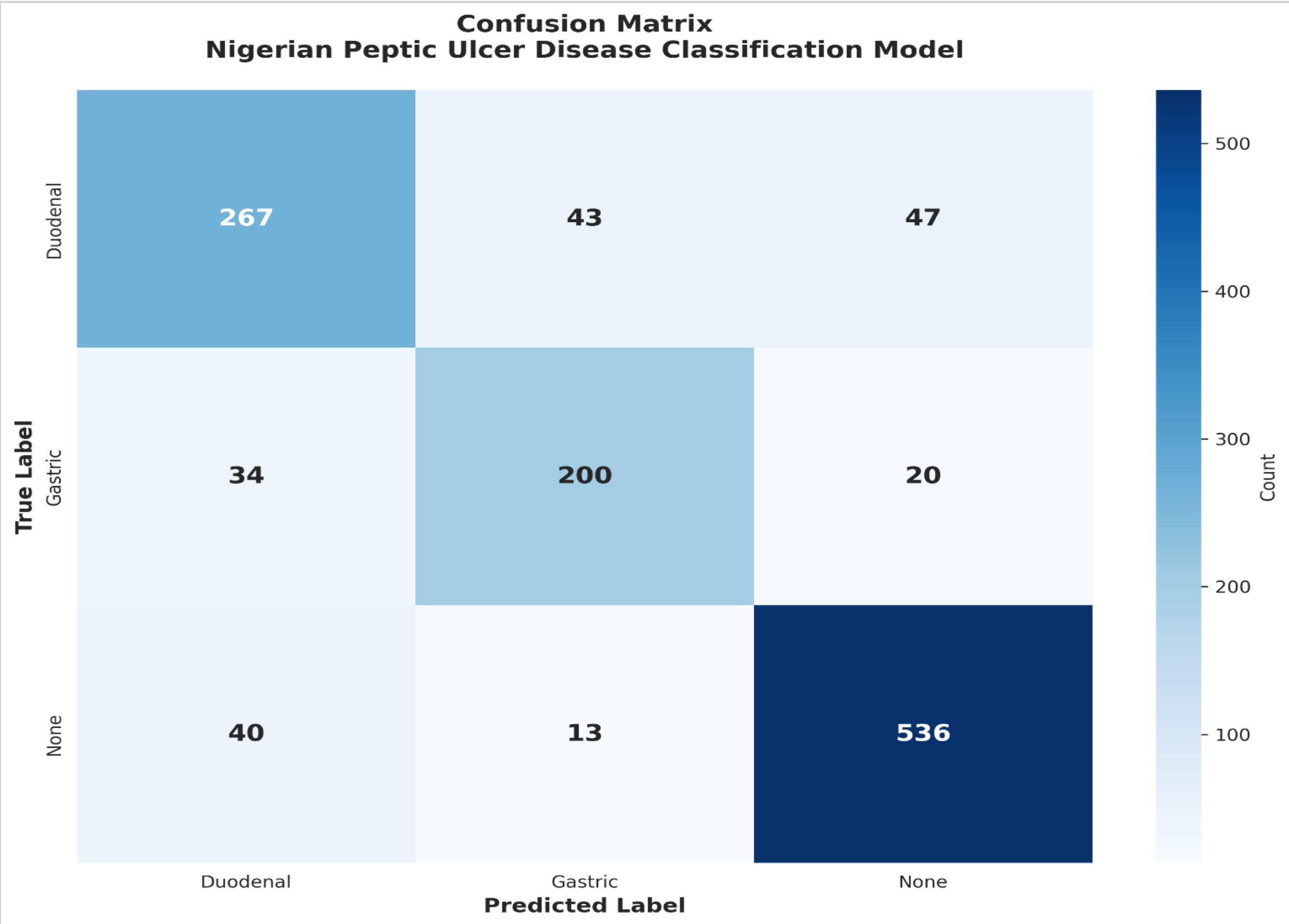


Figure 16: Model Confusion Matrix

Figure 17 shows a bar graph with the four main metrics utilized to evaluate the performance of the Random Forest model on the test data with 1,200 patients' records: accuracy, precision, recall, and F1 score. In the graph, all bars are equally tall and have low variations, which means that the model maintains the high level of correctness and reaches a balance in terms of GASTRIC, DUODENAL, and NONE classes.

Besides accuracy, precision, recall, and F1-score metrics, we use another metric that is widely used to measure a model's performance. It is called AUC-ROC or Area Under the Receiver Operating Characteristic Curve. The ROC curve implies a question "How well is the model capable of identifying actual cases of ulcers without misclassifying healthy patients, as its threshold of classifying the patients as having an ulcer becomes higher or lower?" The y-axis stands for the true positive rate (the percentage of correctly classified ulcers), and x-axis stands for the false positive rate (the percentage of patients who do not have ulcers but were falsely classified by the model as ulcer-positive). AUC stands for the area under the ROC curve and can be measured by one number from 0 to 1. AUC-ROC was calculated through the one-vs-rest method. against the other two. The macro-average of AUC-ROC, as well as class AUC-ROC, were in the category of very strong values.

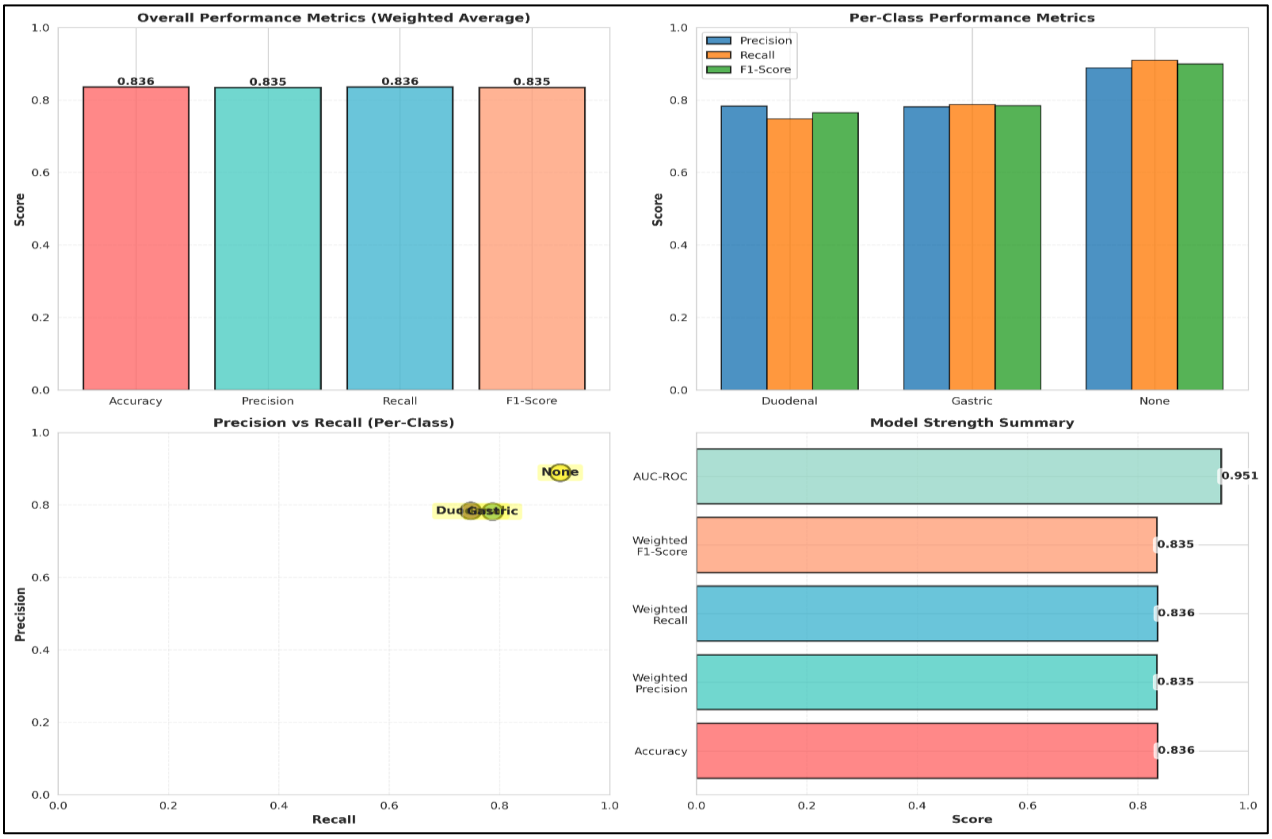


Figure 17: Metrics for Model Chart

### 4.5.2 System Testing

- ❖ Unit Testing: Each backend service and frontend component was tested independently.
- ❖ Integration Testing: The entire process from the form submission to making an inference with the model and storing results in a database was checked.
- ❖ API Testing: Postman was used to check whether DRF API returned the right status codes and JSON structure for different user roles.

### 4.5.3 User Interface Testing

UI components were tested according to 8 Golden Rules of Interface Design by Ben Shneiderman, focusing on consistency and informative feedback.

Diagnosis Wizard: Verified whether the progress bar worked properly and "Alarm Symptoms" gave a prompt visual response. The diagnosis process within a physician account is shown in Figure 18.

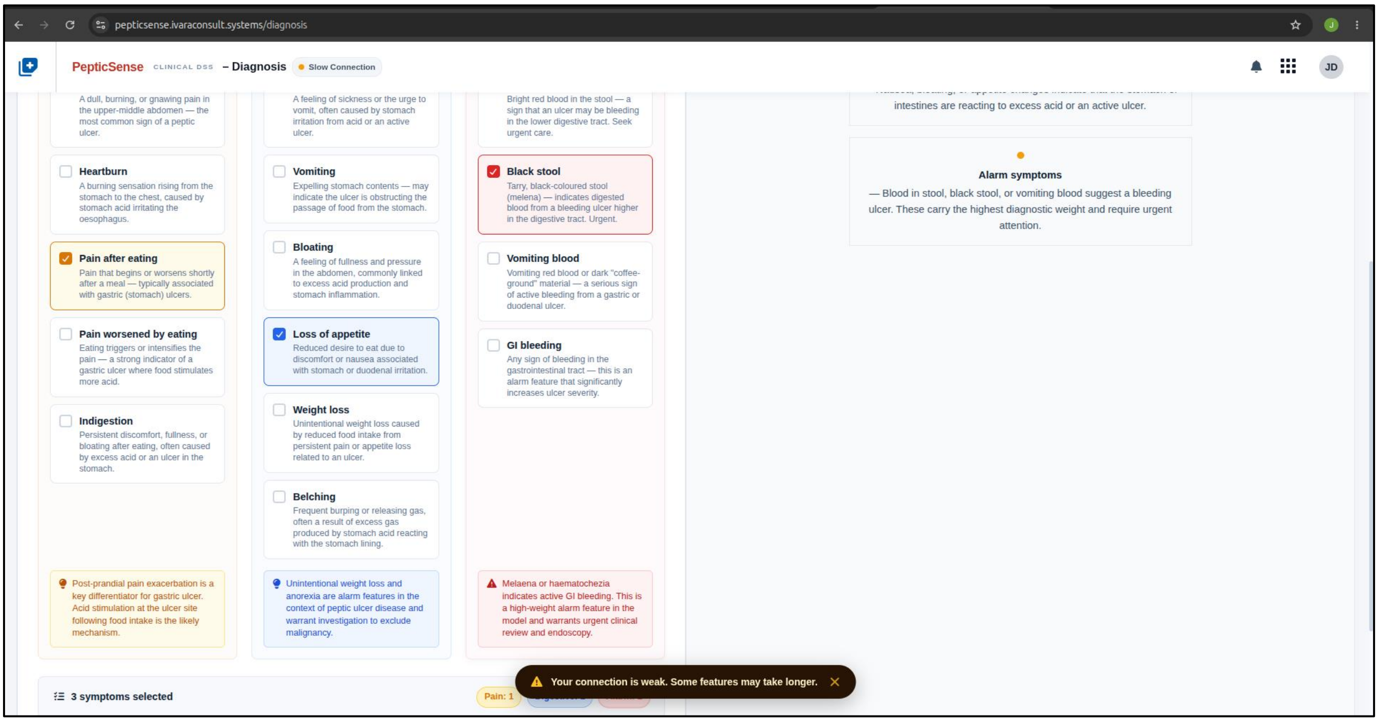


Figure 18: Diagnosis Process in Physician Account - Desktop View

- ❖ Responsive Design: UI components were tested in mobile, tablet and desktop viewports based on the concept of adaptive design of Angular. Tests were done at https://pepticsense.ivaraconsult.systems/

## 4.6 Results Presentation and Analysis

### 4.6.1 Sample Output and Interface Screens

Figure 19 shows a sample prediction result as displayed within a verified physician's account on a desktop view.

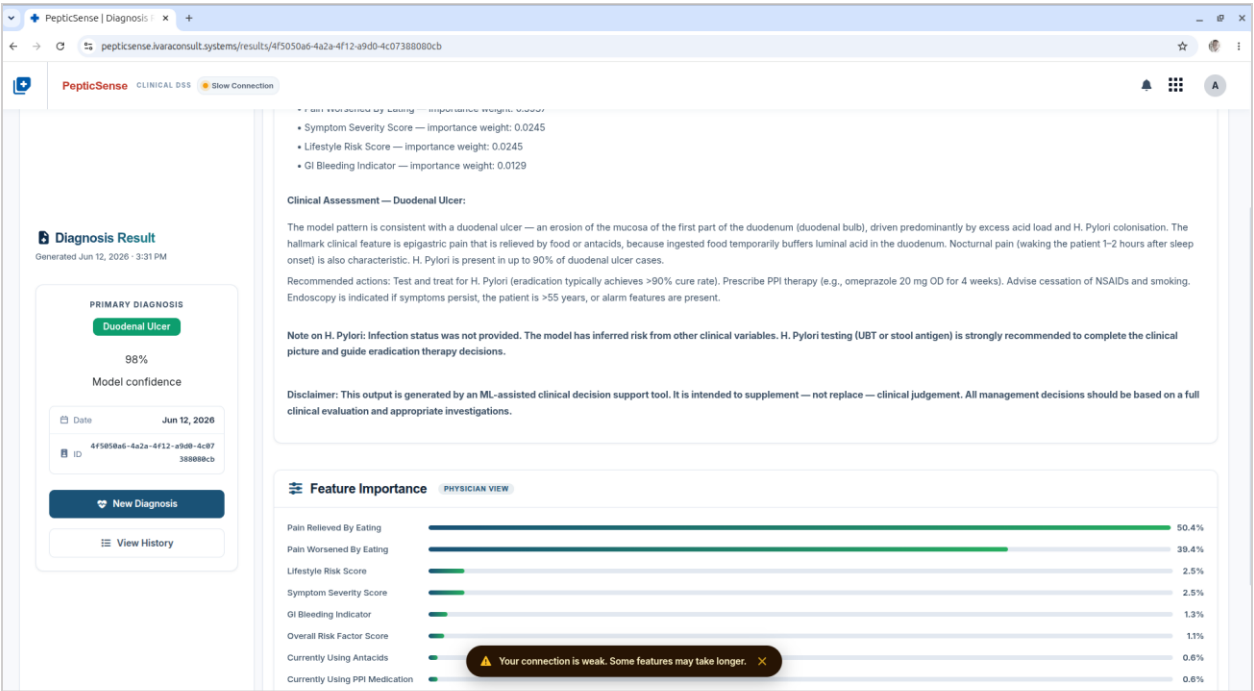


Figure 19. prediction result from Physician account – Desktop view

### 4.6.2 Discussion of Results

- ❖ Performance: The Random Forest model had very high generalization, with a very small difference between training and test accuracies, suggesting that there was no problem of over-fitting.
- ❖ Strengths: The model correctly identifies and gives importance to “Alarm Features,” making it a useful triaging system where endoscopic screening cannot be performed in rural regions.
- ❖ Weaknesses: The model uses patient-provided information, which might be biased. The model is intended to assist, not to substitute clinical decision-making.
- ❖ Goals Achieved: The development of a non-invasive screening system for Peptic Ulcer Disease has been achieved.

# CHAPTER FIVE

# SUMMARY, CONCLUSION, AND RECOMMENDATIONS

## 5.1 Introduction

Introduction of the chapter and restating the aim of the work. It briefly highlights the subject matter, which is a synthesis of the research, conclusions from the results obtained and recommendations for future works or for implementation of the system.

## 5.2 Summary of the Study

The main aim of the research was to provide the solution to the problems of current PUD diagnoses, which were found to be time-consuming, expensive and vulnerable to errors made by the medical personnel, while having the biggest effect on low-income people. In particular, the study aimed at creating and providing an opportunity to use the machine learning-based tool that would allow for fast and affordable classification between the types of ulcers in order to fill the gap found in literature review when there are either general solutions or those based on the data unavailable to the researchers due to privacy regulations. Since it was impossible to use any clinical data due to confidentiality regulations, a synthetic dataset reflecting the pathology of PUD and its high presence of Helicobacter pylori bacteria among the Nigerian population was created.

The proposed solution is called "PepticSense" and it represents a dynamic web-based decision support tool that was built with the help of integrating the Agile development life cycle methodology and the Object-Oriented Analysis and Design (OOAD) one. The combination allowed for the iterative development of features, good collaboration of Django REST Framework (DRF) back-end and Angular front-end sides of the system and care for non-functional requirements like sub-second inference latency, JWT authentication and role-based access. Functionally, the system is designed as a wizard-like multi-step diagnostic test with explainable AI (XAI) outputs suitable for both common users and doctors and a workflow allowing patients to send their diagnostic report to a verified physician for additional consultation.

The implementation of the system (described in Chapter Four) was conducted in accordance with training of a Random Forest classifier on the preprocessed dataset consisting of 6,000 records and 42 features chosen as a result of benchmarking of Decision Tree, Support Vector Machine and Naïve Bayes classifiers. Data preparation involved CRISP-DM-aligned workflow that includes data exploration, data cleaning, engineering of four indicators related to patients' health (symptoms severity, risk factor, lifestyle risk and GI-bleed indicator), categorical variables encoding, correlation-based feature selection and Standard Scaling in order to avoid any variable domination. After that the hyperparameters of the classifier were tuned with the help of GridSearchCV on stratified 80/20 train/test split and the classifier was estimated with accuracy, precision, recall, F1-score, confusion matrices and AUC-ROC. The comprehensive unit, integration, API and user interface tests proved that PepticSense reliably flags high-priority “alarm features,” delivers role-appropriate explanations, and functions effectively as a non-invasive, accessible triage tool within the Nigerian clinical context, thereby fulfilling the objectives set out at the beginning of the study.

### 5.3 Conclusion

Continuing the discussion provided in Section 5.2, the goal of the current research is to build and use a machine-learning-powered tool for the clinical diagnosis of peptic ulcer disease, focusing especially on distinguishing between Gastric and Duodenal ulcers. The general objective formulated in Chapter One, namely "a machine learning approach to clinical diagnosis of peptic ulcer," is completely fulfilled through the design, implementation, and testing of the PepticSense tool.

The following is how each of the four objectives stated at the beginning of the study was reached:

- ❖ A synthetic dataset, designed specifically for the pathology of peptic ulcer, was used together with exploratory data analysis and creating four clinical risk factors to address data scarcity due to patient confidentiality constraints.
- ❖ A dynamic website was built for both patients and clinicians using a decoupled Django REST Framework and Angular infrastructure, including JWT-based authentication and role-based access control.
- ❖ The training of a Random Forest model was conducted and tested in real time, showcasing excellent interoperability among the inference engine, API layer, and the

website, delivering fast, inexpensive classifications of Gastric, Duodenal, and No-Ulcer cases.

- ❖ Peptic ulcer diagnosis and prediction were made completely easy with an interactive and role-aware website and near-zero inference latency, making the tool available across different socio-economic and geographic environments.

Taken together, these results prove that PepticSense works well as a first-line, non-invasive triage tool, especially useful in rural and resource-constrained settings where endoscopy is not readily available. The key weakness of this study, mentioned throughout, is based on the fact that PepticSense uses patient-reported symptoms and a synthetic dataset, introducing subjectivity and restricting external validity. However, the usage of Explainable AI results, warning alarms, and clinical disclaimer somewhat counteracts the above-mentioned constraint, making PepticSense a decision support tool complementing the judgment of a clinician.

### 5.4 Recommendations

On the back of the results obtained from the study as well as the successful implementation of PepticSense, the following recommendations can be made for future research and further development of the tool:

- ❖ Get real clinical data and enable learning continuously: get the necessary approvals and sign the agreements on sharing the data between the clinic at the Godfrey Okoye University and other partner clinics. Make use of the anonymized data available in reality for continuous training of the model, thus making the transition from synthetic data to the actual data-driven production-grade accuracy and reliability possible.
- ❖ Investigate advanced model architectures: compare the existing Random Forest classifier with the deep learning models, such as Recurrent Neural Network (RNN), GRU and Transformer-based models that have shown to achieve better performance in predicting NSAID-induced ulcers with the help of longitudinal electronic health records.
- ❖ Add objective data source: make PepticSense less dependent on subjective data about patients' symptoms through integrating the input from such objective data sources as laboratory information systems, hospital EHRs and APIs from diagnostic devices. This will make it possible to automatically ingest features, such as hemoglobin concentration, H. pylori tests and endoscopies.

- ❖ Scalability and broad adoption: take advantage of the system's cloud agnostic architecture (Docker on DigitalOcean) to extend the reach of PepticSense to more health care facilities, NGOs and tele-health organizations across the country to provide underserved populations in Nigeria with low-cost early screenings.
- ❖ Improve clinical validation and governance: do the prospective clinical validations by practising gastroenterologists, create feedback loop from the Physician's dashboard and establish a process of governance of the model (versioning, monitoring, periodic reevaluation).

## References


[1] A. Divya, B. D. (2022). Disease Prediction Based on Symptoms Given by User Using Machine Learning. *A. Divya, B. Deepika, C. H. Durga Akhila, A. Tonika Devi, B. Lavanya, E. Sravya Teja*.

[2] Ahsan, M. A. (2021). Detecting SARS-CoV-2 From Chest X-ray Using Artificial Intelligence. *IEEE Access*.

[3] Ahsan, M. L. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare (published by MDPI)*.

[4] Ahsan, M. M. (2020). Deep MLP-CNN Model Using Mixed-Data to Distinguish between COVID-19 and Non-COVID-19 Patients. *Deep MLP-CNN Model Using Mixed-Data to Distinguish between COVID-19 and Non-COVID-19 Patients*.

[5] Brownlee, J. (2016). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch.* Melbourne: Machine Learning Mastery.

[6] Brownlee, J. (2018). *Statistical Methods for Machine Learning.* Machine Learning Mastery.

[7] Drucker, H. W. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*.

[8] Houssein, E. H. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*.

[9] Huyen, C. (2022). *Designing Machine Learning Systems.* California: O'Reilly Media, Inc.

[10] Joo Seong Kim, J. K.-H. (2024). Deep Learning-based Prediction of Peptic Ulcer Diseases Caused by Nonsteroidal Anti-inflammatory Drugs Using Longitudinal Electronic Health Records. *Preprint*.

[11] K. Rajendran, M. J. (2020). Predicting Breast Cancer via Supervised Machine Learning Methods on Class Imbalanced Data. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 54-63.

[12] Krishna Kant Singh, M. E. (2021). *Machine Learning and.* Mara Conner.

[13] Krishna Kant Singh, M. E. (2021). *Machine Learning and the Internet of Medical Things in Healthcare.* United Kingdom: Academic Press (an imprint of Elsevier).

[14] McPhee, S. J., & Papadakis, M. A. (2010). *Current Medical Diagnosis and Treatment 2010.* New York: McGraw-Hill Medica.

[15] Mitchell, T. M. (1997). *Machine Learning.* New York: McGraw-Hill.

[16] National Center for Infectious Diseases (U.S.). Division of Bacterial and Mycotic Diseases. Health Communications Activity. (1998/07/01). *Helicobacter pylori: Fact Sheet for Health Care Providers: Upodated July 1998.* Atlanta, Georgia, USA: Centers for Disease Control and Prevention (CDC), National Center for Infectious Diseases.

[17] Nopour, R. (2025). Development of a prediction model for peptic ulcer disease using the fusion of a feature selection stategy and ensemble algorithms. *BMC Gastroenterology*.

[18] Saeedreza Arab, K. R. (2021). A novel fuzzy expert system design to assist with peptic ulcer disease diagnosis. *cognet engineering*.

[19] Sanjay Basu, J. H. (2020). Title: Machine Learning Methods for Precision Medicine Research Designed to Reduce Health Disparities: A Structured Tutorial. *Ethnicity & Disease Publication Date: April 2, 2020*, 217–228.

[20] Shivahare, B. D., Singh, J., Ravi, V., Chandan, R. R., Alahmadi, T. J., Singh, P., & Diwaka, M. (2024). Delving into Machine Learning's Influence on Disease Diagnosis and Prediction. *The Open Public Health Journal*.

[21] Song Y, Z. S. (2021). Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *IEEE/ACM Transactions on Computational Biology and Bioformatics*, 2775 - 80.

## Appendix

### A. Dataset & Preparation Specifications

- Study Focus: Peptic Ulcer Disease (PUD) classification (Gastric vs. Duodenal vs. Normal/No Ulcer) in the Nigerian clinical context.
- Dataset Size: 6,000 patient records.
- Feature Count: 29 raw features mapped down to 14 core clinical input variables.
- Target Classes: 3 classes:
    - 1. GASTRIC (Gastric/Stomach Ulcer)
    - 2. DUODENAL (Duodenal Ulcer)
    - 3. NONE (No significant ulcer pattern / Normal)
- Train/Test Partition: Stratified 80/20 train-test split:
- Training Set: 4,800 records.
- Test Set: 1,200 records.

### B. Model Training Code

The model was trained, tuned, and evaluated using Python. Below is the core training script containing model definitions, hyperparameter tuning using Grid Search with Stratified K-Fold cross-validation, training, and final model serialization:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import StratifiedKFold, GridSearchCV, train_test_split
```

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
import joblib

# 1. Load the preprocessed clinical dataset
df_ml = pd.read_csv('engineered_ulcer_data.csv', na_values=[''], keep_default_na=False)
df_ml = df_ml.dropna()

# Separate features and target
X = df_ml.drop(columns=['Target_Ulcer_Type'])
y = df_ml['Target_Ulcer_Type']

# 2. Stratified Train-Test Split (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

# 3. Model Definition and Initial Fit Comparison
models = {
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42, n_jobs=-1),
    'Decision Tree': DecisionTreeClassifier(max_depth=15, random_state=42),
    'SVM': SVC(kernel='rbf', probability=True, random_state=42),
    'Naive Bayes': GaussianNB()
}

# 4. Hyperparameter Tuning for Production Random Forest Model
rf_params = {
    'n_estimators': [100, 200],
    'max_depth': [10, 15, 20],
    'min_samples_split': [2, 5]
}

rf_grid = GridSearchCV(
    RandomForestClassifier(random_state=42, n_jobs=-1),
    rf_params,
    cv=3,
```

```
    scoring='f1_weighted',
    n_jobs=-1
)
rf_grid.fit(X_train, y_train)

# Save the best estimator as the final production model
best_rf_model = rf_grid.best_estimator_
joblib.dump(best_rf_model, 'nigerian_ulcer_model.joblib')
```

## C. Database Schema Creation Code

The Django models defined in the database schema map directly to the database tables:

```
import uuid
from django.db import models
from conversations.models import ConversationLog



class DiagnosticInput(models.Model):
    """Stores the patient diagnostic form data submitted for prediction."""
    form_id = models.UUIDField(primary_key=True, default=uuid.uuid4, editable=False)
    conversation = models.ForeignKey(
        ConversationLog,
        on_delete=models.CASCADE,
        related_name='diagnostic_inputs',
    )
    patient_age = models.IntegerField()
    symptoms = models.JSONField()
    pain_relieved_by_food = models.BooleanField()
    nsaid_history = models.BooleanField()
    h_pylori_status = models.CharField(max_length=50, blank=True, default='')
    submitted_at = models.DateTimeField(auto_now_add=True)

    def __str__(self):
        return f"DiagnosticInput {self.form_id}"
```

```python
class PredictionRecord(models.Model):
    """Stores the ML prediction result for a diagnostic input."""
    GASTRIC = 'GASTRIC'
    DUODENAL = 'DUODENAL'
    NONE = 'NONE'
    DIAGNOSIS_CHOICES = [
        (GASTRIC, 'Gastric Ulcer'),
        (DUODENAL, 'Duodenal Ulcer'),
        (NONE, 'No Ulcer Detected'),
    ]

    prediction_id = models.UUIDField(primary_key=True, default=uuid.uuid4, editable=False)
    form = models.OneToOneField(
        DiagnosticInput,
        on_delete=models.CASCADE,
        related_name='prediction',
    )
    primary_diagnosis = models.CharField(max_length=50, choices=DIAGNOSIS_CHOICES)
    raw_score = models.FloatField()
    explanation_text = models.TextField(blank=True, default='')
    generated_at = models.DateTimeField(auto_now_add=True)

    def __str__(self):
        return f"{self.primary_diagnosis} ({self.raw_score:.2f})"


class ExplainabilityMetrics(models.Model):
    """Stores feature importance and clinical term mappings for a prediction."""
    explanation_id = models.UUIDField(primary_key=True, default=uuid.uuid4, editable=False)
    prediction = models.OneToOneField(
        PredictionRecord,
        on_delete=models.CASCADE,
        related_name='explainability',
    )
    feature_importance = models.JSONField()
    clinical_terms_map = models.JSONField()
```

```
class Meta:
    verbose_name_plural = 'Explainability metrics'


def __str__(self):
    return f"Explainability for {self.prediction.prediction_id}"
```

## D. Hyperparameter Grid & Final Model Configuration

Table D.1 lists the full hyperparameter search space, while Table D.2 records the final winning configuration selected for the production model.

*Table D.1 — Hyperparameter search space*

| Hyperparameter | Values searched | Purpose |
|---|---|---|
| n_estimators | 100, 200 | Number of trees in the forest; trades off accuracy against training cost. |
| max_depth | 10, 15, 20 | Maximum depth of each tree; controls model complexity and overfitting. |
| min_samples_split | 2, 5 | Minimum samples required to split an internal node; smooths decision boundaries. |
| scoring | f1_weighted | Optimization metric, chosen to handle the 3-class (Gastric / Duodenal / None) target. |
| cv | 3-fold Stratified K-Fold | Preserves class proportions in every fold during cross-validation. |

*Table D.2 — Final winning Random Forest configuration*

| **Parameter** | **Selected value** |
|---|---|
| n_estimators | 100 |
| max_depth | 15 |
| min_samples_split | 2 |
| criterion | gini (default) |
| class_weight | None (classes already balanced via stratified split) |
| random_state | 42 |
| n_jobs | -1 (parallelized across all available CPU cores) |
| **Serialized artefact** | nigerian_ulcer_model.joblb |