# DEVELOPMENT OF PROSTATE ENLARGEMENT RISK PREDICTION AND MANAGEMENT STRATEGIES AMONG ADULT MEN

**BY**

**OKEY GODWIN, IFEANYI**
**GOU/U22/CSC/839**

**DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS,**
**FACULTY OF NATURAL SCIENCES AND ENVIRONMENTAL STUDIES, GODFREY OKOYE UNIVERSITY, ENUGU STATE.**

**IN PARTIAL FULFILLMENT OF THE AWARD OF BACHELOR OF SCIENCE (B.Sc), DEGREE IN COMPUTER SCIENCE.**

**SUPERVISOR: MR UGWU NNAEMEKA**

**JULY, 2026.**

## APPROVAL

This research, titled **Development of Prostate Enlargement Risk Prediction and Management Strategies among Adult Men**, has been assessed and approved by the Department of Computer Science and Mathematics Committees in the Faculty of Natural Sciences and Environmental Studies, Godfrey Okoye University, Enugu, Nigeria.

| | | |
|---|---|---|
| **Mr Ugwu Nnaemeka**<br>**Supervisor** | ..................................<br>Signature | ..................................<br>Date |
| **External Examiner**<br>External Examiner | ..................................<br>Signature | ..................................<br>Date |
| **Dr. Frank Okebanama**<br>HOD, Department of<br>Computer Science and<br>Mathematics | ..................................<br>Signature | ..................................<br>Date |
| **Prof. Ajah I. A.**<br>Dean, Faculty of Computing<br>And Information Technology. | ..................................<br>Signature | ..................................<br>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.

**OKEY GODWIN IFEANYI**
**GOU/U22/CSC/839**

## DEDICATION

This project is dedicated to God Almighty, the source of all wisdom and knowledge. To my beloved family, whose unwavering support and encouragement have been my greatest motivation throughout this academic journey.

# ACKNOWLEDGMENTS

All glory to God Almighty for the strength and wisdom to complete this project. I express my sincere gratitude to my project supervisor, Mr. Ugwu Nnaemeka, for his expert guidance, constructive feedback, and mentorship that shaped this research. I also thank the Head of Department, Dr. Frank Okebanama, and all lecturers in the Department of Computer Science and Mathematics, Godfrey Okoye University, for their academic support. My profound appreciation goes to my parents and family for their love, prayers, and sacrifices, and to my colleagues and friends for their encouragement. Finally, I acknowledge all authors whose works were cited in this project.

## ABSTRACT

The condition commonly known as prostate enlargement is also called Benign Prostatic Hyperplasia (BPH). It is among the most common urological conditions in the adult and older men populations all over the world. It is very important that an early detection of the risks of having BPH is conducted for proper handling of the risk and prevention of complications. However, the current methods for assessing risks are usually centered on consulting a physician and conducting other tests which may not be affordable to everybody. For developing the model, it became necessary to use the OOAD approach, which utilizes Python as the programming language on its back end while HTML, CSS, and Javascript are used as the front-end technologies. Also, guided questionnaires are developed to help in collecting data concerning various aspects such as age, PSA value, IPSS, BMI, nocturia, frequency of urine excretion, poor urine flow, diabetes, and other related factors. In addition, two machine learning algorithms, namely, Random Forest and Logistic Regression, are trained by applying historical datasets acquired from the clinic. Evaluation of the algorithm will involve applying techniques such as five-fold cross-validation as well as metrics including Accuracy, Precision, Recall, F1-Score, Balanced Accuracy, and AUC. A dataset containing 2,097 patient records was used for model training and validation. The developed system achieved an accuracy of 64.68%, precision of 66.08%, recall of 67.87%, F1-score of 66.96%, and AUC of 67.40%. The system provides risk classification, probability scores, management advice, and escalation recommendations. The study demonstrates the potential of machine learning-based digital health solutions for early screening and management of prostate enlargement.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER ONE

# INTRODUCTION

Benign Prostatic Hyperplasia (BPH), commonly referred to as prostate enlargement, is a benign medical condition whereby the prostate gland enlarges gradually. BPH is prevalent among adult males aged 40 years and above, and it is the most frequent cause of LUTS in elderly male patients. The prevalence rate of BPH is reported to be 20–50% among men in their 50s and 60s, and 80% among men in their 80s (Chughtai *et al*., 2020).

If BPH is left untreated or not well-managed, several health complications may arise, such as acute urinary retention, UTIs, and in some extreme cases, kidney dysfunction (Langan, 2020). In addition to its adverse health implications, BPH also imposes a significant economic burden on the healthcare sector.

According to an assessment by Okeke *et al*. (2021) in a tertiary health facility in Eastern Nigeria, roughly ₦4,966,080 is incurred each year to manage BPH among 100 patients in a single health institution.

While BPH is currently treated symptomatically due to limited preventive measures, there is growing advocacy for the need to identify at-risk persons, screen for BPH early, and offer preventive interventions among individuals with high chances of acquiring the medical condition.

## 1.1 Background of the Study

Some factors responsible for the onset of BPH include age and genetic predisposition, which account for 40-70% of cases (Roehrborn, 2021). Nevertheless, some environmental and underlying causes include obesity, insulin resistance, metabolic syndrome (such as hypertension, hypercholesterolemia, and hormonal imbalance), among others.

Several predictive models have been developed recently to forecast BPH. An example is a 2025 paper by Zi *et al*., which used a UK Biobank dataset to develop a machine learning model that predicts the likelihood of BPH based on parameters such as age, hypertension duration, blood glucose concentration, serum urate, and creatinine levels.

Prostate enlargement studies from Nigeria and other parts of Sub-Saharan Africa reveal that age continues to be the biggest risk factor for BPH. Nonetheless, there is no doubt that the genetic aspect cannot be ignored since patients with a family history of the condition have been reported as being at higher risk of developing BPH. Besides, recent urbanisation has brought about changes in lifestyle and diet. Increased intake of processed food and poor eating habits, coupled with obesity, among others, have become common, and all these aspects contribute to the prevalence of BPH (Okeke *et al*., 2021).

Healthcare facilities in Nigeria do not give sufficient attention to non-communicable diseases, as infectious diseases are given priority. Moreover, the use of advanced diagnostic and prognostic measures to predict patients at higher risk of the condition has not been widely adopted due to financial constraints and limited awareness.

Therefore, developing a diagnostic system capable of identifying high-risk patients based on parameters such as age and dietary choices will be very helpful.

## 1.2 Statement of the Problem

The problems associated with BPH are:

1. Late Diagnosis of BPH Risk
2. Inability to Combine Risk Factors
3. Lack of Personalised Management Strategies

## 1.3 Aim and Objectives of the Study

This project aims to develop a prostate enlargement risk prediction and management system for adult males.

The following are the objectives of the study;

1. Determining the challenges in the early detection of prostate enlargement in adult males.
2. Determination of the risk factors related to prostate enlargement.
3. Creation of a predictive model that can incorporate several BPH risk factors for risk analysis.
4. To design personalized management techniques using the risk assessment model.

## 1.4 Significance of the Study

The study will have great significance and value in the medical and technology domains. And some of the recipients of the benefits of the study are:

1. Adult males, who are the target audience of the study, stand to benefit a lot from this study. Having the means to carry out simple health risk assessment, they will be able to gauge their health status, prevent possible health problems and seek medical attention in case it is required. This can help to avert any further health problems.
2. The other parties that will benefit from this research are healthcare providers. They include doctors, nurses and urologists who will be assisted by the proposed technology in making correct health assessment using the patients' data. This not only enhances their ability to diagnose correctly but also personalizes the medical treatment process.
3. The other party that stands to gain from the study are hospitals and healthcare centers where the predictive technology can be used to switch the mode of service delivery from treatment to prevention. For researchers and academics, this study provides useful data, models, and

insights that can support further studies in prostate health and digital healthcare systems. It can also serve as a strong reference point for future research in similar areas.

4. Policymakers can utilize the insights gained from this research for the implementation of better health policies and programs. This research will provide evidence to make decisions regarding early diagnosis, awareness campaigns, and the role of technology in the healthcare sector.
5. Developers of health technologies can draw inspiration from the concepts and models provided in this research paper for the development of innovative health apps.

**1.4 Scope of the Study**

In order to meet the healthcare challenges caused by prostate enlargement in adult men aged 40 years and above, who are highly likely to develop BPH, the scope of the study is purposely set up. In this regard, the study focuses on the role of demographic, clinical, lifestyle and biometric data in risk prediction models which may be used for the estimation of the risk of development of prostate enlargement in a patient. With the help of risk prediction models using predictive analytics and algorithms the study will provide timely evidence-based risk stratification to help in the prevention and treatment of the disease.

This type of systems includes user-friendly interfaces, data management and decision support systems which can be implemented clinically and non-clinically for increased effectiveness and efficiency of healthcare services provided.

Finally, the study entails the implementation of management practices designed to help in implementing evidence-based and preventive interventions and to provide holistic health services with the aim to minimize complications from prostate enlargement.

## 1.6 Limitations of the Study

Limitations of the study include the following:

1. Time-related constraints: Inadequate time may result in incomplete designing, development, testing, validation, and implementation of the solution.
2. Funding constraints: Inadequate funding may prevent obtaining necessary tools such as software, IT infrastructure, data storage, cloud service, and maintenance tools, which would lead to the use of poor technologies.
3. Infrastructure constraints: Unreliable internet connection, inadequate power source, and lack of computing and healthcare IT infrastructure may hinder the implementation of the proposed solution.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.0 Introduction

This chapter will be devoted to the analysis of literature, theories, and models about the issue of benign prostatic hyperplasia (BPH) and the use of digital health technologies for its prevention, treatment, and management. In more detail, the rationale for the development of benign prostatic hyperplasia, risk factors, pathophysiology of the disease, theory of hormonal regulation, and aging will be considered. Digital health technologies including electronic health records, mobile health apps, decision support systems, predictive analytics, and artificial intelligence solutions, and their benefits in early diagnosis, treatment, monitoring, and prediction of the disease will also be analyzed.

## 2.1 Theoretical Background

Fundamental theories and principles that are the basis of the development of a model of the risk prediction and management of prostate enlargement. The theory is based on the interdisciplinary approach, which uses knowledge of medicine, epidemiology, behavioral sciences, and digital technologies in healthcare. The different theories that are the basis for approaches in this study are;

### 2.1.1 Biomedical Explanation of Benign Prostatic Hyperplasia

The focus of the study is on the biomedical explanation of the formation of benign prostatic hyperplasia (BPH), which represents a non-cancerous enlargement of the prostate gland experienced by many male patients from the age of 40 years. According to the hormonal regulation hypothesis, the conversion of testosterone to dihydrotestosterone (DHT) by the 5α-reductase enzyme promotes cell proliferation and gland enlargement. Additionally, hormonal imbalance in the testosterone/estrogen ratio in ageing individuals contributes to prostate gland growth (Chughtai *et al*., 2020; Langan, 2020).

### 2.1.2 Ageing Theory

Ageing theory offers both demographic and physiological reasons for the increase in BPH prevalence. Due to age-related changes in cells, hormonal imbalance, and reduced immune system functioning, ageing makes men vulnerable to prostate enlargement. Age is, hence, regarded as the greatest risk factor that cannot be changed. The prevalence of BPH increases tremendously among males who have crossed 50 years of age (Wei *et al*., 2025).

### 2.1.3 Predictive Modelling and Machine Learning Theory

The technology-based part of the proposed research focuses on machine learning. Classical approaches, such as logistic and Cox regression models, allow estimates of disease risk to be interpreted in terms of clinical and demographic factors. Nevertheless, as the volume of health data grows, modern ML algorithms such as Random Forests, SVMs, ANNs, and gradient boosting have become increasingly prevalent.

These technologies allow establishing non-linear links among multiple predictors and contribute to more accurate risk prediction. Implementation of predictive modelling in a digital platform allows for obtaining personalised risk scores and implementing preventive measures for those who are at risk of developing prostate enlargement.

### 2.1.4 Behavioural Change Theories

BPH management not only involves clinical prediction but also patient engagement and lifestyle changes. This research employs essential behavioural models to promote adherence to preventative and treatment approaches:

- Health Belief Model (HBM): Identifies why people's beliefs about their susceptibility, seriousness, benefits, and barriers affect their decision-making regarding screening and treatment.
- Transtheoretical Model (TTM): States that behaviour change is a process through pre-contemplation, contemplation, preparation, action, and maintenance stages.
- COM-B Model (Capability, Opportunity, Motivation – Behaviour): Argues that behaviour change requires individuals to have adequate capability, opportunity, and motivation.

The theories aid in developing user-centred interventions within the system, such as educational materials, alerts, and tailored suggestions.

### 2.1.5 Integrated Theoretical Framework

By combining biomedical, epidemiological, behavioural, and technological theories, this study has taken an interdisciplinary approach. The combination of all these theories has helped in developing a comprehensive digital system which would be able to:

Predict the risk for developing prostate enlargement on an individual basis.

- Help in clinical decision making;
- Facilitate behavioural changes; and
- Improve preventive health care methods.

## 2.2 Literature Review

Benign Prostatic Hyperplasia (BPH), also known as prostate enlargement, is among the most common urological diseases affecting aging male individuals around the world. The high burden of the disease has led to an increase in research interest in risk assessment, disease prediction, diagnosis, and treatment using advanced computer-based techniques. Some recent studies have shown an increasing application of artificial intelligence (AI) and machine learning (ML) to enhance decision making and predict the outcome of the disease.

Wei *et al*. (2025) analyzed the burden of BPH worldwide through the use of data derived from the Global Burden of Disease (GBD) database in the period ranging from 1990 to 2021 and projected up to 2035. It was found out that despite the low variation in age-standardized incidence and prevalence of the disease over the years, the total number of BPH patients has sharply increased owing to population growth and aging. According to Wei *et al*. (2025), BPH accounts for DALYs in particular for Eastern European and highly populated regions. In addition, projections showed that the incidence and prevalence of BPH are expected to increase over time worldwide.

Machine learning algorithms have been widely applied for prostate disease prediction in recent times. Zhou *et al*. (2025) developed a predictive model utilizing machine learning algorithms to predict the risks of prostate disease based on non-invasive clinical and laboratory indicators. The purpose of the study was to classify patients with and without prostate diseases based on non-invasive indicators prior to any invasion procedure like biopsy. It was evident from the findings that machine learning algorithms could effectively use patient data and create risk predictions in the form of visualization that would help in making medical decisions.

Similarly, Shen *et al*. (2025) have created a predictive model for predicting the probabilities of acute urinary retention after prostate biopsy. Data of 599 patients were analyzed and the advanced feature selection methods LASSO and Boruta algorithms were used. There were six risk factors which were identified by the researchers; BMI, prostatic volume, diabetes status, constipation, IPSS, and residual urine volume. A good prediction was done by the generated nomogram through validations by ROC curve, calibration analysis, and decision curve analysis.

Another research work that involved the use of the UK Biobank dataset for predicting the likelihood of developing BPH was conducted in 2025. In total, more than 210,000 individuals were included in the analysis, with a follow-up time of approximately thirteen years. Over the course of this time, almost 4,800 patients had developed BPH. As the result of the investigation, it has been found that among all other factors, age is the strongest predictor of BPH, whereas additional predictors include the duration of hypertension, blood glucose levels, uric acid level, and serum creatinine level. Among all machine learning models, the best results have been obtained by Light Gradient Boosting Machine (LightGBM). Moreover, a simplified model with the help of a small amount of clinical factors has been proposed.

The efficacy of machine learning approaches in treating BPH was also analyzed through systematic reviews. In their review of studies that employed different machine learning algorithms, including Random Forests, SVMs, ANNs, and Logistic Regression in predicting, diagnosing, and managing BPH, Bouhadana *et al*. (2023) reported that machine learning approaches were better than traditional statistical approaches in predicting disease progression and resistance to treatment. They also reported the predictive importance of non-invasive clinical features, such as PSA level, prostate volume, age, and IPSS scores. Some of the limitations of using machine learning approaches in BPH management include small dataset size, lack of robust validation, and implementation in the clinical setting.

In their analysis of machine learning approaches in predicting surgical outcomes among patients being treated for benign prostatic enlargement, Mourmouris *et al*. (2021) used decision trees, neural networks, and logistic regression models employing preoperative clinical features. They found that the application of machine learning approaches in predicting surgical outcomes resulted in high accuracy and provided clinicians with information about which patients would benefit from surgery without experiencing any postoperative complications.

Additional research demonstrating the usefulness of machine learning for predicting BPH risks was carried out in a study in 2024 that used such clinical factors as PSA levels, size of the prostate gland, body mass index, age, and other co-morbid conditions to predict the course of the disease. Different machine learning models were used, such as Support Vector Machines, Random Forests, and XGBoost. It was found that an ensemble machine learning model, especially XGBoost, produced the best results in terms of accuracy and sensitivity. Therefore, the study revealed that the use of machine learning models for predictions can be beneficial for identifying high-risk patients and developing an individualized treatment strategy.

Artificial intelligence has also been used in the context of prostate imaging. The work of Zhao *et al*. (2025) was a meta-analysis aimed at evaluating the efficiency of AI in image analysis of multiparametric magnetic resonance imaging (mpMRI). The main point of interest in this research was associated with deep learning and radiomics in terms of detecting and predicting prostate diseases. It was found that AI-aided image analysis resulted in higher sensitivity and specificity than the conventional methods of image analysis. However, some challenges were also discovered during this research.

Another systematic review and meta-analysis by Ling *et al*. (2025) has analyzed the prognostic efficacy of various machine learning models for different diseases of the prostate gland (BPH and prostate cancer). The analysis included machine learning models such as Random Forest, Gradient Boosting, Artificial Neural Networks, and deep learning models. The results of the analysis indicated high predictive efficiency, where the Area Under the Curve (AUC) values were in the range of 0.80 to 0.85. The authors discussed the ability of machine learning models to process clinical, laboratory, and imaging data for individualized treatment of patients. Overfitting, heterogeneity among the studied papers, and lack of external validation remain problematic issues in the use of machine learning models.

Overall, the reviewed literature provides evidence of the great potential of machine learning and artificial intelligence models in enhancing the prediction, diagnostics, and treatment of the prostatic enlargement and other related diseases. Most of the papers reported high predictive performance of the proposed techniques and the importance of integration of several clinical and diagnostic factors. However, some challenges in terms of validation, imbalanced datasets, and implementation in the clinical practice are still relevant. These gaps provide a strong justification for developing a robust machine learning-based system for predicting the risk of prostate enlargement among adult men.

### 2.3 Summary of Reviewed Works

**Table 2.1:** Summary of reviewed works on BPH prediction and management

| S/N | Author(s) | Title | Methodology | Similarities | Differences |
|---|---|---|---|---|---|
| 1 | Wei *et al*. (2025). | Global Trend of Benign Prostatic Hyperplasia | Systematic epidemiological analysis using Global Burden of Disease (GBD) data and statistical trend projections. | Both studies focus on BPH as the primary condition in adult males and recognise age as the dominant risk factor. Both aim to inform evidence-based strategies for prevention and management of prostate enlargement. | Wei *et al*. focused on global epidemiological burden and trend projections using GBD data and did not develop a predictive or clinical decision-support tool. The proposed system, by contrast, is a web-based, individual-level risk prediction and management tool built for clinical use in a Nigerian context, integrating personalised risk scoring and management advice. |
| 2 | Zhou *et al*. (2025). | Machine Learning for Prostate Risk Assessment | Supervised machine learning using clinical and laboratory data with model visualisation for clinical interpretation. | Both studies apply supervised machine learning to predict prostate-related conditions using clinical and laboratory variables such as PSA level and patient age as key predictors. Both share the goal of enabling earlier identification of at-risk patients. | Zhou *et al*. focused on distinguishing patients with and without prostate disease prior to biopsy, while the proposed system focuses specifically on predicting BPH risk in a community setting and provides personalised management strategies through a full web-based application accessible without biopsy procedures. |
| 3 | UKBiobank Study (2025). | BPH Risk Prediction Model | Longitudinal cohort study employing statistical analysis and machine learning (LightGBM) with external validation and an online risk calculator. | Both studies develop a web-accessible BPH risk prediction model using machine learning and clinically validated risk factors such as age, PSA level, and metabolic variables. Both aim to provide an online risk calculator to support preventive screening. | The UK Biobank study used a large European cohort of over 210,000 participants and the LightGBM algorithm validated over 13 years. The proposed system is designed for Nigerian healthcare context, uses a Random Forest classifier trained on a smaller local dataset, and integrates personalised management advice and escalation guidance into the output. |
| 4 | Ling *et al*. (2025). | Machine Learning for Prostate Disease Prognosis | Systematic review and meta-analysis of studies employing Random Forest, Gradient | Both affirm the effectiveness of machine learning models, particularly Random Forest, for | Ling *et al*. conducted a meta-analysis evaluating existing ML models across multiple studies rather than building a new system. The proposed system |

| | | | | | |
|---|---|---|---|---|---|
| | | | Boosting, ANN, and deep learning models. | predicting outcomes in prostatic diseases, and both highlight AUC as a key performance metric. Both recognise the value of multi-source clinical data integration for improved prediction accuracy. | implements and deploys an original Random Forest model as a functioning clinical decision-support tool with a front-end questionnaire interface for real-world use. |
| 5 | PubMed Study (2024) | ML-Based Risk Prediction in BPH Patients | Supervised machine learning using SVM, Random Forest, and XGBoost with clinical predictor variables. | Both studies apply ensemble machine learning methods to classify BPH risk using shared clinical features such as PSA level, BMI, and age, and both recognise that ML-based risk prediction can support personalised treatment decisions for BPH patients. | The PubMed study evaluated multiple algorithms including XGBoost and SVM to identify the best performer for general BPH risk progression, while the proposed system deploys Random Forest in a full-stack web application designed for real-world clinical use in Nigeria with an integrated questionnaire, risk classification, and management output. |

### 2.4 Research Gap

Despite extensive research on benign prostatic hyperplasia and advancements in digital healthcare technologies, several critical gaps remain that justify the need for this study, and they are;

1. Most existing studies treat **risk prediction, diagnosis, and management as separate processes**. There is a lack of integrated systems that combine these components into a unified platform. Moreover, this study integrates these different processes.
2. The digital BPH health systems constructed in earlier studies usually ignore the requirements of the targeted population—middle-aged and older male adults. Issues such as poor digital literacy skills, linguistic obstacles, financial constraints, and availability are hardly ever considered. However, this study creates an accessible digital health system that suits the targeted male population.

3. Current models of prediction normally focus on laboratory and clinical predictors but overlook other significant determinants including genetics, hormones, lifestyle, environment, and socio-economic factors. This study incorporates several factors for accurate prediction including demographic, clinical, metabolic, and behavioral characteristics.

# CHAPTER THREE

## SYSTEM ANALYSIS & DESIGN / RESEARCH METHODOLOGY

### 3.0 Introduction

This chapter is dedicated to the introduction of the system analysis, the design framework, and the research methodology that will be needed for the development of the Prostate Enlargement Risk Prediction and Management System. In addition to that, the chapter includes the analysis of the current systems and processes employed to diagnose and manage the health problems in question, as well as the description of the proposed new system.

Furthermore, this chapter focuses on the discussion of the research methodologies, procedures, and methods that were utilized while developing the predictive digital health system. In addition to that, the chapter serves as the technical basis of the paper by putting theories into practice.

### 3.1 System Analysis

It entails the thorough examination of current processes used in tackling the problem in question, listing their strengths, weaknesses and improvements that need to be incorporated in the proposed system. This chapter analyzes the current methodologies used in identifying the risk of prostate enlargement in males and points out its weaknesses. Finally, the goals of the proposed system are stated.

#### 3.1.1 Analysis of the Existing System

The current method used in risk assessment as far as prostate enlargement is concerned in Nigeria and other developing nations involves consulting and examination of men especially using the Digital Rectal Examination method. At Tertiary Hospitals, risk assessment can be achieved by means of PSA blood test coupled with IPSS symptom intensity test. TRUS and Prostate biopsy can equally be carried out depending on availability of facilities. Risk stratification as far as it may be done is manually undertaken by urologists who take into account all the factors mentioned above.

It should be noted here that the presence of such structured tools is rare in the field of community medicine and general medical practice. This suggests that patients visit hospitals only when the condition is really serious; also, there has never been an integration of all the above-mentioned risk factors to calculate the risk involved. In addition, there is not much material to educate and counsel patients expected to suffer from BPH (Okeke *et al*., 2021).

### 3.1.2 Operational Characteristics

i. **Manual Data Collection:** Data is collected through physical forms and interviews, resulting in inconsistent and incomplete data.

ii. **Use of Paper-Based Records**: There is an existing use of paper-based records or poorly organised digital records.

iii. **No Automation**: There is no automation for risk prediction or monitoring.

iv. **Dependence on Symptom Presentation**: The diagnosis of patients is mainly reactive and depends on symptom presentation.

v. **Dependence on Clinician-Centred Decision Making:** Clinical decisions are mainly based on the expertise and experience of individual clinicians.

vi. **Fragmented Care Delivery**: There is no integration between different departments and facilities.

vii. **Poor Use of Digital Infrastructure**: There is poor adoption of health information systems and decision support tools.

### 3.1.3 Limitations of the Existing System

There are a number of limitations associated with the existing system, both in terms of its structure and its functionality. These are as follows:

i. **Late Diagnosis**: The system also suffers from the problem of late diagnosis of patients.
ii. **Lack of Predictive Functionality**: The system lacks a digital system that can be used to predict the risks associated with a patient.
iii. **Poor Health-Seeking Behaviour:** The system also suffers from the problem of poor health-seeking behaviour.
iv. **Lack of Access to Specialists**: The system lacks urologists and other specialists.
v. **Inefficient Data Management**: The system also suffers from inefficient data management.
vi. **Inconsistent Clinical Decisions**: The system also suffers from the problem of inconsistent clinical decisions.
vii. **Lack of Continuous Patient Monitoring**: The system also lacks the feature of continuous patient monitoring.
viii. **High Treatment Costs**: The system also suffers from the problem of high treatment costs.
ix. **Poor Patient Engagement**: The system also suffers from the problem of poor patient engagement.

### 3.1.4 Implications of the Existing System

The implications of the existing system include increased disease, increased complications, decreased quality of life, and increased mortality rates. In addition, the inability of healthcare providers to utilise data-driven insights for the early detection and prevention of diseases is another implication of the system. Furthermore, the inability of the management system to efficiently manage the system is another critical implication of the system.

### 3.1.5 The Need for System Transformation

The need for system transformation stems from the implication of the existing system. Through the analysis of the existing system, it is clear that there is an urgent need for system transformation. The need for an automated risk prediction system, a digital platform, predictive analytics, and a management system stems from the implication of the existing system. The need for system transformation will form the basis of the design of the proposed system.

## 3.2 Analysis of Proposed System

The proposed system will be a digital prostate enlargement risk prediction and management system that will transform the current conventional nature of the healthcare system to a more predictive and preventive nature of digital health care.Unlike the current system, the proposed system will be a patient-centred, data-driven, and technology-enabled model of health care, where the delivery of health care is facilitated by automated processes, data analysis, and intelligent decision-making. The proposed system will be used by the entire health care system, patients, and health care providers through a unified digital health care system.

The proposed system will be a web-enabled system, which means that health care services will be accessed remotely. This is essential for improving the accessibility of health care services, especially in developing countries.

### 3.2.1 Functional Requirement

The proposed system shall be able to:

1. Support user registration and authentication for patients, healthcare
2. Enable secure login and access control using role-based permissions.
3. Allow patient profiles to be created.

4. Support the input of health data, such as demographic, clinical, lifestyle, and medical history data.
5. Automatically predict patient risk using predictive algorithms.
6. Display risk scores and classifications, such as low, medium, and high risk.
7. Enable clinical decision support for healthcare providers.
8. Enable the tracking of patient symptoms.
9. Display recommendations based on patient data.
10. Enable the generation of reports.

### 3.2.2 Non-Functional Requirement

The system should support the following non-functional requirements:

1. Security: Support for data encryption, authentication, authorisation, and secure communication.
2. Privacy: Support for data protection legislation.
3. Scalability: Support for handling an ever-increasing number of users.
4. Reliability: High system availability and robustness.
5. Availability: The system should be available 24/7.
6. Usability: User-friendly interfaces.
7. Performance: High processing speed.
8. Interoperability: The system should be able to support the integration of other health information systems.
9. Maintainability: Ease of updating the system.
10. Portability: The system should be able to support other platforms, such as web-based and mobile platforms.

### 3.3 System Design Methodology

The system was developed using Object Oriented Analysis and Design approach (OOAD). Object-Oriented Analysis and Design is a technique used in the software development industry to analyze and design a system as an object-oriented structure consisting of several objects. Objects have data and functionality characteristics. This approach has provided a very intuitive approach to modeling the real world into the software world as it includes many entities in the real world like actors, datatypes, and processes.

#### 3.3.1 Justification for Methodology

The reasons for adopting OOAD in developing this system include:

1. Multiple Interacting Entities - There exist multiple interacting objects in the system such as patients, administrators, questionnaires, prediction engine, and output of results. All these interacting objects can be modeled as an object that has its own attributes and behaviors.

2. Modularization and Maintenance- This methodology helps to modularize the design, thus making it possible to test and maintain the various elements of the system such as FeatureEngineer, PredictionService, TrainingPipeline, Logistic Regression, and Random Forest Model in isolation.

3. Integration of Machine Learning Elements – With this methodology, it becomes possible to design the system such that all the machine learning elements become part of the overall system architecture. Hence, it becomes easy to incorporate the logistic regression model and the prediction random forest model.

4. Suitability for Python Programming Language – Given that the Python programming language is object-oriented by nature, this methodology can help in designing systems using the Python programming language.

5. Scalability and Extensibility - Further developments in the project can be undertaken using this methodology owing to the fact that there are chances of integrating more features, prediction models, and management modules into the project.

### 3.3.2 Method Of Data Collection

The system used two method in data collection which includes;

1. Collection of Dataset Through History-Based Dataset: The training data will be obtained through the use of the historical dataset provided in data/real_dataset.csv. The dataset consists of ten features proven to be clinically validated such as age, PSA level, IPSS score, BMI, nocturia, frequency of urination, weak stream, family history, diabetes, and BPH risk level. However, for cases when no real dataset is available, the training program generates a simulated dataset based on the same format as the existing dataset.
2. Live Questionnaire-Based Collection of Dataset: For clinical data collection, the live questionnaire will be used at runtime. The information entered by the user includes age, PSA level, IPSS score, BMI, nocturia, daily frequency of urination, weak stream, family history, diabetes, and symptoms. Once all inputs have been sent to the backend API, then every prediction result will be logged in.

   The questionnaire consists of the following items:

1. What is your age?
2. What are your PSA level and IPSS score?
3. What is your BMI, and how frequently do you urinate during the day?
4. Are you experiencing a weak urinary stream?

**Optional Questions (where applicable):**

1. Do you have any chronic risk factors or underlying medical conditions, such as diabetes and family prostrate history?
2. Are you currently experiencing any urgent warning signs or severe symptoms that require immediate medical attention, such as Blood in urine, Fever and Severe pelvic or lower abdominal pain?

### 3.3.3 Dataset Description and Preprocessing

A structured clinical dataset, available at data/real_dataset.csv, was used to train the model. It consists of ten attributes that are demographic, clinical, and symptoms related features collected from adult male patients. The fields along with the description are as follows.

**Table 3.1:** Dataset fields and descriptions

| Field | Description | Data Type |
|---|---|---|
| age | Patient age in years | Continuous |
| PSA_Level | Prostate-specific antigen level in ng/mL | Continuous |
| IPSS_Score | International Prostate Symptom Score (0–35) | Continuous |
| BMI | Body mass index (kg/m²) | Continuous |
| Nocturia | Number of night-time urination episodes | Integer |
| Urinary_Frequency | Number of daytime urination episodes | Integer |
| Weak_Stream | Weak urinary stream severity (0 = No, 1 = Yes) | Binary |
| Family_History | Family history of prostate-related problems (0 = No, 1 = Yes) | Binary |
| Diabetes | Presence of diabetes or metabolic syndrome (0 = No, 1 = Yes) | Binary |
| BPH_Risk | Target classification label (0 = Low risk, 1 = High risk) | Binary |

All the dataset comprised of 2,097 cases, out of which the training dataset accounted for 1,678 cases (about 80%), while the validation dataset had 419 samples (about 20%).

Different preprocessing operations included data type transformation through changing numeric data to float and Boolean data to integer. In addition, the process involved normalization of all the clinical factors to ensure that they have similar scales and creating some more compound factors. The formulas used for normalization included the following:

I. age = (age - 35) / 45
II. psaLevel = psaLevel / 10
III. ipssScore = ipssScore / 35
IV. bmi = (bmi – 18) / 22
V. nocturia = nocturia / 5
VI. frequency = (frequency – 4) / 12
VII. streamStrength = streamStrength / 3

**3.3.4 Model Architecture and Training**

Two classification models were used which are:

I. logistic regression
II. random forest

In the process of building these classifiers, no machine learning packages were used; only native Python mathematical operations were implemented in order to keep up with the dependency-free approach.

To begin with, logistic regression is a linear model where BPH is predicted based on the probability computed through the sigmoid function of the weighted input vector of normalized and engineered features. Moreover, a non-linear classifier random forest involves the training process of several decision trees using randomly selected samples of the training data and averaging their probabilities.

The training process of the model included seven steps:

I. loading and preprocessing the data

II. stratifying the data into the training and validation subsets

III. performing fivefold cross-validation on both classifiers over the training subset

IV. selecting the one that has the highest average accuracy out of two for each fold

V. Training the selected model on the full training subset

VI. computing its performance over the validation subset

VII. Measuring the accuracy, AUC, precision, recall, F1 score, and balanced accuracy and saving model artifacts and performance results.

Random forest was selected as the deployed model based on its superior cross-validated accuracy. The validation results are presented in the Table below.

**Table 3.2:** Deployed model performance metrics on the validation set

| **Metric** | **Result** |
|---|---|
| Selected Algorithm | Random Forest |
| Accuracy | 64.68% |
| Area Under the ROC Curve (AUC) | 67.40% |
| Precision | 66.08% |
| Recall (Sensitivity) | 67.87% |
| F1 Score | 66.96% |
| Balanced Accuracy | 64.49% |
| Classification Threshold | 0.50 |
| Training Set Size | 1,678 records |
| Validation Set Size | 419 records |

### 3.4 System Modelling Using UML

The system modelling was done through diagrams similar to those used in UML. These include the use case, activity, and class diagrams. The diagrams shown below include the use case diagram, activity diagram, and class diagram.

### 3.4.1 Use Case Diagram

The use case diagram identifies the two primary actors in the system — the User (patient or screening candidate) and the Administrator (researcher or system operator) — and maps each actor to the system functions they interact with.

**Table 3.3:** Use case descriptions for the Prostate Enlargement Risk Predictive System

| Actor | Use Case | Description |
|---|---|---|
| User / Patient | Register/Login | Creates an account or signs in to access the system's prediction and guidance features. |
| User / Patient | Answer Questionnaire | Responds to the guided questionnaire by providing clinical and demographic information required for risk assessment. |
| User / Patient | View risk prediction | Views the predicted risk classification and probability score |
| User / Patient | View management Guild | Reviews personalised care and lifestyle recommendations based on the predicted risk level. |
| Doctor/Researcher | Register/Login | Creates an account or signs in to access administrative and research functionalities. |
| Admin / Researcher | Upload Questionnaire | Creates, updates, and manages questionnaire items used for data collection and risk assessment. |
| Admin / Researcher | Retrain model | Executes the training pipeline to update the predictive model with new data |
| Admin / Researcher | Generate evaluation report | Reviews the performance metrics produced after model training |
| Admin / Researcher | Review prediction records | Accesses historical prediction logs for analysis and quality review |

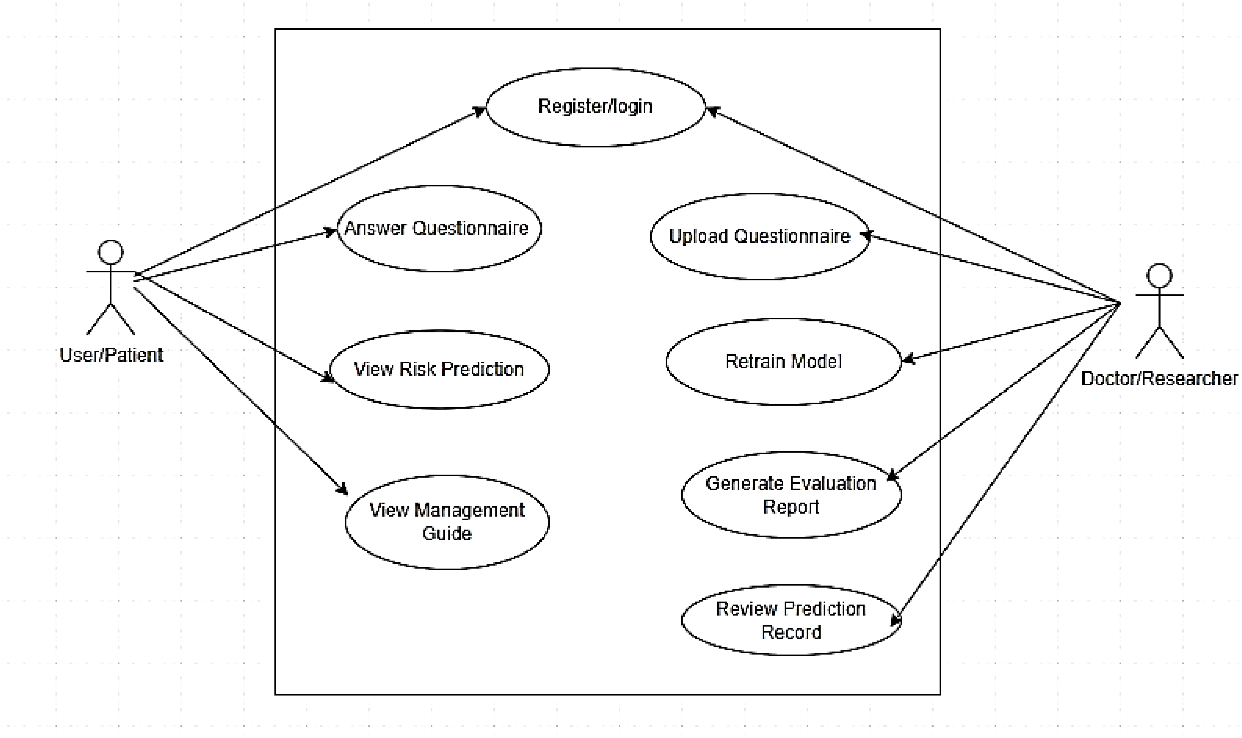


**Figure 3.1:** Use Case diagram of the Prostate Enlargement Risk Predictive System

### 3.4.2 Activity Diagram

The activity diagram illustrates the workflow starting from the opening of the web interface by the user up until the delivery of the prediction result. Specifically, the above-mentioned workflow includes the following stages:

1. The user opens the browser interface.
2. The user fills in the questionnaire step-by-step.
3. The client-side validation of filling out the questionnaire is carried out on each step. In case the data is invalid, a warning message appears, and the user should make amendments accordingly.
4. The valid JSON input is uploaded to the /api/predict endpoint.

5. Normalization is conducted according to the same formulas that were utilized for training. Predictions are also performed at this stage.
6. Probability calculation occurs. The result is assigned to the category of Low, Moderate, or High.
7. The result summary is generated along with management recommendations and the next steps to be taken.
8. The result is recorded in the JSONL file.
9. The prediction result is sent to the frontend part.

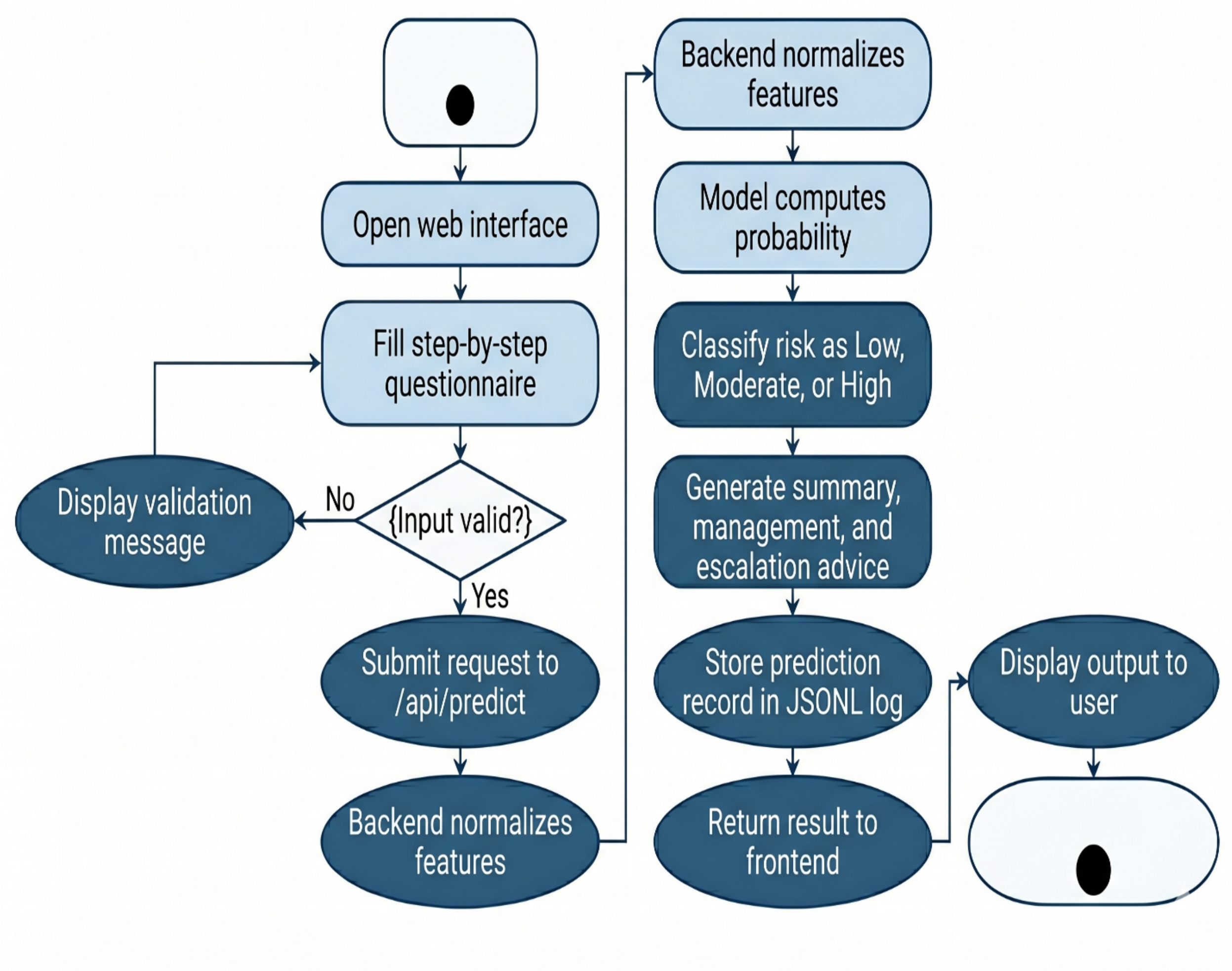


**Figure 3.2:** Activity diagram of the Prostate Enlargement Risk Predictive System workflow

### 3.4.3 Class Diagram

The class diagram illustrates the logical modules present in the software along with their interaction. There are five main classes whose functionalities are explained in the table that follows.

**Table 3.4:** Class diagram module descriptions

| Class / Module | Key Methods / Attributes | Responsibility |
|---|---|---|
| AppHandler | do_GET(), do_POST(), _write_json(payload, status) | Handles incoming HTTP requests and routes them to the appropriate service |
| PredictionService | predict(payload), build_profile(probability, flags), save_prediction_record(), load_recent_records(limit) | Executes inference, generates the risk profile and advice, and manages record storage |
| FeatureEngineer | normalize_row(features), to_feature_vector(features), to_engineered_feature_vector(features), describe_modifiers(features) | Applies normalization and feature engineering transformations to raw questionnaire inputs |
| ModelArtifact | algorithm, feature_order, weights / trees, metrics, train_size, validation_size | Stores the serialised trained model and its evaluation metadata in model.json |
| TrainingPipeline | load_real_dataset(), generate_synthetic_dataset(), split_dataset(), evaluate_scored_rows(), write_report(), save_model() | Implements the end-to-end model training, comparison, evaluation, and persistence workflow |

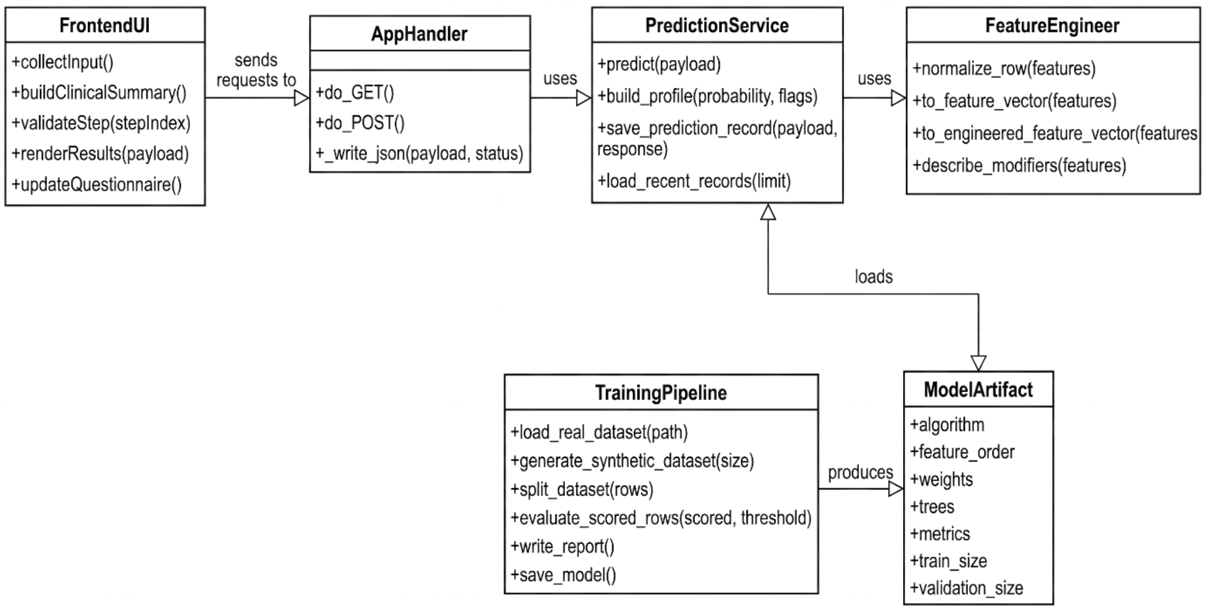


**Figure 3.3:** Class diagram showing module relationships

### 3.5 System Design

The design of the proposed solution will emphasize the development of a highly efficient, dependable, and maintainable web application that is meant to provide predictions regarding the risk of developing prostatic hyperplasia. The system will comprise three main components – the frontend, a backend created using Python programming language, and a prediction engine with machine learning algorithms enabling screening of BPH patients.

### 3.5.1 Input Design

Input design will be conducted through a questionnaire with six steps designed to eliminate the possibility of mistakes by the user and ensure proper gathering of relevant information. Every item is being validated prior to the submission to prevent errors.

Table 3.5**:** Input field design for the guided questionnaire

| S/N | Input Field | Type | Purpose |
|---|---|---|---|
| 1 | Age | Number | Captures patient age |
| 2 | PSA Level( Postrate Specific Antigen ) | Number | Captures prostate-specific antigen value |
| 3 | IPSS Score(international Postrate System Score) | Number | Captures symptom severity |
| 4 | BMI(Body Mass Index) | Number | Captures body mass index |
| 5 | Nocturia | Number | Documents night time urination events |
| 6 | Frequency | Number | Documents day time urination frequency |
| 7 | Weak Stream | Select | Shows presence of weak primary stream |
| 8 | Family History | Checkbox | Shows presence of inherited risk factor |
| 9 | Diabetes | Checkbox | Shows presence of metabolic risk factor |
| 10 | Hematuria | Checkbox | Indicates blood in urine |
| 11 | Fever | Checkbox | Shows possible infection |
| 12 | Severe Pain | Checkbox | Shows a critical warning sign |

### 3.5.2 Output Design

The output module aims at being clear and understandable. The system will produce relevant results based on the input information contained in the user-submitted questionnaires.

The system will show:

- Predicted risk level (Low, Moderate, or High)
- Risk percentage score
- Prediction summary
- Extracted patient feature summary
- Management strategies

**Table 3.6:** Output Element design

| Output Element | Description |
|---|---|
| Risk Label | Final class label returned by the model |
| Probability Score | Percentage estimate of risk |
| Summary | Short interpretation of the prediction |
| Feature Summary | Human-readable list of submitted factors |
| Management Advice | Suggested care or follow-up steps |

### 3.5.3 Database Design

For the storage of all data within the system, a relational database management system (RDBMS) has been selected. The database architecture employs the SQLite RDBMS during development time and can easily be switched to PostgreSQL at its operational time. The database tables are listed below:

**Table 3.7:** Users Table

| S/N | Column Name | Data Type | Constraint | Description |
|---|---|---|---|---|
| 1 | user_id | INTEGER | PRIMARY KEY AUTOINCREMENT | Unique auto-incrementing identifier for each registered user |
| 2 | full_name | VARCHAR(100) | NOT NULL | Full name of the registered user |
| 3 | email | VARCHAR(100) | NOT NULL UNIQUE | Email ID employed as the login username; needs to be unique for all users |
| 4 | password_hash | VARCHAR(225) | NOT NULL | Password stored in a hashed form for validation |
| 5 | role | VARCHAR(25) | NOT NULL DEFAULT 'user' | Role assigned to the account: user (patient) or admin (researcher/operator) |
| 6 | created_at | DATETIME | NOT NULL | Timestamp recording when the account was created |
| 7 | last_login | DATETIME | NULL | Timestamp of the most recent successful login session |

**Table 3.8:** Questionnaire table

| S/N | Column Name | Data Type | Constraint | Description |
|---|---|---|---|---|
| 1 | questionnaire_id | INTEGER | PRIMARY KEY AUTOINCREMENT | Unique identifier for each uploaded questionnaire record |
| 2 | user_id | INTEGER | FOREIGN KEY REFERENCES users | References the admin user who uploaded or last updated the questionnaire |
| 3 | title | VARCHAR(150) | NOT NULL | Descriptive title of the questionnaire |
| 4 | questions | TEXT | NOT NULL | JSON-serialised list of question items included in the questionnaire |
| 5 | version | VARCHAR(20) | NOT NULL | Version label for tracking revisions to the questionnaire over time |
| 6 | uploaded_at | DATETIME | NOT NULL | Timestamp recording when the questionnaire was uploaded or last updated |
| 7 | is_active | BOOLEAN | NOT NULL DEFAULT 1 | Indicates whether this questionnaire is currently active: 1 = Active, 0 = Inactive |

**Table 3.9:** Dataset table

| S/N | Column Name | Data Type | Constraint | Description |
|---|---|---|---|---|
| 1 | dataset_id | INTEGER | PRIMARY KEY AUTOINCREMENT | Unique auto-incrementing identifier for each dataset record |
| 2 | age | INTEGER | NOT NULL | Patient age in years |
| 3 | psa_level | FLOAT(4.2) | NOT NULL | Prostate-Specific Antigen concentration expressed in ng/mL |
| 4 | ipss_score | INTEGER | NOT NULL | International Prostate Symptom Score used to evaluate urinary symptoms |
| 5 | bmi | FLOAT(4.2) | NOT NULL | Body Mass Index of the patient in kg/m² |
| 6 | nocturia | INTEGER | NOT NULL | Number of night-time urination episodes |
| 7 | urinary_frequency | INTEGER | NOT NULL | Frequency of urination reported by the patient |
| 8 | weak_stream | BOOLEAN | NOT NULL | Indicates the presence of weak urine stream: 1 = Yes, 0 = No |
| 9 | family_history | BOOLEAN | NOT NULL | Indicates whether there is a family history of BPH: 1 = Yes, 0 = No |
| 10 | diabetes | BOOLEAN | NOT NULL | Indicates whether the patient has diabetes: 1 = Yes, 0 = No |
| 11 | bph_risk | INTEGER | NOT NULL | Target risk classification generated for BPH prediction |

**Table 3.10:** Model_artifact table

| S/N | Column Name | Data Type | Constraint | Description |
|---|---|---|---|---|
| 1 | model_id | INTEGER | PRIMARY KEY AUTOINCREMENT | Unique identifier for each trained machine learning model |
| 2 | dataset_id | INTEGER | FOREIGN KEY REFERENCES dataset | References the dataset used for model training |
| 3 | algorithm | VARCHAR(20) | NOT NULL | Name of the machine learning algorithm used for training |
| 4 | feature_order | TEXT | NOT NULL | Stores the ordered list of features used during training |
| 5 | metrics | TEXT | NOT NULL | Stores model evaluation metrics such as accuracy and F1-score |
| 6 | train_size | INTEGER | NOT NULL | Number of records used for training the model |
| 7 | validation_size | INTEGER | NOT NULL | Number of records used for model validation and testing |

**Table 3.11:** predictions_record table

| S/N | Column Name | Data Type | Constraint | Description |
|---|---|---|---|---|
| 1 | prediction_id | INTEGER | PRIMARY KEY AUTOINCREMENT | Unique identifier for each prediction record |
| 2 | user_id | INTEGER | FOREIGN KEY REFERENCES users | References the registered patient who submitted the prediction request |
| 3 | model_id | INTEGER | FOREIGN KEY REFERENCES model_artifact | References the trained model used to generate the prediction |
| 4 | timestamp | DATETIME | NOT NULL | Date and time the prediction was generated |
| 5 | features | TEXT | NOT NULL | JSON-serialised patient feature values submitted via the questionnaire |
| 6 | prediction | VARCHAR(20) | NOT NULL | Predicted BPH risk classification generated by the model |
| 7 | profile | VARCHAR(225) | NOT NULL | Summary profile or interpretation associated with the prediction result |

### 3.5.4 System Architecture Design

The architecture of the system make use of a simple client-server-machine learng structure. It made up of three major layrs:

1. **Presentation Layer (Frontend)**

This layer hanles user ineraction and data input. It is developd using:

- index.html
- styles.css
- app.js

Users interact with the questionnaire form through the browser interface, where they submit age, PSA level, IPSS score, BMI, and urinary symptoms.

**2. Application Layer (Backend)**

This laer contains the business logicy and API processing. It is implemnted using:

- servyer.py

The backend receivs user input, valdates the data, loads the trained moel, perfoms a prediction, applies rule-based warnig flags, and retrns the final result.

It also manges:

- Prediction record storage
- API response generation
- Health checks
- Retrieval of prediction history

**3. Intelligence/Data Layer**

This layer contains the machine lerning model and supporing data files. It icludes:

- model.json
- train_model.py
- evaluation_report.json
- files inside the data directory

The training pipeline uses the real dataset to train the prediction model, evaluate performance, and generate reports for system validation.

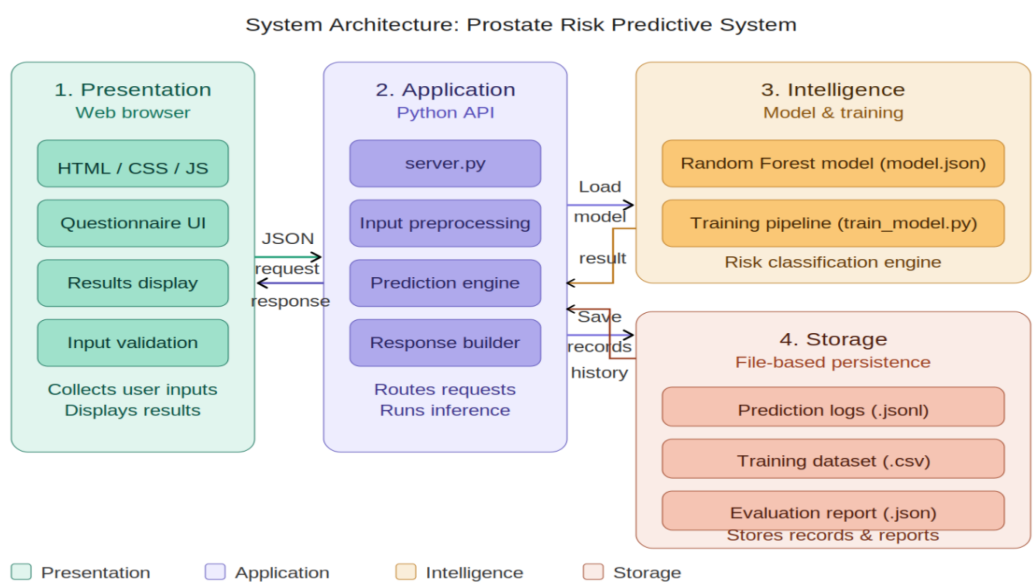


**Figure 3.4:** System architecture diagram shwing the three-layer client-server structure

### 3.5.5 Algorithm Design

The prediction prcess follows a structured mach7ine learning workflow that combines staatistical prediction with cliniical rule-based intertpretation.

**Prediction Workflow**

1. User navigates to the web interface

2. User fills out the interactive form
3. The system verifies all the input values
4. Inputs are pushed to /api/predict
5. The backend prepares the features
6. The model gets loaded from model.json
7. The machine learning algorithm predicts the risk probability of BPH
8. The risk level is classified as Low, Moderate, or High
9. Business rules identify warnings like high PSA or intense symptoms
10. The output prediction is gotten

# CHAPTER FOUR

# SYSTEM IMPLANTATION

## 4.0 Introduction

This chapter is going to address the implementation of the Prostate Enlargement Risk Predictive System which was developed in Chapter Three. The development environment used to develop the system, the data set used to develop the predictive system, pre-processing of the dataset, training and evaluating of the model, implementation of the predictive system modules and the tests done to ensure that the system meets all the requirements will be addressed in this chapter. Finally, this chapter will conclude by addressing the results obtained from the system.

## 4.1 Development Environment and Tools

In this task, the development environment used was selected considering the need of simplicity and portability of the system. This is because the system needs to be implemented in educational computer systems. This system has been designed in such a way that it uses minimum dependencies, meaning that it requires only Python 3 and web browser

### 4.1.1 Programming Language and Libraries

Programming languages used include Python 3 for backend software development and machine learning; and HTML, CSS, and JavaScript languages for user interface implementation. The reasons for selection of programming languages are provided below.

1. The choice of Python 3 was driven by its ubiquitous use in the academic environment, the presence of complete functionality through the standard library (thereby eliminating the

need for any third party installations), and frequent use in machine learning. The Python modules used included http.server, which is needed for the implementation of the backend API; json module required for managing the model artifacts and records; csv, to read the training data file; math, for sigmoid activation and probability computation; and pathlib for handling paths.

2. The choice of HTML, CSS, and JavaScript as the frontend programming languages was dictated by the use of the three programming languages in the modern browsers, absence of installation of any program in the user's computer, and creation of cross-platform user interface without any dependency at all.
3. The JavaScript Fetch API was used for sending the data between frontend and backend, thereby avoiding the use of any additional libraries for sending HTTP requests.

**Table 4.1:** Programming languages and tools used in the implementation

| S/N | Language / Tool | Purpose in the System |
|---|---|---|
| 1 | Python 3 | Backend HTTP server, prediction processing, model training, evaluation, and report generation |
| 2 | HTML5 | Structure and layout of the web-based questionnaire interface |
| 3 | CSS3 | Styling, responsive presentation, and visual feedback for form validation |
| 4 | JavaScript (ES6) | Client-side step navigation, input validation, API communication, and result rendering |
| 5 | JavaScript Fetch API | Sends questionnaire data as JSON to the backend prediction endpoint |
| 6 | JSON / JSON Lines | Serialisation format for the model artifact, evaluation reports, and prediction log |

### 4.1.2 Development Platform and Hardware Requirements

The program was developed and tested on an ordinary PC, having installed there an up-to-date operating system for personal computers. To start the server, you should run the file "server.py", but to access the program, one needs only to open any browser and point it to localhost:8000.

**Table 4.2:** Hardware and software requirements for the system

| **Component** | **Minimum Requirement** | **Recommended Requirement** |
|---|---|---|
| Processor | Dual-core CPU | Intel Core i3 or equivalent |
| RAM | 4 GB | 8 GB or more |
| Storage | 500 MB free disk space | 1 GB free disk space |
| Display | 1024 × 768 resolution | 1366 × 768 or higher |
| Web Browser | Chrome, Firefox, or Edge | Latest version of Chrome or Edge |
| Python Runtime | Python 3.x | Python 3.8 or later |
| Network | Not required for local use | Required only for cloud or networked deployment |

### 4.2 System Implementation and Modules

The system was implemented as a modular three-layer application. The five principal implementation modules and their responsibilities are summarised below.

**Table 4.3:** System implementation modules and responsibilities

| **Module** | **File(s)** | **Responsibility** |
|---|---|---|
| Presentation Module | index.html, styles.css, app.js | Renders the guided questionnaire, validates input, sends requests, and displays prediction results |
| API and Processing Module | server.py | Handles HTTP requests, normalises features, applies the trained model, generates advice, and logs records |
| Training and Evaluation Module | train_model.py | Loads the dataset, performs preprocessing, trains and compares algorithms, and generates evaluation reports |
| Model Artifact Module | model.json | Stores the serialised trained model, feature order, decision trees, and validation metrics for runtime loading |
| Data Storage Module | data/, evaluation_report.* | Stores the training dataset, prediction history, and evaluation output files |

The frontend questionnaire takes twelve inputs from patients in six stages and then makes a call to the Fetch API function in order to post a JSON data object to /api/predict. The server-side server.py program normalizes the received values, processes them using the trained random forest algorithm imported from model.json, predicts the probability, creates a management and escalation plan for the predicted probability, writes a new record in prediction_records.jsonl, and returns the whole response object in JSON format.

### 4.3 Testing and Evaluation

Various tests were performed in order to make sure that each of the modules works properly both on its own and in the context of an integrated system, as well as to evaluate the performance of the system as a whole in meeting the requirements outlined in Chapter Three.

#### 4.3.1 Evaluation Metrics

The performance of the predictive model was assessed using the following binary classification metrics:

**Table 4.4:** Evaluation metrics used for model assessment

| **Metric** | **Definition** |
|---|---|
| Accuracy | Proportion of total predictions that are classified correctly |
| AUC (Area Under ROC Curve) | Measure of the model's ability to discriminate between positive and negative cases across all classification thresholds |
| Precision | Proportion of cases predicted as high risk that are truly high risk |
| Recall (Sensitivity) | Proportion of truly high-risk cases that are correctly identified by the model |
| F1 Score | Harmonic mean of precision and recall, balancing both metrics |
| Balanced Accuracy | Mean of per-class recall, adjusted for class imbalance |

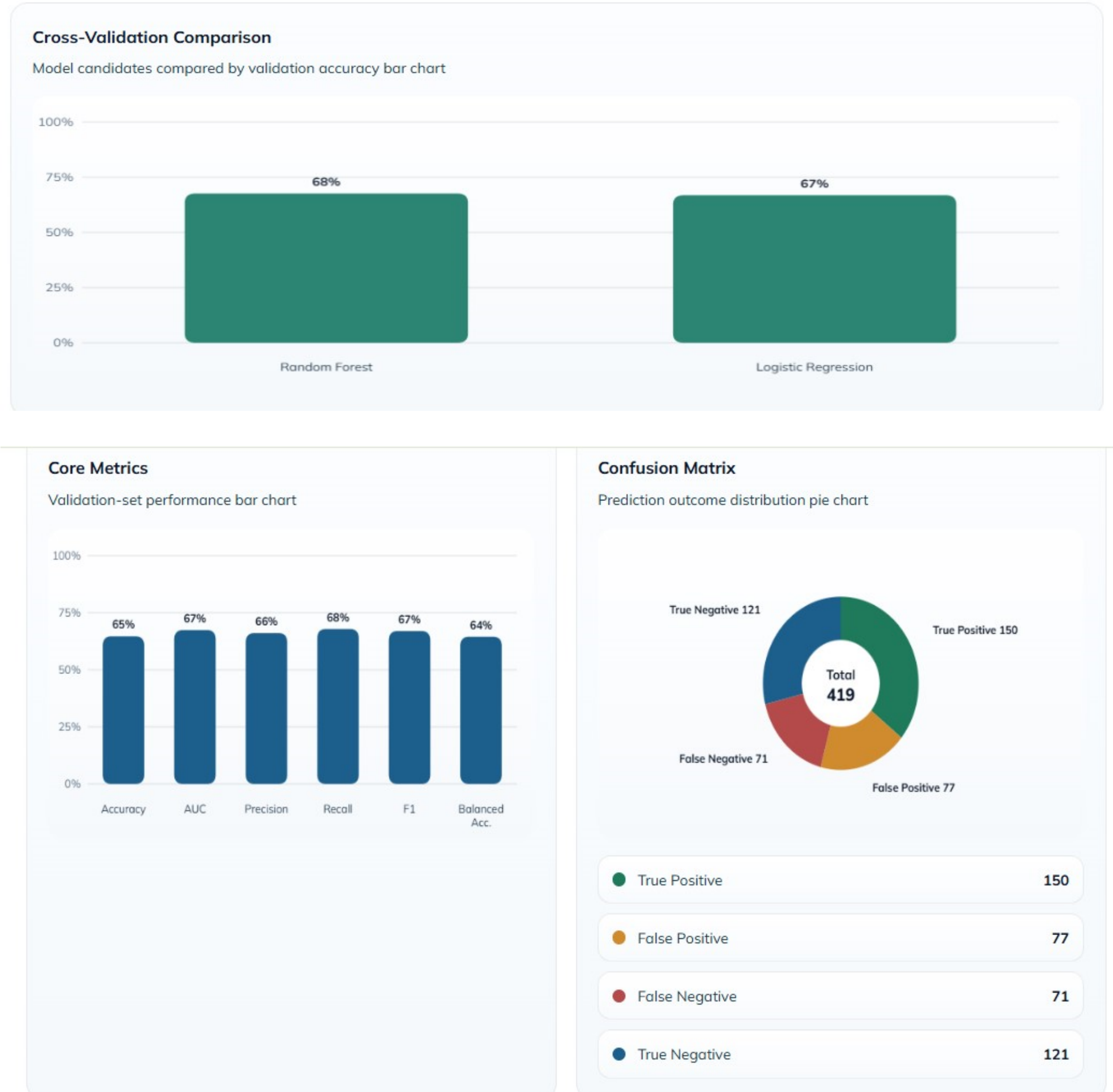


**Figure 4.1:** Sample of Model Evaluation metrics of the system

Based on the results of the confusion matrix from the validation data set, there were 150 True Positives, 121 True Negatives, 77 False Positives, and 71 False Negatives. This indicates that the classifier can discrimqinate reasonably well but still has an acceptaple amount of False Nnegatives.

### 4.3.2 System Testing

System testing connfirmed that all modunles operate correctly together across the complete prediction workflow. The test cases performed and their outcomes are summarised below.

**Table 4.5:** System test cases and results

| Test Case | Verification Target | Expected Outcome | Result |
|---|---|---|---|
| Server startup | server.py starts without errors | System accessible on port 8000 | Pass |
| Health endpoint | /api/health returns 200 OK | Backend confirms service availability | Pass |
| Valid prediction request | /api/predict accepts correct JSON | Prediction response returned with risk label and score | Pass |
| Input validation — missing field | Submission with missing required field | 400 error with informative message | Pass |
| Input validation — out of range | PSA or IPSS value outside allowed range | Validation message displayed, form not submitted | Pass |
| Record logging | Successful prediction stored in log | New timestamped record in prediction_records.jsonl | Pass |
| Records retrieval | /api/records returns history | Recent prediction records returned as JSON array | Pass |
| Retraining pipeline | train_model.py executes on updated dataset | model.json and evaluation reports regenerated | Pass |

### 4.3.3 User Interface Test and Walk-Through

The walk-through of the user interface testing is done step-by-step from accessing the landing page until prediction results are displayed. The following process is tested through the walk-through:

1. The user uses the browser interface and views the brief introduction as well as the clinical warnings on the landing page.
2. The user enters the age.
3. The user enters his/her PSA levels and IPSS score.

4. The user enters his/her body mass index, nocturia frequency, and daily urine void frequency.
5. The user indicates that the stream is weak.
6. The user verifies if there are risk factors such as family history and diabetes.
7. The user indicates the presence or absence of any red flag symptoms like hematuria, fever, and pain.
8. The user clicks the submit button and gets the prediction result in terms of the risk label, probability, features, and recommendation.

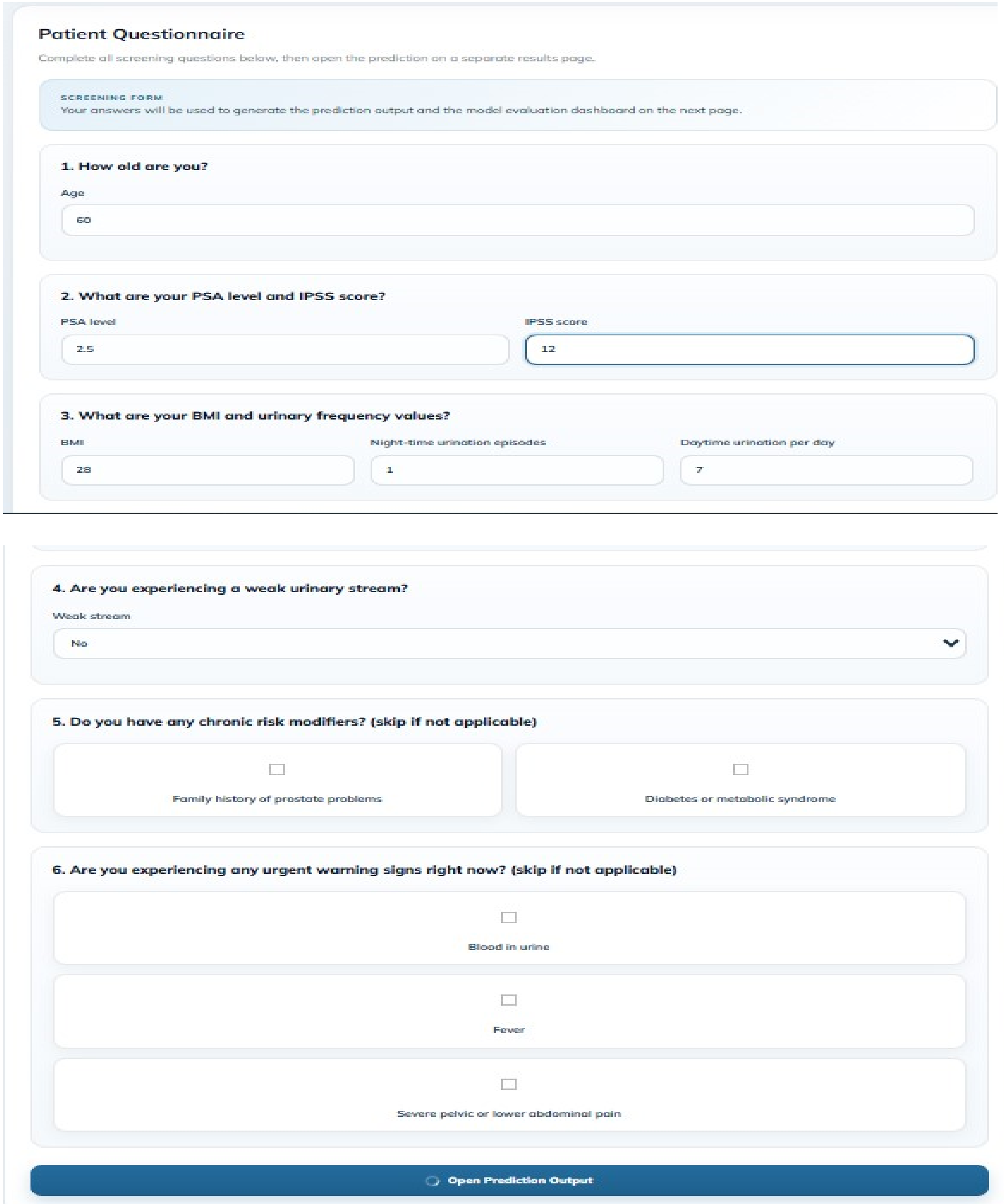


**Figure 4.2:** Sample system patient Questionaire intput

## 4.4 Results' Presentation and Analysis

Presentation and analysis of the obtained results' significance to the objectives of the project are provided below.

### 4.4.1 Sample Outputs And Screenshots

As the result of the questionnaire data entry, the system provides the following output components:

**Table 4.6:** System output components

| Output Component | Description |
|---|---|
| Risk Label | Final classification: Low, Moderate, or High |
| Probability Score | Model's estimated BPH risk likelihood as a percentage |
| Prediction Summary | Plain-language interpretation of the result |
| Extracted Features | Human-readable list of submitted clinical values |
| Management Strategies | Personalised care, monitoring, and referral recommendations |
| Escalation Advice | Urgent care instructions when warning sign flags are positive |
| Model Metadata | Algorithm, training size, AUC, and accuracy for transparency |

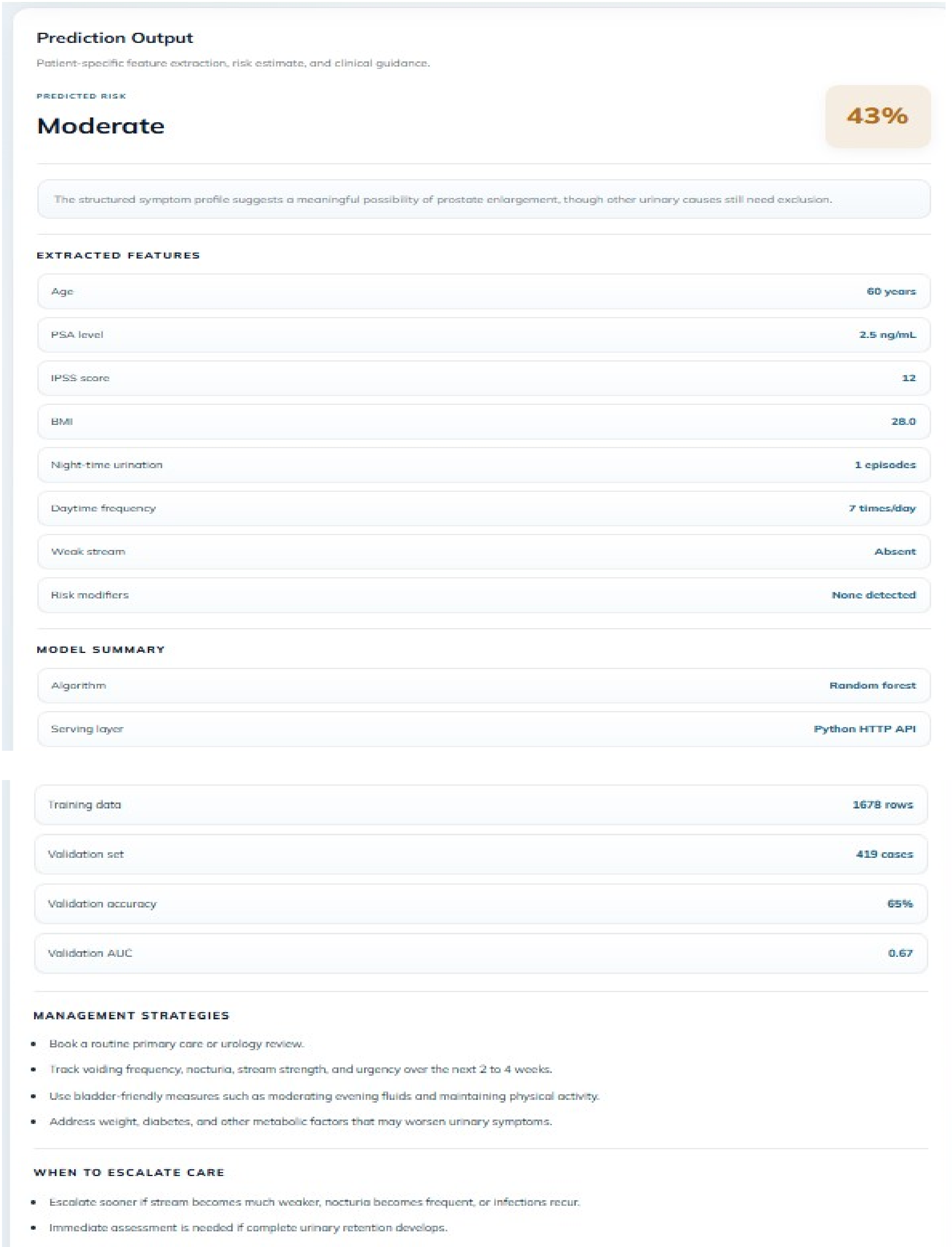


**Figure 4.3:** Sample system output for a moderate-risk patient profile

### 4.4.2 Discussion of Results and System Performance

The constructed classifier based on the random forest achieved the accuracy score of 64.68%, AUC of 67.40%, and F1 score of 66.96% on the test set used for validation of model performance. The achieved numbers demonstrate a clear advantage over the chance level of 50%, which means that the system was able to identify the predictive capacity of the nine input clinical variables.

In the comparison of performance metrics calculated from the cross-validation procedure for two models, the algorithm based on logistic regression showed a higher score of AUC and F1 measures. Nevertheless, random forest had a better score regarding the accuracy metric. Therefore, the achieved numbers confirm that the data set is quite informative for making a prediction about BPH development.

The main limitation of the application of the model in practice is the result of 71 false negatives obtained through the analysis of the confusion matrix on the sample size of 419 patients. False negatives, that is, cases in which the patients at risk were estimated as low-risk, can entail significant problems owing to the delay in rendering adequate medical attention. In addition, this corresponds to the initial assumption used in building the model, which assumes that the model will work as an auxiliary decision-making tool, rather than a diagnostic one. Recommendations on management, suggesting that the patients from moderate-risk categories get clinically examined, serve as compensation for this flaw.

# CHAPTER FIVE

# SUMMARY, RECOMMENDATIONS, AND CONCLUSION

## 5.0 Introduction

In this chapter, the following will be done in order to give an overview of all the works that have gone into the completion of this project, followed by some conclusions concerning the design, development, and evaluation of the Prostate Enlargement Risk Predictive System. Recommendations regarding how this project can be improved upon and further developed will be outlined in detail, as well as the knowledge gained from the study.

## 5.1 Summary

The main purpose of this project is solving one of the significant gaps in the management of prostates in Nigeria. This gap includes the lack of availability of a user-friendly tool for predicting and managing the early stages of benign prostatic hyperplasia (BPH).

As described in Chapter One, the context of this research has been established in terms of clinical issues and public health challenges. Benign prostatic hyperplasia is found to affect more than 50% of individuals within the age group 50-60, and almost 80% of males above 80 years of age. Three major issues that are identified in the literature include the following: diagnosis at a late stage, problems in taking into account several factors that contribute to the issue, and the lack of personalization in treatments provided to patients.

Chapter Two reviewed existing research on BPH epidemiology, risk factors, clinical score metrics such as the International Prostate Symptom Score (IPSS), and existing research on predicting BPH risk via machine learning techniques. This literature review found that the risk factors for BPH that are supported by clinical validation include age, PSA level, IPSS score, body mass index, nocturia, frequency of urination, weak stream, family history, and diabetes. Moreover, machine learning techniques have been shown to be effective at clinical screening similar to the one at hand.

Chapter Three outlined the design process following an iterative prototyping approach. UML-style diagrams were used to design the use cases, workflow of activities, class diagrams, and deployment diagram. The system was developed in accordance with a three-tier architecture: a guided questionnaire using the browser for the presentation tier, Python HTTP API for the application tier, and the intelligence tier which comprises the model, the reports, and the prediction history.

Details regarding the implementation of the system are presented in Chapter Four. The development environment used in the creation of the system entails the use of Python 3 as the programming language for the backend of the system as well as the machine learning algorithms, and HTML, CSS, and JavaScript for the frontend interface. Two classifiers were trained and compared in terms of performance by running five-fold cross-validation on a set of 2,097 clinical cases. Finally, the best performing classifier in terms of cross-validated accuracy was selected as the one to be deployed in the system, and this is the random forest algorithm which achieved an accuracy of 64.68%, an AUC score of 67.40%, F1 of 66.96% and balanced accuracy of 64.49%. All the modules used in the system functioned perfectly in validation tests.

**5.2 Conclusion**

In relation to the development of the Prostate Enlargement Risk Predictive System via this project, it will be observed that an effective and straightforward risk predictor can be built without the use of any advanced or clinical inputs. In this sense, the risk predictor software fulfills all the specified objectives as mentioned in Chapter One.

The classification performance provided by the predictor model, due to its training and testing using clinical data, is better than a random guess and shows excellent discrimination regarding the risk profiles of patients. In addition to the risk label and risk probability, many other details such as patient characteristics, risk-based methods for managing the condition, and emergency protocols are included in the risk prediction output.

The process of prototyping through iterations employed by our group in developing this project proved to be very effective. Each part of the system, whether it is the question-and-answer part, the predictor itself, or the training process, has been implemented step-by-step. Moreover, a modular approach allows us to make changes to any element without affecting other parts of the program.

Finally, one can conclude that our system functions well, has a proper modular design, and can serve as a great academic prototype for detecting cases of prostate enlargement.

**5.3 Recommendation**

Based on the results and limitations of this research, some suggestions have been developed for the future use of the developed system:

- Proper clinical validation: clinical validation is to be done based on the data obtained from healthcare institutions of Nigeria. Proper clinical validation will enhance the validity of the system based on the implementation on different datasets.
- Improving the training data: the present training data set can be improved by adding more information related to different patients to make the model more generalizable. Further information regarding clinical measures, including results of digital rectal examination and the volume of prostates, as well as the medication records of patients, will assist in developing prediction features.

- Model investigation: it is further recommended that advanced research be carried out on using other modeling techniques, including gradient boosting, neural networks, and stacking, in contrast to logistic regression and random forest comparison.
- Modifications of system architecture in order to integrate the system with EHRs: Modifications need to be performed on the architecture of the system in order to enable the integration of the algorithm with any hospital information systems and EHRs in order to log the predictions made using the algorithm in patients' medical records for follow-up.
- Enhancing mobility: A mobile application or a responsive design of an interface is recommended to enhance mobility, as adult males might prefer mobile phones over desktop computers while seeking healthcare services in their communities.
- Security features: The implementation of the proposed system will require authentication procedures and logging for compliance with patient data protection requirements and ethical norms of healthcare IT use.
- Translations into other languages: In order to ensure effective performance of the algorithm among communities in Nigeria and West Africa where the English language is less common, translation of the questionnaire into some of the Nigerian languages (Yoruba, Hausa, Igbo, etc.) is recommended.
- Learning and model improvement: An infrastructure for linking verified clinical outcomes to predictions and improving the algorithm is needed.

## REFRENCES


Bouhadana, D., Nguyen, D. D., Zorn, K. C., Bhojani, N., & Classé, J. (2023). Machine learning in benign prostatic hyperplasia: A systematic review. *Prostate Cancer and Prostatic Diseases,* 26(1), 1–9. https://doi.org/10.1038/s41391-022-00587-6

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chughtai, B., Forde, J. C., Thomas, D. D. M., Laor, L., Hossack, T., Woo, H. H., Te, A. E., & Kaplan, S. A. (2020). Benign prostatic hyperplasia. *Nature Reviews Disease Primers,* 2, 16031. https://doi.org/10.1038/nrdp.2016.31

Emberton, M., Cornel, E. B., Bassi, P. F., Fourcade, R. O., Gómez, J. M. F., & Castro, R. (2020). Benign prostatic hyperplasia as a progressive disease: A guide to the risk factors and options for medical management. *International Journal of Clinical Practice,* 62(7), 1076–1086. https://doi.org/10.1111/j.1742-1241.2008.01785.x

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Foster, H. E., Barry, M. J., Dahm, P., Gandhi, M. C., Kaplan, S. A., Kohler, T. S., Lerner, L. B., Lightner, D. J., Parsons, J. K., Roehrborn, C. G., Sarma, A. V., Wilt, T. J., & McVary, K. T. (2021). Surgical management of lower urinary tract symptoms attributed to benign prostatic hyperplasia: AUA Guideline. *Journal of Urology*, 200(3), 612–619. https://doi.org/10.1097/JU.0000000000000319

Langan, R. C. (2020). Benign prostatic hyperplasia. *Primary Care: Clinics in Office Practice,* 46(2), 223–232. https://doi.org/10.1016/j.pop.2019.02.003

Ling, Q., Liu, Z., Chen, X., & Ye, S. (2025). Machine learning models for prognosis prediction in prostatic diseases: A systematic review and meta-analysis. *Frontiers in Oncology,* 15, 1489203. https://doi.org/10.3389/fonc.2025.1489203

McVary, K. T., Roehrborn, C. G., Avins, A. L., Barry, M. J., Bruskewitz, R. C., Donnell, R. F., Foster, H. E., Gonzalez, C. M., Kaplan, S. A., Penson, D. F., Ulchaker, J. C., & Wei, J. T. (2021). Update on AUA guideline on the management of benign prostatic hyperplasia. *Journal of Urology*, 185(5), 1793–1803. https://doi.org/10.1016/j.juro.2011.01.074

Mourmouris, P., Tzelves, L., Berdempes, M., Lazarou, L., Skolarikos, A., & Mitropoulos, D. (2021). The use of machine learning in benign prostatic hyperplasia: A systematic review of the literature. *Prostate International,* 9(4), 172–176. https://doi.org/10.1016/j.prnil.2021.04.001

Okeke, C. J., Nwofor, A. M. E., Ikuerowo, S. O., & Obiora, C. C. (2021). Economic burden of benign prostatic hyperplasia treatment in a tertiary hospital in South-East Nigeria. *Nigerian Journal of Clinical Practice,* 24(3), 356–361. https://doi.org/10.4103/njcp.njcp_320_20

Parsons, J. K. (2020). Modifiable risk factors for benign prostatic hyperplasia and lower urinary tract symptoms: New approaches to old problems. *Journal of Urology,* 178(2), 395–401. https://doi.org/10.1016/j.juro.2007.03.103

Roehrborn, C. G. (2021). Benign prostatic hyperplasia: An overview. *Reviews in Urology,* 7(Suppl 9), S3–S14.

Shen, L., He, J., Dou, Y., Wu, Y., Xiong, M., Wu, J., & Gao, X. (2025). Development and validation of a machine learning model for predicting acute urinary retention after prostate biopsy. *BMC Urology,* 25, 37. https://doi.org/10.1186/s12894-025-01712-w

Wei, Y., Huang, S., Zhang, Z., Li, H., & Wu, Z. (2025). Global, regional, and national burden of benign prostatic hyperplasia from 1990 to 2021 with projections to 2035: A systematic analysis. *BJU International,* 135(2), 210–220. https://doi.org/10.1111/bju.16542

World Health Organization. (2021). *Noncommunicable diseases: Key facts.* https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Zhao, L., Chen, Z., Wang, Y., Liu, J., & Zhang, H. (2025). Artificial intelligence in prostate multiparametric MRI: A meta-analysis of diagnostic accuracy. *European Radiology*, 35(1), 89–101. https://doi.org/10.1007/s00330-024-11012-3

Zhou, J., Chen, W., Luo, X., Zhang, Q., & Li, F. (2025). A machine learning-based non-invasive prediction model for prostate disease assessment prior to biopsy. *Cancer Medicine,* 14(3), e70041. https://doi.org/10.1002/cam4.70041

Zi, H., Wang, Z., Long, Q., Liu, H., Wang, X., Gao, P., & Tian, J. (2025). Machine learning-based prediction model for benign prostatic hyperplasia: A study based on UK Biobank. *Frontiers in Medicine*, 12, 1521823. https://doi.org/10.3389/fmed.2025.1521823